LIÈGE université
Sciences
Mathématique

Contributions to spatial data analysis and Stein's method

Dissertation présentée par

Marie Ernst

en vue de l'obtention du grade de Docteur en Sciences

Sous la direction de Gentiane Haesbroeck & Yvik Swan

– Décembre 2019 –

Members of the committee

Advisors:

Prof. Gentiane Haesbroeck	Université de Liège
Prof. Yvik Swan	Université de Liège
	Université Libre de Bruxelles

Chair:

Prof. Philippe Lambert	Université de Liège

Members:

Dr. Ben Berckmoes	University of Antwerp
Prof. Christophe Ley	Ghent University
Prof. Gesine Reinert	University of Oxford

Summary

During my PhD thesis, I had the opportunity to work on two distinct thematics, namely robustness and Stein's method. Those two domains are my two supervisors' respective area of expertise. My thesis comprises two parts, each part corresponds to one of them.

The first part is based on robustness applied to spatial data. In the spatial context, two types of outliers may be detected, the local and the global ones. A local outlier corresponds to an observation for which the non-spatial attributes are significantly different from the ones of its neighbours. On the other hand, the global outliers have an atypical behaviour with respect to the whole dataset (Haslett et al., 1991). A literature review of the existing multivariate detection techniques is conducted in Chapter 1. Then, adaptations of the technique of Filzmoser et al. (2014) is proposed. The different proposals are compared by means of real data examples and simulations. This chapter corresponds to the published article Ernst and Haesbroeck (2017). As a matter of fact, this article is mentioned in different papers dealing with outlier detection with respect to some spatio-temporal and multivariate contexts.

Spatial techniques are useful only if the data are spatially autocorrelated. In Chapter 1, we consider a multivariate measure of autocorrelation based on the determinant of robust and regularized estimations of covariance matrices. Nevertheless, as discussed in Archimbaud et al. (2018), such estimations of covariance matrices (robust or not), can be problematic if the dimension increases. Indeed, as described in this paper, when the observations are reduced to a subspace of $\mathbb{R}^p$, the estimation of the covariance matrix does not allow to easily distinguish the characteristics of outliers from the regular observations.

A first natural guideline is then to study the spatial autocorrelation in the univariate setting. Therefore, we analyse tests of spatial autocorrelation which are the most widely used in practice. The considered tests are based on Moran's index (Moran, 1950), Geary's ratio (Geary, 1954) and Getis and Ord's statistic (Getis and Ord, 1992). In Chapter 2, the lack of robustness of classic tests is explained. Then, robust

alternatives of these tests are proposed. Finally, the power of different tests (classic and robust) are compared using simulations.

After the developments of Chapter 2 in the univariate setting, we wish to extend them into the multivariate case. Therefore, we had a look at multiple testing. In order to control the global level, classic tools are usually used in multiple testing. They are based on a level correction for each test (Benjamini and Hochberg, 1995, Benjamini and Yekutieli, 2001), a modification of the test statistics (see for instance Cai and Liu, 2016, for correlation tests) or even a transformation of the data in order to remove the dependency between them (Leek and Storey, 2008). The first method can usually be applied to any configuration, without taking care of specific dependence structures in the data. On the opposite, the last one requires the knowledge of the dependence structure. In order to safeguard the initial information (data and test statistics), we wonder about the multivariate distribution of p-values with respect to the initial dependency. An alternative is studying the distribution of the number of rejections. This variable has a discrete distribution which can be written as a sum of dependent indicators. This question was the starting point of Part II of the thesis, in which the original goal was to study discrete distributions using Stein's method. However, it occurs at this point that we stumbled on several new results which lead to several publications but also astray from our initial plan.

Consequently, Part II of the thesis focuses on Stein's method. This methodology is based on the characterisation of distributions by means of linear operators. A brief overview of this technique is given at the beginning of Chapter 3, followed by the general context considered hereafter. Then, we obtained several new probabilistic representations of inverse Stein operators (i.e., solutions to Stein equations) which opened the way to a wealth of new manipulations.

Chapter 4 details an application towards the important topic of variance bounds. We provide a generalization of Klaassen's variance bounds of arbitrary univariate targets under minimal assumptions. Our results hereby contain basically the entire literature on the topic, in a unified framework containing, in particular, both continuous and discrete distributions alike.

Chapter 5 deals with infinite covariance expansions and follows naturally from Chapter 4. In this chapter, a probabilistic representation of Lagrange's identity is used to obtain Papathanasiou-type variance expansions of arbitrary order. The expansions hold again for arbitrary univariate target distribution under weak assumptions, in particular they hold for continuous and discrete distributions alike. The weights are studied under different sets of assumptions either on the test functions or on the underlying distributions. These three chapters correspond to the submitted articles Ernst et al. (2019a,b).

The formalism introduced previously can also be used to deduce upper bounds on distances between distributions. Indeed, the total variation, Wasserstein and Kolmogorov distances may be expressed in terms of solutions to Stein equations. To bound the distances, a vast literature studies bounding the derivatives of solutions, called Stein's factors. They were introduced by Stein (1972) for Gaussian case and by Chen (1975) for Poisson distribution. Chapter 6 deduces Stein's factors from the developments of previous chapters. Upper bounds on Stein's factors and on different distances are obtained and compared to results which are already available in the literature for discrete and continuous distributions. The results of this chapter are the object of the preprint article Ernst and Swan (2019).

Finally, Stein's method can also be used to define other types of distances between distributions, for instance the Fisher information distance or the Stein discrepancy. A generalized Stein discrepancy is also useful in statistics. This concept introduced notably in Gorham and Mackey (2015) measures the dissimilarity between two distributions. Moreover, as it can be written as an expectation over one of the two distributions, the discrepancy can easily be empirically estimated using samples drawn from the second distribution. This particularity allows the definition of a goodness-of-fit test for any distribution with Stein operator. The papers Liu et al. (2016) et Chwialkowski et al. (2016) constructed such test for continuous distributions. We extend it to any univariate distribution under minimal assumptions. Lastly, the generalized Stein discrepancy could also be used to estimate parameters of a distribution with a "moment-type" method. The example of the K-distribution is used as illustration. The two applications are developed in Chapter 7 as well as some other perspectives.

Summary in French - *Résumé*

Au cours de mes années de doctorat, j'ai eu l'opportunité de travailler dans deux thématiques assez distinctes, d'un côté la robustesse et de l'autre la méthode de Stein. Ces deux domaines sont liés aux expertises respectives de mes deux promoteurs et composent les deux parties de cette thèse.

La première partie de cette thèse est basée sur l'étude de robustesse pour des données qui sont liées par une dépendance particulière, à savoir les données spatiales. Dans le contexte spatial, il est possible de détecter différents types d'observations atypiques, à savoir les atypiques locaux et globaux. Un atypique local a des valeurs observées sur les variables non-spatiales qui diffèrent fortement de celles des localisations voisines tandis que les atypiques globaux ont quant à eux un comportement atypique vis à vis de l'ensemble des données observées (Haslett et al., 1991). Une revue de la littérature des techniques de détection multivariée est faite dans le Chapitre 1. Ensuite, une nouvelle approche basée sur l'amélioration d'une technique existante, Filzmoser et al. (2014), est proposée. Les différentes procédures sont comparées à l'aide d'exemples et de simulations. Ce Chapitre correspond à l'article publié Ernst et Haesbroeck (2017). Celui-ci est d'ailleurs cité dans différents articles qui traitent des détections d'atypiques dans différents contextes spatio-temporels multivariés.

Ces techniques sophistiquées ne sont intéressantes qu'en présence de données autocorrélées spatialement. Dans le Chapitre 1, nous considérons une mesure d'autocorrélation multivariée basée sur l'estimation robuste et régularisée de déterminants de matrices de variance-covariance. Cependant, comme discuté dans Archimbaud et al. (2018), les estimations, robustes ou non, des matrices de variance-covariance peuvent poser problème lorsque la dimension du problème augmente. En effet, comme décrit dans Archimbaud et al. (2018), lorsque les observations considérées se situent dans un sous-espace de $\mathbb{R}^p$, l'estimation de la matrice de variance-covariance ne permet plus de distinguer efficacement les caractéristiques particulières d'observations atypiques de celles des autres observations.

Il est assez naturel d'étudier de plus près le problème d'autocorrélation spatiale

dans le contexte univarié. C'est pourquoi nous analysons les tests d'autocorrélation spatiale univariés les plus couramment utilisés par les praticiens, à savoir les tests basés sur les indices de Moran (Moran, 1950), Geary (Geary, 1954) et Getis et Ord (Getis et Ord, 1992). Dans le Chapitre 2, nous démontrons le manque de robustesse des tests classiques en présence d'observations atypiques. Des versions robustes de ceux-ci sont alors proposées. La puissance des différents tests (versions classiques et versions robustes) sont ensuite comparées à l'aide de simulations.

Après ce passage en univarié, nous souhaitons étendre les développements du Chapitre 2 au cas multivarié. C'est pourquoi nous nous intéressons ensuite au problème des tests multiples. Afin d'assurer un niveau global satisfaisant, les outils classiquement utilisés dans les tests multiples sont basés sur une correction des niveaux de chaque test (Benjamini et Hochberg, 1995, Benjamini et Yekutieli, 2001), sur une modification des statistiques de test (voir par exemple Cai et Liu, 2016, pour des tests de corrélation) ou encore une transformation des données afin de retirer la dépendance entre observations (Leek et Storey, 2008). Les premières méthodes sont généralement applicables dans toutes les situations sans tenir compte d'une structure de dépendance spécifique dans les données initiales tandis que la dernière nécessite quant à elle la connaissance de la structure de dépendance. Afin de préserver autant que possible les conditions initiales (pas de transformation des données ni des statistiques de test), nous nous sommes interrogés sur la distribution des p-valeurs multivariées en fonction de la dépendance initiale ou encore, sur la distribution du nombre de rejets, qui peut être exprimé comme une somme d'indicatrices dépendantes, à savoir des distributions discrètes. C'est à partir de cette problématique que nous avons décidé d'étudier de plus près les distributions discrètes à l'aide de la méthode de Stein. Cependant, notre travail sur la méthode de Stein a soulevé de nouvelles questions qui ont menés à plusieurs publications, mais qui nous a éloigné de cet objectif initial.

La Partie II de la thèse est ainsi consacrée à la méthode de Stein. Cette méthode est basée sur l'exploitation de la caractérisation d'une distribution à l'aide d'opérateurs. Pour démarrer le Chapitre 3, nous décrivons brièvement l'historique de la méthode et nous définissons le contexte général dans lequel nous travaillons. Ensuite, nous développons de nouvelles représentations des opérateurs inverses de Stein. Celles-ci sont intrinsèquement utiles pour le développement d'outils au sein de la méthode de Stein.

A l'aide de la méthode de Stein, des identités de variance et covariance sont construites dans le Chapitre 4. Nous obtenons une généralisation des bornes de variance de type Klaassen pour des distributions univariées arbitraires (discrètes ou continues). Notre résultat englobe de nombreux articles liés à ce sujet.

Ensuite, le Chapitre 5 développe des expansions infinies de covariance, ce qui suit

naturellement le chapitre précédent. Dans ce chapitre, une identité probabiliste de Lagrange est utilisées afin de définir une expansion de variance, d'ordre arbitraire, dans le style de Papathanasiou. De nouveau, ces résultats sont valables pour des distributions univariées arbitraires sous des conditions relativement faibles. Les différentes fonctions de poids qui interviennent dans l'expression sont détaillées pour différentes distributions discrètes et continues. Ces trois chapitres correspondent aux deux articles soumis Ernst et al. (2019a,b).

Les différentes quantités définies précédemment permettent également de déduire des bornes sur les distances entre différentes distributions. En effet, les distances entre distributions de probabilité (distance en variation totale, distance de Wasserstein et distance de Kolmogorov) peuvent s'exprimer à l'aide des solutions d'équations de Stein. Afin de borner ces dernières, tout une littérature s'intéresse à déterminer des bornes sur les différentes dérivées des solutions, appelées facteurs de Stein. Ceux-ci ont été introduit par Stein (1972) pour la distribution normale et par Chen (1975) pour la distribution de Poisson. Dans le Chapitre 6, nous utilisons le formalisme introduit dans le Chapitre 3 afin de développer les solutions des équations de Stein et d'en déduire des facteurs de Stein. Des bornes sur ces facteurs de Stein et sur différentes distances ont été obtenues et comparées avec les résultats déjà disponibles dans la littérature, que ce soit pour des distributions discrètes ou continues. Les résultats de ce chapitre sont repris dans l'article prépublié Ernst et Swan (2019).

Pour conclure cette partie, nous présentons deux applications aux statistiques ainsi que d'autres perspective dans le Chapitre 7. L'artillerie de la méthode de Stein permet de définir d'autres types de distances entre distributions, à savoir les notions de distance de Fisher généralisée et de divergence de Stein. Une généralisation de la notion de divergence de Stein peut également être utilisée dans le contexte statistique. Ce concept introduit notamment dans Gorham et Mackey (2015) permet de mesurer la dissimilarité entre deux distributions. Comme cet objet peut être exprimé comme une espérance liée à une seule des deux distributions, la divergence peut facilement être estimée empiriquement à partir d'échantillons. Cette caractéristique permet de définir notamment un test d'ajustement pour n'importe quelle distribution. Les articles Liu et al. (2016) et Chwialkowski et al. (2016) utilisent d'ailleurs cette mesure afin de construire un test d'ajustement pour des distributions continues. Nous proposons d'étendre ce développement afin de construire un test d'ajustement pour toute distribution univariée qui respecte des conditions minimales. La mesure de divergence peut également être utilisée afin de construire des estimateurs du type "estimateurs des moments" pour une distribution donnée. L'exemple de la K-distribution permet d'illustrer cette piste de recherche. Pour finir, différentes autres perspectives liées à la méthode de Stein sont brièvement exposées.

Acknowledgments - *Remerciements*

Ces années de doctorat ont été parsemées de nombreuses expériences de vie enrichissantes professionnellement mais également personnellement, qui m'ont permis d'avancer. Je tiens à remercier toutes les personnes qui y ont contribué de près ou de loin.

L'aventure de ce doctorat n'aurait pu démarrer sans ma promotrice, Gentiane Haesbroeck. Je la remercie vivement de m'avoir proposé de travailler avec elle dans le domaine de la statistique qui m'était alors relativement inconnu. Cela a toujours été un plaisir de travailler avec elle, que ce soit pour la recherche ou l'enseignement. Je la remercie pour son soutien, sa disponibilité et son écoute tout au long de ces années passées au B37.

Le remerciement suivant est évidemment adressé à mon co-promoteur, Yvik Swan, qui, pour sa part, m'a ouvert la porte du monde des probabilités. Sa passion pour les mathématiques a mené à une collaboration très riche et agréable, que ce soit pour la recherche, l'enseignement ou pour la vulgarisation des mathématiques via l'initiative MATh.en.JEANS. Ce projet, qu'il a lancé à Liège, m'a occupée et amusée pendant de nombreuses heures lors de mes années d'assistanat. Je le remercie pour sa disponibilité et sa confiance qui m'ont accompagnée pendant ses années passées à Liège.

I would like to thank Gesine Reinert for the fruitful collaboration on Stein's method. I really enjoyed working with her. I also thank Philippe Lambert who agreed to be a member of my thesis committee. Finally, I thank Ben Berckmoes and Christophe Ley who, along with Professors G. Haesbroeck, Y. Swan, G. Reinert and P. Lambert, made me the honour of being members of the jury.

During my PhD, I had the opportunity to present my work in several international conferences. These travels were possible thanks to partial funding via the Interuniversity Attraction Pole StUDyS P7/06 of the Belgian State and a Welcome Grant of the Université de Liège.

Même si le doctorat est principalement une aventure en solitaire, il n'aurait pas été réalisable sans l'accompagnement de mes collègues et amis au quotidien. Je commence par remercier Stéphanie, ma "soeur de thèse", qui a égayé mes vendredis après-midi et qui est la meilleure comparse pour les conférences. Disponible et à l'écoute, elle a toujours réussi à me motiver dans les moments de doute. Je remercie également tous mes collègues d'avoir parcouru un bout de chemin ensemble, et plus particulièrement le groupe des "matheuses bavardes" pour leur amitié.

Enfin, j'aimerais remercier ma famille et mes amis qui m'ont accompagnée durant ces années et qui m'ont permis de vivre pleinement en dehors de la thèse. J'adresse une pensée particulière à ceux qui ne sont plus là. Pour finir, je souhaite remercier plus particulièrement Jérôme d'avoir relu plusieurs passages de ce manuscrit et, surtout, pour sa présence à mes côtés, son soutien inconditionnel, sa patiente, son écoute et ses encouragements permanents.

Contents

Committee **i**

Summary **v**

Summary in French - *Résumé* **ix**

Acknowledgments - *Remerciements* **xiii**

Contents **xv**

I Spatial dependence

1 Outliers in spatial multivariate data **3**

1.1 Introduction 3
1.2 Local detection in spatial data 5
1.3 Local adaptation of the detection technique of Filzmoser et al. (2014) 8
1.3.1 Local structure 8
1.3.2 Restriction to homogeneous neighbourhoods 10
1.3.3 Modification of the reference set and of the comparison function 11
1.3.4 Tuning parameters 12
1.4 Examples 14
1.4.1 Social data in France 16
1.4.2 Geochemical data 18
1.4.3 Cancer data in France 21
1.4.4 Preliminary conclusion 22
1.5 Simulations 23
1.5.1 Spatial units 24
1.5.2 Spatial correlation structure 24
1.5.3 Contamination set-up 25
1.5.4 Performance measure 26

1.5.5 Results . . . 28
1.6 Conclusion . . . 29

2 Robustness of tests for spatial autocorrelation 37
2.1 Introduction . . . 37
2.2 Spatial autocorrelation indexes . . . 38
2.3 Tests for spatial autocorrelation . . . 41
2.4 Robustness of tests . . . 44
2.5 Robust versions of Moran's tests . . . 50
2.5.1 Moran index based on ranks . . . 51
2.5.2 Moran index using robust regression . . . 56
2.6 Simulation study . . . 59
2.6.1 Simulation setting . . . 59
2.6.2 Results . . . 60
2.7 Conclusion . . . 61
A Appendix . . . 68

General conclusion and perspectives 77

II Stein's method

Motivation: binomiality 81

3 Stein differentiation 85
1 Introduction . . . 85
2 Stein operators and Stein equations . . . 87
3 Representations of the inverse Stein operator . . . 94
4 Sufficient conditions and integrability . . . 98
5 The inverse Stein operator . . . 104

4 First order covariance identities and inequalities 107
1 Introduction . . . 107
2 Covariance identities and inequalities . . . 112
3 About the weights . . . 117
3.1 Score function and a Brascamp-Lieb inequality . . . 117
3.2 Stein kernel and Cacoullos' bound . . . 118
3.3 Discussion . . . 119

5 Infinite covariance expansions 125
1 Introduction . . . 125

2 A probabilistic Lagrange inequality . . . 128
3 Papathanasiou-type expansion . . . 130
4 About the weights in Theorem 5.3.1 . . . 132
4.1 General considerations . . . 132
4.2 Handpicking the test functions . . . 136
4.3 Illustrations . . . 137
A Appendix: proofs . . . 141

6 Stein factors and distances between distributions 153
1 Introduction . . . 153
2 Stein operators, equations and solutions . . . 157
2.1 The solutions to Stein equations . . . 162
2.2 Stein factors . . . 166
3 Bounds on IPMs and comparison of generators . . . 170
A Some more proofs . . . 183

7 General conclusions and perspectives 187
1 Kernelized Stein Goodness-of-fit tests . . . 189
2 A generalized MOM estimator . . . 195
3 Perspectives . . . 198

Bibliography 205

List of Figures 223

List of Tables 225

PART I

Spatial dependence

CHAPTER 1

Comparison of local outlier detection techniques in spatial multivariate data

1.1 Introduction

Spatial data are characterized by statistical units, with known geographical positions, on which non-spatial attributes are measured. Due to their respective positions, one expects some dependence between the statistical units under consideration, as Tobler's first law of geography states: *Everything is related to everything else, but near things are more related than distant things.*

Spatial data may be corrupted by atypical observations and following Haslett et al. (1991), one usually distinguishes two types of outliers. An observation might be an outlier in the traditional way, i.e., it lies far from the majority of the other data points in the space of the non-spatial attributes. In spatial statistics, such an observation is called a *global outlier*. An observation might also simply have non-spatial attributes with significantly differing values with respect to its neighbours. Such an observation is a *local outlier*. A local outlier might also be global. The observations can then be categorized into four groups: the local outliers, the global outliers, the local and global outliers and the regular observations. In practice, it is important to be able to identify these four groups. To illustrate these various types of outliers in spatial data, let us consider a real example in one dimension. The amount of annual waste per capita has been measured on the 262 Walloon municipalities in Belgium, as illustrated on Figure 1.1. Clearly, the municipalities Lens, Froidchapelle, Waterloo, La Hulpe and Braine-le-Château are global outliers as their observed values (colored in white) are outlying with respect to the majority of the data observations. Braine-le-Château is also a local outlier as it differs strongly from its neighbours.

Finally, Eupen is a local outlier but not a global one; its observed value is coloured in pink while the values of the surrounding municipalities are coloured in green.

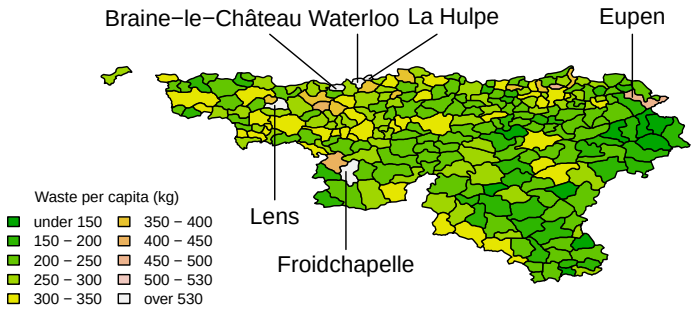


Figure 1.1: Illustration of local and global outliers in a univariate setting using waste per capita (kg) in Walloon municipalities: Braine-le-Château (global and local), Eupen (local), Froidchapelle (global), La Hulpe (global), Lens (global) and Waterloo (global).

Detecting global outliers is usually performed by means of classical detection techniques, like those using Mahalanobis-type distances based on robust estimations of location and covariance. This part of the detection problem is not further discussed. Instead, focus is on the local detection. Schubert et al. (2014) reviewed the local detection techniques available in the literature, with a particular emphasis on spatial data. However, in their Section 5.6, they stress that most known techniques are only able to detect local outliers when a single non-spatial attribute is observed. However, they mention that the methods advocated by Sun and Chawla (2004) and Chawla and Sun (2006) might be applied in the multivariate case even though these authors did not clearly acknowledge that fact. Moreover, the detection technique LOF and its variants, that are fully described in Schubert et al. (2014), are illustrated in the multivariate setting in Breunig et al. (2000) and widely available in the `R` package `dprep`. Other local detection techniques exist and may be applied in the multivariate case. Indeed, Chen et al. (2008), Filzmoser et al. (2014) and Harris et al. (2014), all develop detection techniques applicable to multivariate non-spatial attributes, but

these were not considered in the review of Schubert et al. (2014). Therefore, the main aim of this chapter is to provide a complement to the paper of Schubert et al. (2014) by expressing these known multivariate methods along the lines of the general framework based on the context and model functions introduced in Schubert et al. (2014). The literature review stretches to 2016, the year of submission of our paper Ernst and Haesbroeck (2017). The papers published afterwards are discussed in the conclusion. In parallel, a second contribution is to slightly adapt the procedure of Filzmoser et al. (2014) in order to increase its local characteristics. The main advantage of the adapted procedure is to improve the exploratory analysis of data by providing additional insights on the detected local outliers.

More specifically, Section 1.2 reviews the existing multivariate local techniques following Schubert et al. (2014)'s approach while Section 1.3 describes the local adaptations that can be applied to Filzmoser et al.'s technique in order to improve its local nature. Then, the different proposals are compared by means of real data examples (Section 1.4) and by means of simulations (Section 1.5). Some conclusions follow in Section 1.6.

1.2 Local detection in spatial data

Schubert et al. (2014) decompose any local detection procedure in basically two steps: first, a kind of outlyingness measure (typically a distance) is computed for each spatial unit and secondly, this measure is compared with those computed on other spatial units to decide whether it is outstanding.

Schubert et al. (2014) describe the computation of the outlyingness measure by means of a *model function* which is applied on a possibly restricted set of observations (usually the neighbours of the spatial unit under consideration). This subset of data points is called the *context set* and when it does not contain all the data points, the outlyingness measure has a *local* flavor. In the outlier detection technique reviewed by Schubert et al. (2014), most (but not all) context sets are local.

The comparison step, performed by means of a *comparison function*, is based on the measures derived on a given set of units, this set being possibly different from the set of neighbours. In Schubert et al. (2014) terminology, this other subset is called the *reference set*. It might contain all the data points (yielding a *global* comparison) or a subset of these (corresponding to a *local* comparison). At the end of the process, the detection technique yields either a binary classification (clean-outlier) or an outlier score (describing the degree of outlyingness of the observation).

Let us note that, when the comparison is made on a local level, Kriegel et al. (2011) stress the necessity to add a normalization or regularization step in order to get outlier scores that are comparable and interpretable. However, as we focus on

techniques resulting on a binary classification, we do not consider this additional refinement of the process.

In order to describe the two-step components of the multivariate local detection techniques outlined in the Introduction, some definitions and notations need to be introduced. Let $z_1, \ldots, z_n$ denote the p-dimensional observations associated with some spatial coordinates $s_1, \ldots, s_n$, i.e., z_i is the observed value of $Z(s_i)$ where Z is a p-variate random vector. Focusing on local outlyingness requires to define a *neighbourhood* for each observation. Let $\mathcal{N}_i$ denote the neighbourhood of the location s_i, $1 \leq i \leq n$. Any type of neighbourhood might be considered since, following Schubert et al. (2014)'s philosophy, the choice of the neighbourhood should be made independently from the choice of the detection technique. For instance, it could be decided to construct neighbourhoods containing a fixed number of observations, k say, the observations selected in $\mathcal{N}_i$ being the $k-1$ nearest neighbours of s_i. The closeness is assessed by means of an appropriate distance (e.g. Euclidean distance or orthodromic track) computed on the spatial coordinates. To keep the choice of the neighbourhoods unspecified, the number of neighbours in $\mathcal{N}_i$ is denoted as n_i throughout the text.

Most techniques compute, at some point in their process, a Mahalanobis-type distance. Let μ and Σ denote respectively a p-dimensional vector (a center) and a $p \times p$ positive-definite matrix (a variance-covariance structure) and consider a p-variate observation z. The squared distance between z and μ while taking into account the correlation structure inherent to Σ is denoted as

$$d_{\mu,\Sigma}(z) = (z-\mu)^T \Sigma^{-1} (z-\mu).$$

In practice, estimations of μ and Σ are required in order to compute these distances. Classically, the sample mean and covariance matrix are used but, in a perspective of outlier detection, robust alternatives should be favored. All the techniques reviewed in this section rely on the Minimum Covariance Determinant (MCD) estimator (Rousseeuw, 1985). For a random sample $\{z_1, ..., z_n\}$, with $z_i \in \mathbb{R}^p$, the MCD estimator is determined by selecting a subset of h observations (with $n/2 \leq h \leq n$) which minimizes the generalized variance among all possible subsets of size h. The MCD location and scatter estimations are then given by the sample mean and the sample covariance computed from this subset.

Following a similar presentation as in Table 4 of Schubert et al. (2014), here is now the description of the context and reference steps of the three listed multivariate techniques.

1. Median Algorithm, (Chen, Lu, Kou, and Chen, 2008)

   This proposal is a multivariate extension of the univariate approach (described also in Chen et al., 2008, but already introduced in Lu et al., 2004) based on the detection of the outlying distances computed between the observed non-spatial attribute of a spatial unit and the median of that attribute over its neighbours.

Context	Model function
$\mathcal{N}_i$	Computation of h_i which is the difference (in $\mathbb{R}^p$) between z_i and the vector of marginal medians computed on z_j with $s_j \in \mathcal{N}_i$.
Reference	Comparison function
Global	Computation of the distances $d_{\hat{\mu},\hat{\Sigma}}(h_i), 1 \leq i \leq n$, where $\hat{\mu}$ and $\hat{\Sigma}$ are the MCD location and dispersion estimators computed on $h_1, \ldots, h_n$. Comparison of these distances with a F-quantile.

2. Detection technique of Filzmoser, Ruiz-Gazen, and Thomas-Agnan (2014)

   In some cases, as explained by Schubert et al. (2014), the context or the comparison steps might be divided into several sub-steps, not based on the same context or reference sets. This happens in the approach of Filzmoser et al. (2014) as a global estimation step needs to be carried out before working inside each neighbourhood.

Context	Model function
Global	Robust estimation of the center and dispersion of $\{z_1, \ldots, z_n\}$ by means of the MCD estimator; yielding $\hat{\mu}$ and $\hat{\Sigma}$.
$\mathcal{N}_i$	Computation of the n_i distances $d_{z_i,\hat{\Sigma}}(z_j)$ with $s_j \in \mathcal{N}_i$. Computation of the isolation degree of s_i.
Reference	Comparison function
Global	Comparison of the isolation degrees and selection of the largest ones.

3. Geographically weighted detection (Harris, Brunsdon, Charlton, Juggins, and Clarke, 2014)

   When the dimension is large, an additional (and global) step is advocated by the authors in the context framework, as described below.

Context	Model function
Global (optional)	Reduction of the dimension with robust PCA.
$\mathcal{N}_i$	Application of a Geographically Weighted PCA in $\mathcal{N}_i$ Computation of score distances (SD), orthogonal distances (OS) and component scores (CS).
Reference	Comparison function
Global	Comparison of the univariate measures SD, OS, and CS with theoretical quantiles or empirical quantiles.

One can see that all methods work locally, at least partially, for the context part of the procedure while the comparison step is operated on a global level. The number of steps performed on a local level yields the so-called *degree of locality* of the search procedure, as defined in Schubert et al. (2014). The above techniques have a single local step in their process. Let us note also that Harris et al. (2014) as well as Filzmoser et al. (2014) distinguish local and global outliers and separate the search of the two types of outliers. Only the local detection is taken into account in the description here. Chen et al. (2008) do not mention the different types of outliers and detect all of them indifferently.

The local nature of the detection technique of Filzmoser et al. (2014) is restricted to a single step and this local step is preceded by a preliminary global step in order to compute the overall correlation structure of the data. Transforming this initial step into a local one is one of the elements implemented in the adaptation, as outlined in the next section.

1.3 Local adaptation of the detection technique of Filzmoser et al. (2014)

1.3.1 Local structure

In Filzmoser et al. (2014), the model function used in the neighbourhood $\mathcal{N}_i$ is based on the computation of the pairwise squared distances

$$d_{z_i,\hat{\Sigma}}(z_j) = (z_j - z_i)^T \hat{\Sigma}^{-1}(z_j - z_i) \text{ with } s_j \in \mathcal{N}_i \tag{1.1}$$

which rely on the robust estimation of the *global* correlation structure. The global correlation structure (as well as the global center $\hat{\mu}$) is also at the core of the computation of the isolation degree as this degree is a quantile of a decentralized χ^2 distribution with non-centrality parameter given by $d_{\hat{\mu},\hat{\Sigma}}(z_i)$. Using the same overall structure implicitly assumes that the data are stationary, but may prove to be inefficient when the neighbourhoods have different shapes.

Therefore, in order to increase the local nature of the procedure, we suggest to plug locally estimated covariance matrices into the definition of the pairwise squared distances (1.1), yielding so-called local squared distances

$$d_{z_i,\hat{\Sigma}_i}(z_j) = (z_j - z_i)^T \hat{\Sigma}_i^{-1}(z_j - z_i) \text{ with } s_j \in \mathcal{N}_i$$

where $\hat{\Sigma}_i$ is estimated using only the attribute values of the statistical units belonging to $\mathcal{N}_i \cup \{s_i\}$. Now, some care is required in the estimation process as the number of observations included in each neighbourhood, n_i, may be small (typically a fraction of the sample size) and, in high-dimensional cases, the number of units in $\mathcal{N}_i \cup \{s_i\}$ may even be smaller than the dimension p.

To ensure the positive-definiteness of the estimated covariance matrix, using regularized estimators is an option that is suggested in the literature (Witten and Tibshirani, 2009, Friedman et al., 2008). Moreover, as detection of outliers is at sake here, robustness should also be advocated. Therefore, the regularized version of the Minimum Covariance Determinant estimator outlined in Fritsch et al. (2011) is used for the local and robust estimation of the covariance matrix in each neighbourhood.

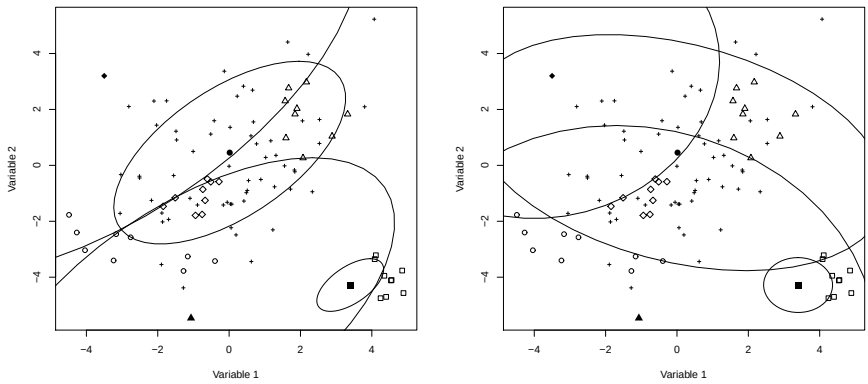
The regularized MCD estimator is obtained by the maximization of the penalized negative log-likelihood function restricted to a subset of $n/2 \leq h \leq n$ observations, i.e.,

$$\log|\Sigma| + \frac{1}{n}\sum_{z_j \in H}(z_j - \mu)^T \Sigma^{-1}(z_j - \mu) + \lambda \mathrm{Tr}\Sigma^{-1}$$

As explained in Fritsch et al. (2011), the FAST-MCD algorithm of Rousseeuw and Driessen (1999) may be adapted in order to compute the regularized version of the MCD estimator. As a final remark concerning the use of a robust and regularized estimator in the detection procedure, let us note that it does not relate to the regularization step suggested in Kriegel et al. (2011).

As illustration, let us consider the artificial data set, named `dat`, of the `R` package `mvoutlier`. It consists of $n = 100$ observations distributed according to the bivariate normal distribution contaminated by some outliers (see the scatter plot of the non-spatial attributes on Figure 1.2). Filzmoser et al. (2014) highlighted the four observations represented by full symbols on Figure 1.2 (panel a). Each of these observations has nine neighbours represented by identical (but empty) symbols. One can see that the full diamond and triangle are local and global outliers, the full circle is a local outlier and the full square is a global outlier. On panel a, the way the detection technique of Filzmoser et al. (2014) works can be visualized. For any full symbol, say z_i, the isolation degree may be obtained by computing the confidence level of the ellipse centred at z_i and shaped according to the global structure, the ellipse being inflated until it covers the attributes of its *next neighbour* (i.e., the neighbour whose

non-spatial attribute z_j lies the closest to z_i). On panel b, the same approach is followed, but this time the structure of the ellipse varies from one point to the next since it is locally estimated by means of the regularized MCD estimator.



(a) (b)

Figure 1.2: Artificial data set `dat` of the `R` package `mvoutlier`: (a) illustration of Filzmoser et al.'s detection technique and (b) illustration of its adaptation.

1.3.2 Restriction to homogeneous neighbourhoods

A second adaptation can be added to the methodology in order to take into account the possible heterogeneity of the attributes of the spatial units included in a given neighbourhood. Indeed, as also stressed in Chawla and Sun (2006), an observation should not be classified as a local outlier if its non-spatial attributes differ from those of its neighbours because they are simply lying in an unstable area. Therefore, only the spatial units whose neighbourhoods consist of spatial units with non-spatial attributes that are sufficiently concentrated in $\mathbb{R}^p$ could be considered in the detection process.

To measure the concentration in the space of the non-spatial attributes, the volume of the ellipsoid centred at the robust estimation of the local mean, shaped according to the locally estimated covariance matrix, could be computed. The volume of such an ellipsoid is proportional to the determinant of its structure matrix (the proportionality factor depending only on the dimension p and not on the number of points in the neighbourhood). However, as the sizes of the neighbourhoods might vary, comparing the determinants of the locally estimated covariance matrices is not

appropriate. The ellipsoid must be scaled (inflated or deflated) to make the different cases more comparable.

More precisely, here is the approach that has been followed to measure the concentration inside the neighbourhoods. Assume that the i-th neighbourhood is under consideration (with sample size n_i) and let $\hat{\mu}_i$ and $\hat{\Sigma}_i$ be the regularized MCD estimations derived on z_j, $s_j \in \mathcal{N}_i \cup \{s_i\}$. The estimated covariance matrix $\hat{\Sigma}_i$ is characterized by a given *size*, i.e., its determinant, and a given *shape* defined by the matrix $\hat{V}_i$ given by $\hat{\Sigma}_i / \sqrt[p]{\det \hat{\Sigma}_i}$. Using the shape matrix instead of the covariance matrix in the construction of the ellipsoids yields ellipsoids of comparable volumes for all the neighbourhoods (as the determinants of all shape matrices are equal to 1). An appropriate measure of concentration in the i-th neighbourhood, c_i say, may then be defined as follows:

$$c_i = \frac{1}{h_i} \sum_{j:s_j \in H_i} d_{\hat{\mu}_i, \hat{V}_i}(z_j)$$

where H_i is the optimal subset corresponding to the regularized MCD estimations computed on $\mathcal{N}_i \cup \{s_i\}$, this subset containing non-outlying observations by construction. Multiplying the shape matrix $\hat{V}_i$ by c_i (which may be interpreted as a deflation or inflation factor), the volume of the ellipsoid becomes proportional to $\det(c_i \hat{V}_i) = c_i^p$. Another option for computing this measure of concentration would be to replace the mean operator by the median, i.e.,

$$c_i = \underset{j:s_j \in \mathcal{N}_i}{\text{median}}\ d_{\hat{\mu}_i, \hat{V}_i}(z_j).$$

This yields results that are quite similar, as illustrated in Section 1.4.1.

Finally, the resulting volumes, or equivalently the resulting mean squared distances c_i, are then ranked from the smallest (i.e., most concentrated ellipsoid) to the largest (i.e., the biggest ellipsoid). Only the spatial units having neighbourhoods characterized by a volume ranked among the $\lceil \beta \times n \rceil$ smallest (for an appropriate value of β as discussed in the next Subsection) are further considered in the local detection technique. The set of spatial units selected for the final step of the detection is denoted as $\mathcal{D}$.

1.3.3 Modification of the reference set and of the comparison function

The replacement of the global estimation $\hat{\Sigma}$ by a local one and the restriction of the search to the spatial units having a homogeneous neighbourhood have a direct impact on the final steps of Filzmoser et al.'s technique. Indeed, the distributional result allowing to compute the isolation degree by means of a quantile of a decentralized

χ^2 distribution is no longer valid. This parametric approach has been replaced by a non-parametric one using the local distances between each observation and its next neighbour, the *next neighbour* of a given observation being the neighbour whose non-spatial attributes lie closer to the non-spatial attributes of the observation. The closeness is therefore measured in the space of the non-spatial attributes and not in the space of the spatial coordinates. When this distance is large, the corresponding observation is tagged as a local outlier.

In summary, the adapted Filzmoser et al. technique, referred to as the *regularized spatial detection technique* from now on, might be described by the following steps:

Context	Model function
$\mathcal{N}_i$	Robust and regularized estimation of the center and dispersion of the data $\{z_j, s_j \in \mathcal{N}_i \cup \{s_i\}\}$; yielding $\hat{\mu}_i$ and $\hat{\Sigma}_i$. Computation of the deflation factor c_i.
Global	Ranking of c_i, $1 \leq i \leq n$, and selection of the units s_i corresponding to the $\lceil \beta \times n \rceil$ most homogeneous neighbourhoods.
Reference	**Comparison function**
$\mathcal{N}_i$ with $s_i \in \mathcal{D}$	Computation of the squared distances of the closest neighbours $\min_{s_j \in \mathcal{N}_i} d_{z_i, \hat{\Sigma}_i}(z_j)$.
$\mathcal{D}$	Comparison of the distances and selection of the largest ones.

Let us observe that there are now two distinct local steps in this procedure, increasing by 1 the degree of locality of the initial procedure.

1.3.4 Tuning parameters

There are several parameters that need to be chosen in order to apply the regularized spatial detection technique. First, the local estimation step requires the tuning of two parameters: the coverage of the MCD estimator (i.e., the number h of observations included in the MCD calculations) and the regularization parameter λ. Then, a fraction β has to be chosen in order to keep only the most concentrated neighbourhoods.

1. Coverage of the regularized MCD estimator
   Usually, the coverage is chosen according to the breakdown point one wants to achieve by taking $h = \lceil n \times (1-\alpha) \rceil$ where $0 < \alpha < 1/2$ is the chosen breakdown value. The breakdown point is, roughly speaking, the smallest fraction of contamination which renders the estimations meaningless. Under regularization,

as the sample size n might be smaller than the dimension p, defining h as above is misleading as p might be bigger than n. In fact, one can show that the breakdown point of the regularized MCD estimator is given by $\min(h, n-h+1)/n$. As it is quite natural not to expect more than 25% of outliers inside each neighbourhood, it was decided to set the coverage rate to the proportion 0.75 (i.e., $h_i = \lceil (n_i + 1) \times 0.75 \rceil$) for all the local estimations. Of course, to achieve robustness, it is necessary to have $h_i < n_i + 1$, which is only guaranteed if the size of the neighbourhood is at least equal to 3. Therefore, spatial units which are quite isolated and have less than three neighbours cannot be considered by the regularized spatial detection technique, unless their neighbourhoods are inflated until reaching the minimum required number of neighbours.

2. Regularization parameter λ

   The regularization parameter was locally set following a suggestion outlined in Fritsch et al. (2011). Indeed, as the penalty function considered in their paper is based on the trace of the concentration matrix (the inverse of the covariance matrix), a value of λ equal to $\mathrm{tr}\Sigma/np$ would yield an unbiased estimation of the trace of the covariance matrix. Inspired on this idea, λ may be locally set in each neighbourhood to the value $\widehat{\mathrm{tr}\Sigma_i}/h_i p$ where $\widehat{\mathrm{tr}\Sigma_i}$ should be robustly estimated. To do so, it is sufficient to get robust estimations of each marginal variance. Let $\hat{\sigma}_{i\ell}$ denote the marginal median absolute deviation of the ℓ-th coordinate of Z computed in the ith neighbourhood. Then, $\sum_{\ell=1}^{p} \hat{\sigma}_{i\ell}^2$ does the job.

3. Homogeneity proportion

   As explained above, only the spatial units corresponding to a given proportion (β say) of the most homogeneous neighbourhoods are further analysed in the regularized spatial detection technique. Taking β too large (i.e., keeping spatial units whose neighbours have a heterogeneous pattern) tends to increase the false detection rate. On the other hand, taking β too small might be too restrictive if some small neighbourhoods contain several local outliers. To enrich the exploration analysis of the data, we advise to choose a whole range of β values and to visualize the results on adjusted boxplots. More specifically, select a grid of proportions for β, for example 0.1, 0.25, 0.5, 0.75, 0.9, and for each of these, plot the "next distances" of each spatial unit by means of a boxplot, adjusted to take into account the asymmetry of the distribution of the distances. Full details on the construction of these adjusted boxplots are available in Hubert and Vandervieren (2008) but it is interesting to note that the fence of the

boxplot is defined by

$$[Q_1 - 1.5\ e^{-4\mathrm{MC}} IQR; Q_3 + 1.5\ e^{3\mathrm{MC}} IQR]$$

where MC denotes the medcouple, which measures the skewness of the sample and is given by

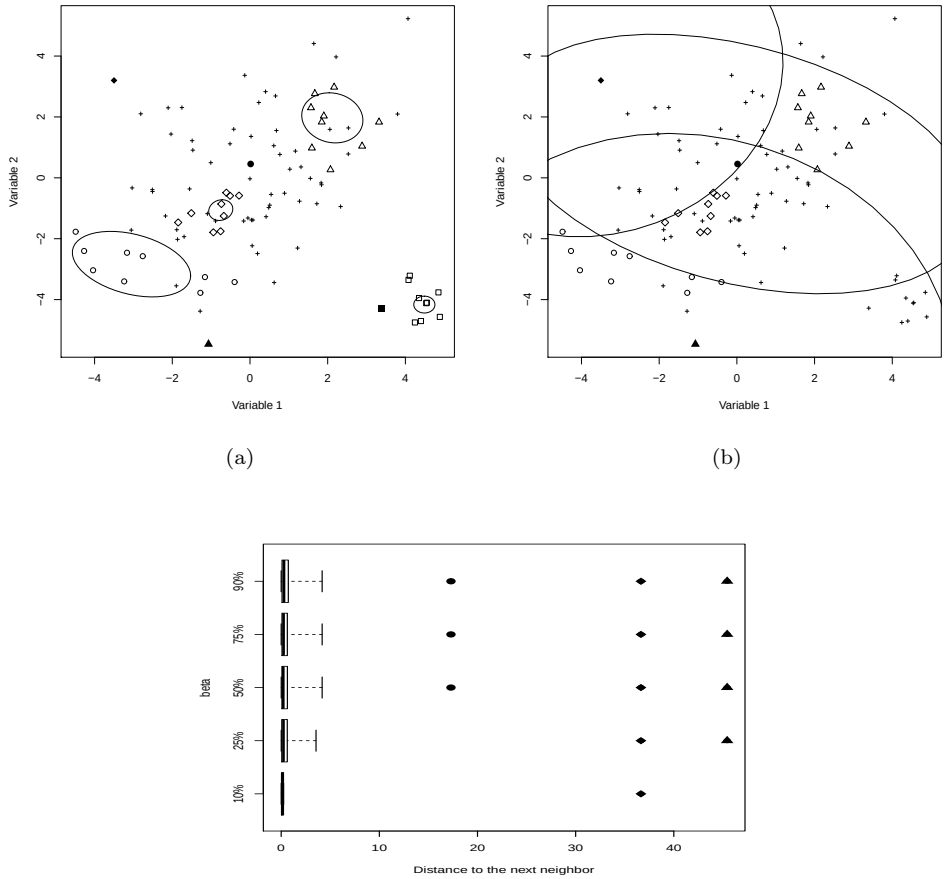
$$\mathrm{MC} = \underset{z_i \leq Q_2 \leq z_j}{\mathrm{median}} \frac{(z_j - Q_2) - (Q_2 - z_i)}{z_j - z_i}.$$

The next distance associated with s_i corresponds to the smallest distance among the locally estimated squared distances $d_{z_i,\hat{\Sigma}_i}(z_j)$, with $s_j \in \mathcal{N}_i$. The units having much bigger next distances than the others may then be flagged as local outliers, a natural cutoff being given by the upper whisker of the adjusted boxplot. The simultaneous consideration of several values of β allows to measure the degree of outlyingness of the observations and to visualize the potential impact the local heterogeneity might have on this outlyingness.

Going back to the artificial data `dat` of the `R` package `mvoutlier`, Figure 1.3 illustrates the effect of the choice of β on the results of the detection. Looking first at the boxplots (panel c), one can see that choosing $\beta = 0.1$ (first boxplot) only yields the full diamond as local outlier. When $\beta = 0.25$, the full triangle is also classified among the local outliers. Finally, when $\beta \geq 0.5$, a third local outlier (the full circle) appears. The plot on panel a shows that the structure and the homogeneity of the neighbourhoods vary for the different symbols, illustrating the necessity to adjust the structure locally and to focus only on the stable areas. On panel b, the local distance of the closest neighbour is illustrated using the ellipse centred at the observation z_i and inflated until reaching its closest neighbour.

1.4 Examples

In this section, examples already considered in the literature are exploited to compare the detection techniques reviewed or introduced in Sections 1.2 and 1.3. All the computations are done in the statistical software `R`. The detection technique of Filzmoser et al. (2014) was applied via the procedures `locoutPercent`, `locoutneighbor` and `locoutSort` of the package `mvoutlier` and the outliers are detected visually by means of Filzmoser et al.'s suggested graphical display. The procedure of Harris et al. (2014) was partially re-implemented using, as a core component, the procedure `gwpca` of the package `GWmodel`. All the available tools (i.e., the score distances SD, the orthogonal distances OD and the component scores CS) were computed and compared to empirical quantiles (the theoretical quantiles advocated by Harris et al. (2014) for the two measures SD and OD are too small in most examples, due probably to the non normality of the data). Note that the use of the procedure `gwpca` restricts the



(a) (b)

(c)

Figure 1.3: Illustration of the regularized spatial detection technique on the artificial data `dat` of the `R` package `mvoutlier`: (a) comparison of the homogeneity of the neighbourhoods, (b) representation of the ellipses, centred at z_i and inflated until reaching their closest neighbours and (c) boxplots of "next distances" for varying values of β.

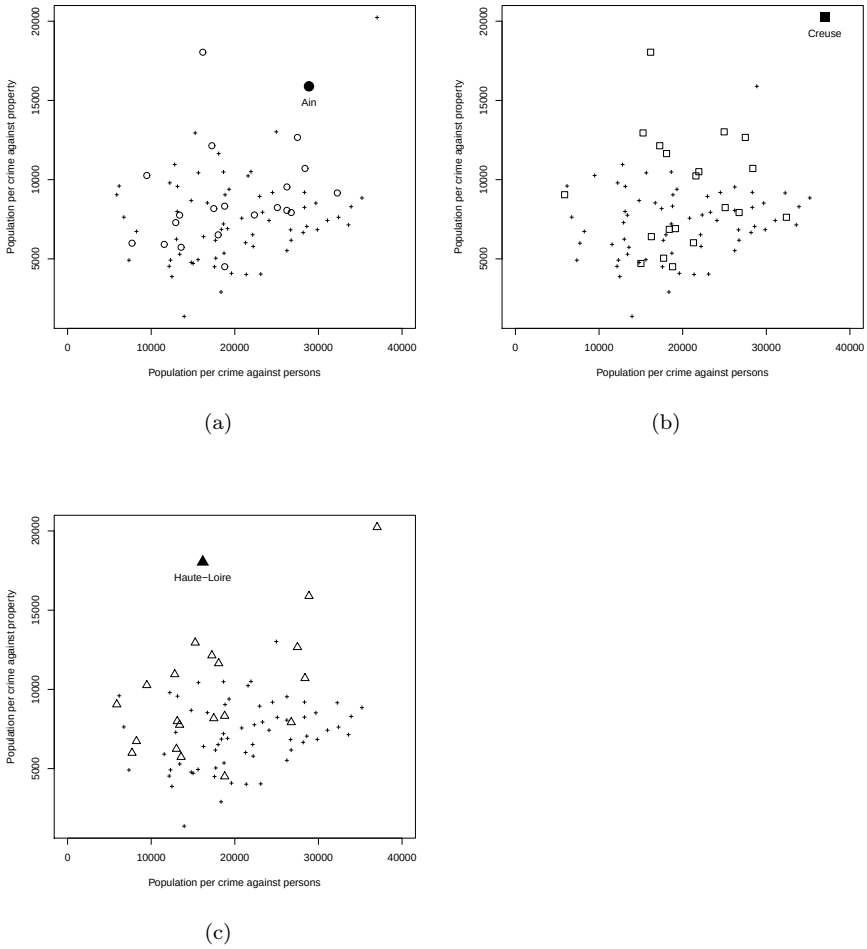
application of Harris et al.'s technique to cases where the neighbourhoods contain the same number of neighbours or are constructed by means of a given critical distance. Chen et al. (2008)'s technique was implemented in `R` as no public procedure could be found.

1.4.1 Social data in France

Dray and Jombart (2011) revisited social data measures on 85 departments in France in 1830. The data set is available in the `R` package `Guerry`. For illustrative reasons, only the two following variables are selected here: *population per crime against persons* and *population per crime against property*. Moreover, as in Filzmoser et al. (2014), the neighbourhoods are constructed as the set of the 20 closest neighbours.

On Figure 1.4 (showing the scatter plot of the two non-spatial attributes), the results obtained by the technique advocated by Chen et al. (2008) is illustrated. The three full symbols represent the three local outliers (Ain, Creuse and Haute-Loire) found by that technique. The corresponding empty symbols highlight their neighbours. There are two worth noting points: first, these outliers clearly lie far from the bulk of the data, implying also a global outlyingness (they are, together with the department of Correze, classified solely as "global" outliers by Filzmoser et al. (2014), as shown on Figure 1.6). Then, the important dispersion among the non-spatial attributes of the neighbours of the detected points questions the relevance of their local outlyingness. The technique of Harris et al. (2014) detects the same three outliers as well as the department of Correze (on Figure 1.5), for which the lack of homogeneity inside the neighbourhoods is again quite visible.

The technique of Filzmoser et al. is illustrated on Figure 1.6 and pinpoints a single local outlier: the Rhone department. Again, one may argue that Rhone is maybe not so outlying when one takes into account the dispersion inside its neighborhood. In fact, it turns out that the homogeneity is quite weak inside each neighbourhood, implying that the notion of local outlyingness is not so clear for that particular data set. Nevertheless, the regularized spatial detection technique detects only one local outlier (Loire Inférieure) when β is set to 0.25. Figure 1.7 (panel a) illustrates the relative homogeneity of the neighbourhood of the detected point while the adapted boxplots (panel b) show that the next distances of the selected point lie outside the outer fences of the box for that particular value of β. If β is set to 0.75 or 0.9, one or two other local outliers (Ain and Creuse) are also detected but their neighbourhoods undoubtedly are heterogeneous. Therefore, one needs to decide whether labelling these observations as *local* outliers is really appropriate. As explained in Section 1.3.2, the homogeneity measures are based on the computation of the mean of a subgroup of (uncontaminated) local distances, but the median of the local distances



(a) (b)

(c)

Figure 1.4: Social data: detection based on Chen et al. (2008). Representation of three "local" outliers and their neighbours. They are clearly global outliers and the neighbourhoods obviously lack homogeneity.

might have been used instead. Panel c of Figure 1.7 illustrates the results obtained for that other option. We can see that the two main local outliers are detected again but at slightly different levels of homogeneity (i.e., at different values of β).

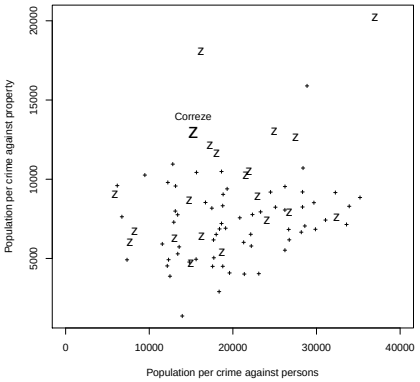


Figure 1.5: Observation classified as a local outlier by the technique of Harris et al. (2014) in addition to the three illustrated on Figure 1.4. The neighbourhood is quite heterogeneous and the observation is labelled as global outlier for Filzmoser et al. (2014).

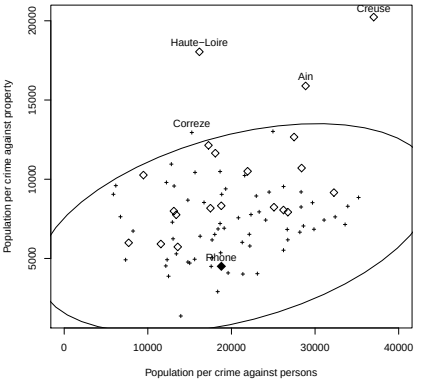
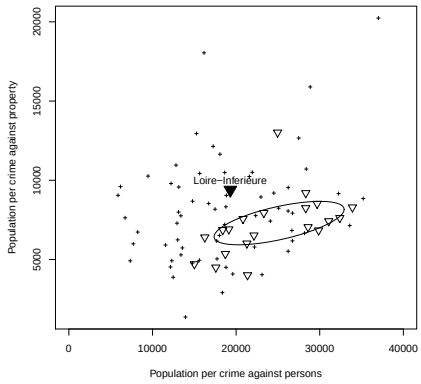


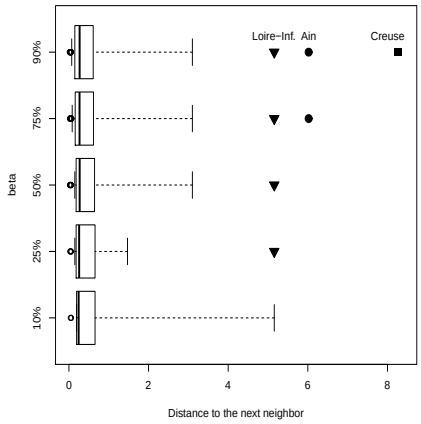
Figure 1.6: Detection of a local outlier (Rhône) and illustration of the global outliers for the technique of Filzmoser et al. (2014). The local outlyingness of this observation may not be relevant considering the dispersion inside its neighbourhood.

1.4.2 Geochemical data

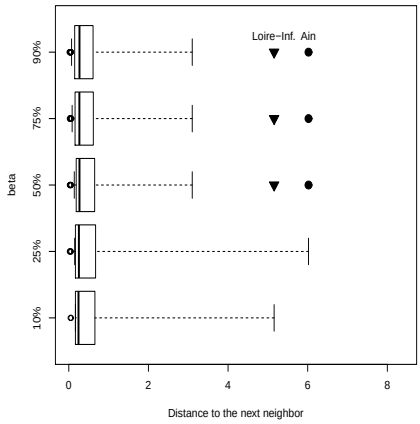
The Baltic Soil Survey data (BSS data, available in the `R` package `mvoutlier`) were collected in agricultural soils from Northern Europe (total area of about 1 800 000 km^2, 768 sampling sites taken on an irregular grid). Only the ten elements from the top layer (Al_2O_3, Fe_2O_3, K_2O, MgO, MnO, CaO, TiO_2, Na_2O, P_2O_5 and SiO_2) are considered here and, after the application of the isometric log-ratio transformation (as the data are compositional), this yields a dimension $p = 9$. The neighbourhoods are the same as those constructed by Filzmoser et al. (2014), i.e., they correspond to the sets of the 10 nearest neighbours. With such a small sample size inside the neighbourhoods, the technique of Harris et al. (2014) requires first the application of the global step based on a robust Principal Component Analysis, presented as optional in Section 1.2. Five principal components are kept through the detection process.



(a)



(b)



(c)

Figure 1.7: (a) The observation *Loire-Inférieure* is detected by the regularized spatial detection technique with $\beta = 0.25$. The boxplots of "next distances" are given for varying values of β when using the mean of the distances (plot on panel b) or the median (plot on panel c). There are slight differences according to the homogeneity measure but the same main outliers are detected.

Figure 1.8 (panel a), representing the different locations where the data were collected, summarizes the results of the different detection techniques. The most locally outlying spatial units detected by Chen et al. (2008) are plotted as full circles while crosses are used for Filzmoser et al. (2014) and empty squares represent the results of the detection of Harris et al. (2014). Based on the new approach, three local outliers (full triangles) are spotted. It is interesting to note that most outliers found by Chen et al. (2008), Harris et al. (2014) and Filzmoser et al. (2014) do not belong to the set $\mathcal{D}$ of spatial units lying in the most homogeneous areas and could not therefore be pointed out by the spatial regularized technique. Moreover, once again, the local outliers pinpointed by Chen et al. (2008) would be tagged as global by the full detection technique of Filzmoser et al. (2014).

(a) (b)

Figure 1.8: Detection of local outliers on geochemical data measures in Northern Europe. (a) Detected local outliers for each technique: full circle (Chen et al., 2008), crosses (Filzmoser et al., 2014), empty squares (Harris et al., 2014) and full triangles (regularized spatial detection technique). (b) The swapping of these two highlighted observations is entirely detected by the regularized spatial technique but not by the others.

To further analyze the performance of their detection techniques, Filzmoser et al. (2014) contaminated the data by exchanging the non-spatial attributes of two spatial units. Their contamination is not detected by Chen et al. (2008) nor by Harris et al.

(2014) and by the regularized spatial technique, as the considered neighbourhoods do not belong to the homogeneous set $\mathcal{D}$. However, swapping the observations obtained at the two locations highlighted on panel b of Figure 1.8 makes Filzmoser et al. (2014) detection partially fail (as it detects only one of the local outliers) while the regularized spatial technique works fine. The swapping of these two locations is based on a "contamination" technique advocated by Harris et al. (2014). A robust Principal Component Analysis is applied on the non-spatial attributes and the observations of $\mathcal{D}$ with the smallest score and the largest score on the first principal component are swapped. This contamination procedure is used again in the simulation study (see Section 1.5).

1.4.3 Cancer data in France

This data set contains five variables: the male lung cancer mortality rate (standardized over the age range 35–74 and over the 2-year period, 1968–1969), the cigarette sales and the percentages of employed males in specific types of industry (metal, mechanic and textile), the variables being measured at the scale of 82 French departments. These data come from Richardson et al. (1992). This time, following Richardson et al.'s way of proceeding, two departments are considered neighbours if they share a boundary. This implies that the spatial units get different numbers of neighbours (numbers ranging from one to eight). As the regularized detection technique requires at least three observations in order to estimate the local structure in the neighbourhoods, the nine departments having only one or two neighbours are neglected in the detection analysis, but they are kept when playing the role of neighbour for another spatial unit. Also, a second limitation in the study of this data set comes from the fact that Harris et al.'s procedure cannot be applied as currently implemented because the neighbourhoods have varying sizes and are not defined in terms of a critical distance. Therefore, only the three other techniques are considered.

These techniques detect different outlying departments as illustrated on the map of France in Figure 1.9: Vosges (which has six neighbours) and Aube (with five neighbours) for Chen et al. (2008); Bas-Rhin and Calvados (both having three neighbours) for Filzmoser et al. (2014); Nord ($n_i = 3$, $\beta \geq 0.25$), Hautes-Pyrénées ($n_i = 3$, $\beta \geq 0.25$), Indre-et-Loire ($n_i = 5$ and $\beta \geq 0.25$) and Tarn ($n_i = 5$, $\beta = 0.5$) for the regularized adaptation (recall that the consideration of increasing values of β allows the user to be less and less restrictive on the homogeneity of the neighbourhoods).

Marginal scatterplots (not shown) of the non-spatial attributes illustrate quite clearly that the neighbourhoods of Vosges, Aube and Bas-Rhin are not spatially homogeneous. It might be excessive to call them "local outliers". Moreover, the two first (i.e., those found by Chen et al. (2008)) belong again to the list of global outliers.

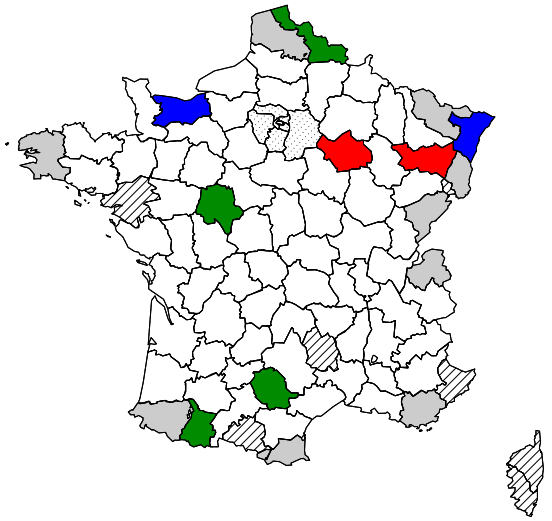


Figure 1.9: Map of France with outlying departments colored in red (Chen et al., 2008), blue (Filzmoser et al., 2014) or green (Regularized spatial technique). The grey-shaded areas represent those departments excluded from the detection procedure. In the data set, the hatched departments are not included and the dotted ones around Paris are aggregated.

1.4.4 Preliminary conclusion

As a preliminary conclusion, one might say that it is difficult, using real data, to put forward a detection technique which performs best. Indeed, it is not clear which observations are really locally outlying and most detected observations are different using one technique or another. The next section resorts to simulations to get a more objective comparison, local outliers being known in advance as they are inserted in clean data sets.

Nevertheless, most examples illustrate the fact that Chen et al. (2008)'s procedure mixes up global and local outliers. Another comment is the fact that the new regularized spatial detection technique gives additional insights on the outlyingness of the observations thanks to the possible consideration of several values of β. Using a fixed and unique value of β might be a bit too restrictive as one would then focus

only on a subsample of potential outliers (only on those lying in the most homogeneous neighbourhoods). Most local outliers found by the other techniques would not be found by the new procedure if restricted to the $\beta \times 100\%$ most homogeneous neighbourhoods for a fixed β. We recommend to vary the values of β and to keep in mind the corresponding interpretation. When using the other techniques, one needs to decide whether a spatial unit living in an unstable area should really be tagged as a *local* outlier. Finally, it is worth stressing once again that the first step of the regularized spatial detection technique is local, even if the sample size of the neighbourhoods is small (possibly smaller than the dimension) while the detection technique of Filzmoser et al. (2014) starts with a global estimation step (avoiding the local problem) and that of Harris et al. (2014) requires the application of an additional global step to handle data sets where the dimension is big.

1.5 Simulations

In this section, simulations are conducted in order to provide an objective comparison of the three detection techniques reviewed in Section 1.2. As additional information, the impact of the suggested adaptations presented in Section 1.3 is analysed. Harris et al. (2014) had already resorted to simulations, but with the sole objective of comparing variants (theoretical or empirical quantiles as cutoffs, different construction of the weight matrix inducing the neighbourhoods,...) of their own technique. Therefore, their simulation study is extended here to envelop the other detection techniques as well, while using only their default proposal instead of all their variants (i.e., theoretical quantiles are used as cutoffs for the measures SD and OD while empirical quantiles are computed for CS and the neighbourhoods are based on a given number of closest neighbours). In Section 1.4 devoted to the examples, the local outliers detected by Filzmoser et al.'s technique (as well as by its adaptation) are found by means of the visual analysis of some graphical displays. In simulations, this visual detection is no longer possible and needs to be automated. Filzmoser et al.'s procedure is automated as follows: when the isolation degree is three times bigger than the expected value (taken equal to $1/k$ where k is the number of neighbours), then the observation is tagged as a local outlier. For the regularized spatial detection technique, observing the changes for varying values of β and globalizing the detection is the best option. In the simulations, the search is decomposed into separate detections for the different fixed values of β. Taking a proportion β smaller than 0.4 might therefore be a bit too restrictive for the Gaussian process used in the simulations (as further explained later). Therefore, β is set to 0.4, 0.6, 0.8 and 1 and for each choice of β, a binary classification of the observations is derived. The observations in $\mathcal{D}$ having next distances outside the fence of the adjusted boxplot of

Hubert and Vandervieren (2008) are tagged as local outliers for that specific β.

To perform simulations in a spatial context, one needs to define the spatial units as well as a spatial correlation structure for the non-spatial attributes. Also, as outlier detection is the main interest here, contamination has to be introduced in the data. Finally, an objective way to measure the performance of the detection techniques should be defined. The choices made for these four aspects of the simulation study are further developed in the following subsections. Then, the results are outlined and discussed.

1.5.1 Spatial units

To mimick a practical situation, adapting Harris et al.'s idea to the Belgian setting, the first round of simulations is performed on spatial units consisting of the $n = 262$ municipalities of the Walloon region in Belgium (the municipalities, characterized by their longitude and latitude, are already illustrated in Figure 1.1 and may be visualized again on the different panels of Figure 1.10). In parallel, a more rigid configuration consisting of a 20×20-cell grid is considered for a second set of simulations. In both cases, neighbourhoods are constructed by means of the $n_i = k = [0.05 \times n]$ closest neighbours (even for the cells lying at the border of the square). Of course, in the grid case, the neighbourhoods show a more regular pattern than in the case of Wallonia.

1.5.2 Spatial correlation structure

The simulated data come from a p-dimensional Gaussian and second-order stationary process with mean vector zero and covariance matrix given by, following the notations of Gneiting et al. (2010),

$$C(h) = \begin{pmatrix} C_{11}(h) & \dots & C_{1p}(h) \\ \vdots & \ddots & \vdots \\ C_{p1}(h) & \dots & C_{pp}(h) \end{pmatrix}$$

with

$$C_{ii}(h) = \sigma_i^2 M(h|\nu, a),\ i = 1, \dots, p$$

and

$$C_{ij}(h) = \rho_{ij}\sigma_i\sigma_j M(h|\nu, a),\ i, j = 1, \dots, p\ (i \neq j)$$

where $M(h|\nu, a)$ is the spatial correlation at a distance h based on the Matérn function. The restriction to a constant spatial scale parameter a and a constant smoothness parameter ν yields the so-called *parsimonious* multivariate Matérn model, as already used by Harris et al. (2014). Consistently with the purpose of extending

their results, similar values for the tuning parameters a and ν and for the elements of the cross-covariance matrix, $\rho_{ij}, 1 \leq i, j \leq p$ and $\sigma_i, 1 \leq i \leq p$, are chosen. First, the spatial scale parameter is set to 1. Then, for $p = 5$, the following cross-covariances and variances is used:

$$\Sigma = \begin{pmatrix} 70 & 60 & 60 & 65 & 65 \\ 60 & 90 & 75 & 70 & 70 \\ 60 & 75 & 95 & 60 & 55 \\ 65 & 70 & 60 & 75 & 60 \\ 65 & 70 & 55 & 60 & 85 \end{pmatrix}$$

where $\Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$ with $\rho_{ii} = 1 \ \forall \ i$. For smaller values of the dimension p, the upper-left $p \times p$ square sub-matrix of Σ is chosen.

As far as the smoothness parameter is concerned, note first that larger values correspond to smoother variations. The choice of ν is illustrated in the bivariate case in Figure 1.10 using the real spatial locations of the Walloon municipalities. On the two upper panels, ν is set to 0.5, while it is equal to 2.5 (value suggested in Harris et al., 2014) on the lower panels. Both choices are further considered in the simulation study as, to our point of view, they provide different homogeneity patterns even though, for each case separately, the neighbourhoods have comparable homogeneity. Under such a scheme, restricting the detection to a small number of the most homogeneous neighbourhoods is counterproductive. This explains why β is set to values above 40% in the simulations.

1.5.3 Contamination set-up

Simulating the data by means of the Gaussian and second-order stationary process detailed above should yield data free of local outliers. Global outliers are possible though, but these are not under consideration here (unless they are local at the same time). As introduced in the example of Subsection 1.4.2, the contamination process defined in Harris et al. (2014) is used for the simulations. In order to reach a given percentage of local contamination, 5% say, more care needs to be given to the construction of these local outliers. Indeed, as already stressed by Harris et al. (2014), it may happen that, as suggested, the contamination method ends up with the swap of a bunch of neighbours. If they have close attributes in the beginning and are swapped all together, they keep similar values and cannot be considered local outliers, while the contamination procedure labels them as such. If that unfortunate swapping happens too often, the detection techniques is considered inefficient for finding the local outliers, while there are none to find. This potential problem is illustrated on Figure 1.11 using the grid case in two dimensions. On panel a, the clean case is represented while the 5%-contaminated configuration can be looked at

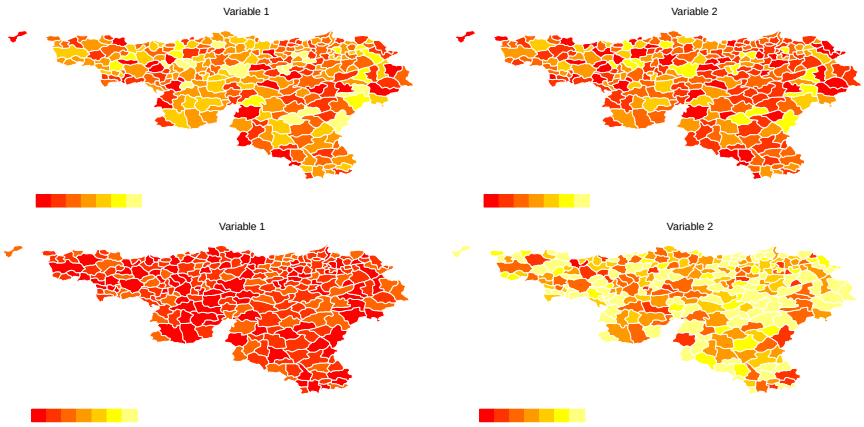


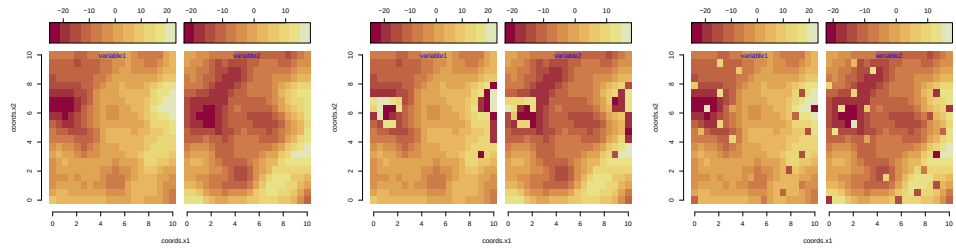
Figure 1.10: Gaussian and second-order stationary process based on the parsimonious Matérn model for $\nu = 0.5$ (upper panels) and $\nu = 2.5$ (lower panels) to illustrate different homogeneity patterns.

on panel b. Clearly, patches of neighbours are swapped without introducing the expected 5% of local contamination in the data.

Harris et al. (2014) solve this problem by not focusing on the highest and lowest scores. Here, another way of choosing the units that will be swapped has been designed. Again, the highest scores correspond to potential candidates to swap with the smallest ones, but the selection proceeds one at a time and discards any spatial unit lying in the neighbourhood of another one which was previously selected. For the clean set-up illustrated on Figure 1.11 (panel a), this adaptation of Harris et al.'s proposal yields the contaminated configuration on panel c, where the patches of outliers have disappeared. Unlike in the geochemical application where the contamination step is restricted to units lying in $\mathcal{D}$, the local outlier inserted in the simulated data sets might lie in more heterogeneous neighbourhoods.

1.5.4 Performance measure

To measure the performance of a local outlier detection technique, misclassification error rates may be computed. A good detection technique should not only detect the local outliers, but also avoid to falsely detect good observations as local outliers. The concepts of "false positive" (i.e., a regular observation classified as a local outlier)



(a) (b) (c)

Figure 1.11: Clean (panel a) and 5%-contaminated (panels b and c) 2-dimensional Gaussian process on the grid; (b) the contamination procedure is unrestricted and (c) the contamination has the additional constraint of discarding neighbours.

and "false negative" (i.e., a local outlier which remains undetected) are informative for that matter. Table 1.1, built following Cerioli and Farcomeni (2011)'s notations, summarizes the possible combinations of the outcomes of the detection technique (regular/outlier) with respect to the real category of the data points.

Table 1.1: Contingency table of the real category of the observations with respect to the classification resulting from the application of an outlier detection technique.

Reality	Classified as Regular	Outlier	Total
Regular	TN	FP	M_0
Outlier	FN	TP	M_1
Total	$n-R$	R	n

Several summary measures based on the false positive and false negative error rates are listed in Cerioli and Farcomeni (2011). Here, referring to the notations used in Table 1.1, the performance measures that is computed over the simulations are the *false positive error rate* defined by FP/M_0, the *false negative error rate* given by FN/M_1, as well as the agreement measure given by Cohen's Kappa statistics, as advocated by Harris et al. (2014). Ideally, the two rates should be equal to 0, while the Kappa statistics should be close to 1. Departure from 0 for the FN error rate indicates a *swamping effect* and departure from 0 for the FP error rate is the sign of a *masking effect*. Kappa statistics focuses on the other hand on the capacity of the detection technique to correctly identify the good and the bad observations.

1.5.5 Results

For $p = 2$ and $p = 5$, 500 data sets were simulated according to the spatial model detailed above, using both the regular grid and Wallonia as spatial domains and using the two specified values for the smoothness parameter ($\nu = 0.5$ and $\nu = 2.5$). The three techniques, as well as the adaptation of Filzmoser et al.'s technique (via four independent applications in order to compare the different β values), were applied to each simulated data set and the three summary measures were recorded for each as well.

Figure 1.12 shows, using boxplots, the results of the computation on the 500 simulated data sets of the false positive, false negative error rates and of the Kappa measures, under the 2D configuration (using the two spatial domains and for the two ν values), while Figure 1.13 yields the same results for the 5D set-up. Tables 1.2 (for $p = 2$) and 1.3 (for $p = 5$) provides the averages of these three summary measures computed over the 500 runs.

Let us first look at the results concerning the false positive error rates. One can see that, whatever the dimension, the technique of Harris et al. (2014) wrongly flags too many good observations as local outliers; the percentage of false positives is only approximately under control when the regular grid with $\nu = 0.5$ is exploited. In the three other settings, the average false positive error rates are well above 50% and the boxplots lie in the upper part of the figures. Chen et al. (2008)'s technique does a bit better, even though it does not handle so well the Wallonia spatial domain with $\nu = 2.5$. The two remaining techniques perform well on that criterion whatever the dimension. It is interesting to note (but not surprising) that there is not any effect of the choice of β on that summary measure (except when all neighbourhoods are searched, in which case the false positive error rate increases a bit).

As far as the false negative error rates are concerned, Chen et al. (2008)'s procedure yields the best results. Its boxplots lie below the others and may even be completely degenerated at 0 in some configurations (meaning that all local outliers are found). The other techniques seem to partially suffer from some masking effect as not all local outliers are detected on average. The detection method of Harris et al. (2014) corresponds to the second best option on that criterion, while the regularized spatial detection technique discarding at least 40% of the neighbourhoods is clearly the worst, which is expected as the contamination was not restricted to the most homogeneous neighbourhoods. Nevertheless, taking $\beta = 0.8$ provides more protection against the masking effect and ends up with comparable results with respect to Filzmoser et al. (2014)'s technique. This large value of β can be justified by the strong homogeneity of all neighbourhoods obtained by the Gaussian process. Therefore, working only on a small proportion of neighbourhoods (which might be illuminating in real data analysis) is too restrictive in this simulated set-up.

Turning now to the Kappa statistics, one notes that the boxplots based on the method of Harris et al. (2014) lie well under the other boxplots, due to the lack of correct identification of the good observations. Chen et al. (2008) procedure corresponds to Kappa values mainly above 0.5, except under the specific configuration pointed out before (i.e., Wallonia with smoothness parameter 2.5). It even provides the highest association measures when the spatial domain is the regular grid with smoothness parameter 0.5. In most configurations (except the regular grid with $\nu = 0.5$), one can see that Filzmoser et al. (2014)'s technique outperforms the techniques of Harris et al. (2014) and Chen et al. (2008) as well as the regularized spatial method with $\beta = 0.4$. Now, as far as the latter technique is concerned, there is clearly an improvement in this overall measure of association when β increases. A value $\beta = 0.8$ seems to provide, for the considered simulation schemes, the best compromise: it allows to keep low FP rates, while limiting the FN rates and reaching reasonably high Kappa values, close to the best performing technique (Filzmoser et al. (2014) or Chen et al. (2008) depending on the configuration). The overall good performance of the technique of Filzmoser et al. (2014) may be partly explained by the configurations used in the simulation study. The application of a global estimation step based on the MCD estimator when the data are generated according to a second-order stationary Gaussian process is quite appropriate.

1.6 Conclusion

This chapter provides a review of the techniques allowing to detect local outliers in the spatial context using multivariate non-spatial attributes. It can be seen as a complement to Schubert et al. (2014)'s paper and in that spirit, the same approach is followed, i.e., each method is described by means of context and comparison sets and functions. Even if real data examples, and even sometimes simulations, have already been used in the literature to illustrate the performance of these known detection techniques, these attempts mainly focused on one of the proposals at a time and not on all of them simultaneously. An objective comparison of these techniques using the same data as illustrative examples and using the same simulation study was one of the objectives of this work.

As an additional objective, an adaptation of the technique of Filzmoser et al. (2014) is suggested. While the original detection method is appealing and easy to apply, its initial and global estimation step prevents it from being fully local. Also, the notion of local outlyingness is not so clear and a possible restriction of the search to the observations belonging to the most concentrated neighbourhoods is discussed through the chapter.

As a conclusion, one can say that it is difficult to determine the "best" detec-

tion technique when using real data. The truth concerning the observations that are really local outliers is not known and most methods detect different observations. Nevertheless, the results of the techniques may be interpreted as they are based on well-defined contexts and comparison functions. In our opinion, it is appropriate, when dealing with real data, to think about the adequacy of tagging as local outlier an observation whose neighbourhood does not show any sign of homogeneity. The regularized spatial detection technique that we constructed based on Filzmoser et al. (2014)'s proposal allows to restrict the search to the most homogeneous neighbourhoods. A thorough exploratory analysis of the data, linked to the results displayed by the adjusted boxplots of the next distances, provides efficient ways to detect local outliers.

Outlier detection is a broad field which continues to attract a lot of attention in several fields. This topic has generated a wide range of publications since 2016, last possible year of the publications included in the review considered in this chapter (year of the submission of the paper Ernst and Haesbroeck, 2017). To fill in the gap between 2016 and 2019, we briefly highlight hereafter the most relevant publications that were published during that period in the general context of spatial data and in data mining.

First, some of the techniques that we present in this chapter have been improved or modified later on. More specifically, the use of a spatially weighted principal component analysis to detect spatial outliers, as described in Harris et al. (2014, 2015) is still under study. For instance, Lin (2019) proposes a partial geographically weighted PCA with globally standardized data and geographically weighted spatial association to carry the detection of multivariate spatial outliers. Singh and Lalitha (2018) construct an algorithm similar to Chen et al. (2008) which is based on a so-called "location quotient" in place of the difference between the observation and the vector containing the marginal medians over its neighbourhood.

More globally, Zimek and Filzmoser (2018) provide a general overview of outlier detection techniques from the data mining and the statistical point of view. Moreover, they present the open-source data mining framework (ELKI, introduced in Schubert et al., 2015) which contains, among others, the algorithms of many standard methods for outlier detection. On the other hand, Rottoli et al. (2018) reference detection techniques adapted to the spatial context. To deal with a massive amount of possible algorithms, they introduce a *knowledge discovery process* for detection of local spatial outliers. The goal is to separate the different steps in a standardized way in order to allow the use of specific algorithms for each step.

In the data mining field, we point out some papers which deal with spatial outlier detection. First, Schubert et al. (2015) follow the general framework that we used

in this chapter (Schubert et al., 2014) and apply it to streaming data. A second paper, Kamble and Doke (2017), reviews and compares several outlier detection approaches applied to data mining. Finally, Sijin et al. (2017) focus on the particular case of keyword search. They compare the advantages and disadvantages of several techniques. These techniques use notably antihub algorithms, angle based outlier detections and the reverse nearest neighbour search to detect spatial local outliers in the keyword search.

Other research papers focusing on a specific application field are also doveted to the development of techniques for the detection of local outliers. It is not possible to list them all but the domain of flow probability distributions has attracted our attention. Y. Djenouri and his co-authors propose to model the sequence of traffic flow sequences as probability distributions of flows. They are interested in detecting local outliers in the distributions. First, Djenouri and Zimek (2018) summarise outlier detection techniques available for urban traffic data and the paper Djenouri et al. (2019) is a survey on outlier detection algorithms (including our proposed technique). Then, Djenouri et al. (2018) propose an adapted framework to detect flow probability distribution outliers. Their conclusion seems quite natural: their technique outperforms the other ones by considering multivariate distributions (i.e., the correlation structure between the different flows) instead of dealing with univariate settings for each flow separately.

Lastly, several other fields may rely on spatial outlier detection in practice. We may cite for instance O'Leary et al. (2016) for an application to urban air quality (even if they mention detection techniques adapted to the multivariate setting, only univariate tools are used for the detection); Petri (2017) which applies detection on bio-molecular data (comparison of the methods to detect correlation pattern in RNA measurements).

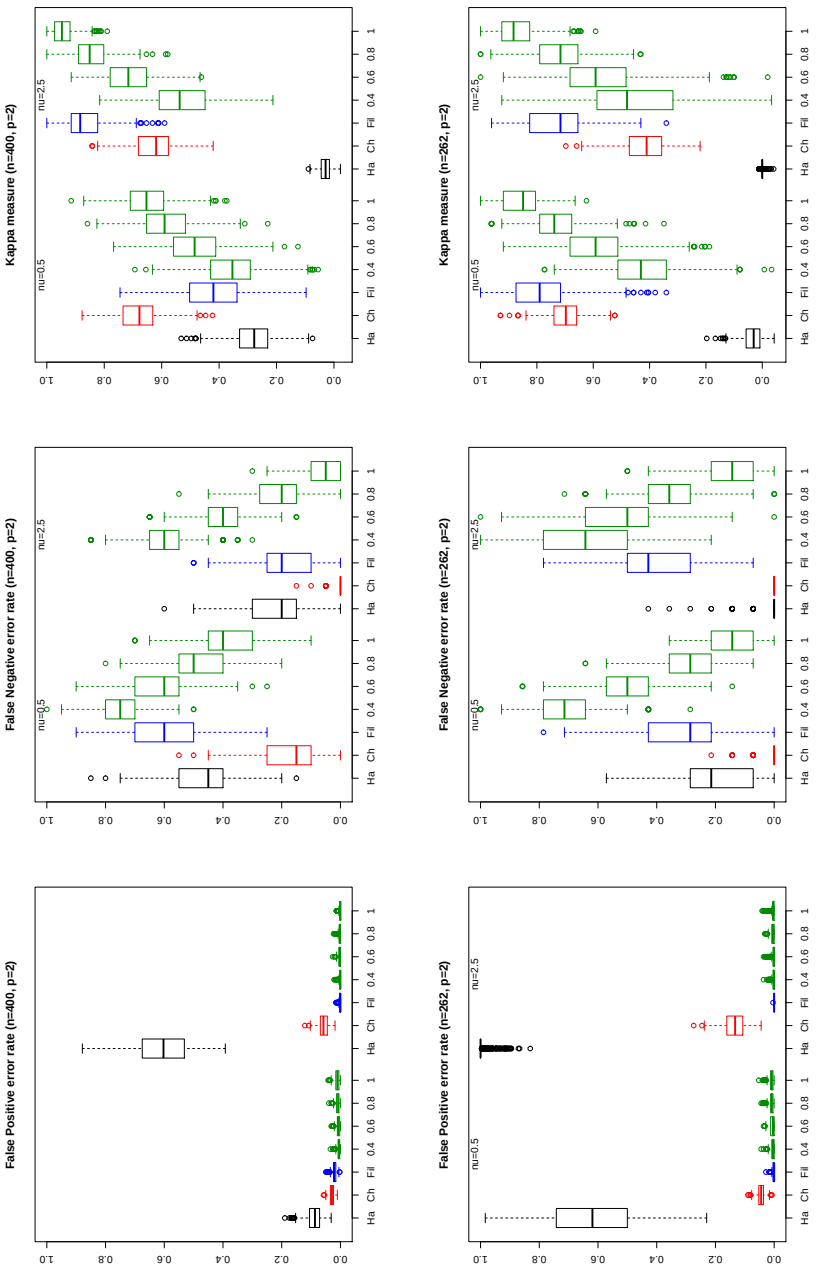


Figure 1.12: False positive (left panels), false negative error rates (middle panels) and Kappa statistics (right panels) for the 500 simulations with $p = 2$ tested by the four different techniques (**Harris et al., 2014** (Ha), **Chen et al., 2008** (Ch), **Filzmoser et al., 2014** (Fil) and **the regularized spatial technique** which is tested for $\beta = 0.4$, 0.6, 0.8 and 1). The upper panels correspond to the regular grid ($n = 400$) while the lower ones are based on the Walloon municipalities ($n = 262$). In each case, two values of ν are tested (0.5 and 2.5).

Table 1.2: Means of false positive, false negative error rates and the Kappa measures for 500 bivariate simulations.

			Harris et al. (2014)	Chen et al. (2008)	Filzmoser et al. (2014)	
Grid	$\nu = 0.5$	FP	0.090	0.029	0.021	
		FN	0.470	0.170	0.592	
		Kappa	0.282	0.678	0.423	
	$\nu = 2.5$	FP	0.604	0.058	0.001	
		FN	0.215	0.002	0.197	
		Kappa	0.030	0.627	0.870	
Wallonia	$\nu = 0.5$	FP	0.620	0.044	0.003	
		FN	0.204	0.007	0.306	
		Kappa	0.035	0.700	0.778	
	$\nu = 2.5$	FP	0.991	0.135	8×10^{-6}	
		FN	0.009	0	0.394	
		Kappa	0	0.420	0.737	
			Regularized spatial technique			
			$\beta = 0.4$	$\beta = 0.6$	$\beta = 0.8$	$\beta = 1$
Grid	$\nu = 0.5$	FP	0.006	0.006	0.006	0.012
		FN	0.739	0.612	0.488	0.385
		Kappa	0.359	0.480	0.583	0.650
	$\nu = 2.5$	FP	0.002	0.002	0.002	0.002
		FN	0.606	0.405	0.220	0.062
		Kappa	0.530	0.709	0.844	0.948
Wallonia	$\nu = 0.5$	FP	0.005	0.005	0.005	0.010
		FN	0.682	0.502	0.308	0.120
		Kappa	0.425	0.591	0.736	0.852
	$\nu = 2.5$	FP	0.003	0.003	0.003	0.004
		FN	0.669	0.532	0.366	0.153
		Kappa	0.443	0.581	0.724	0.873

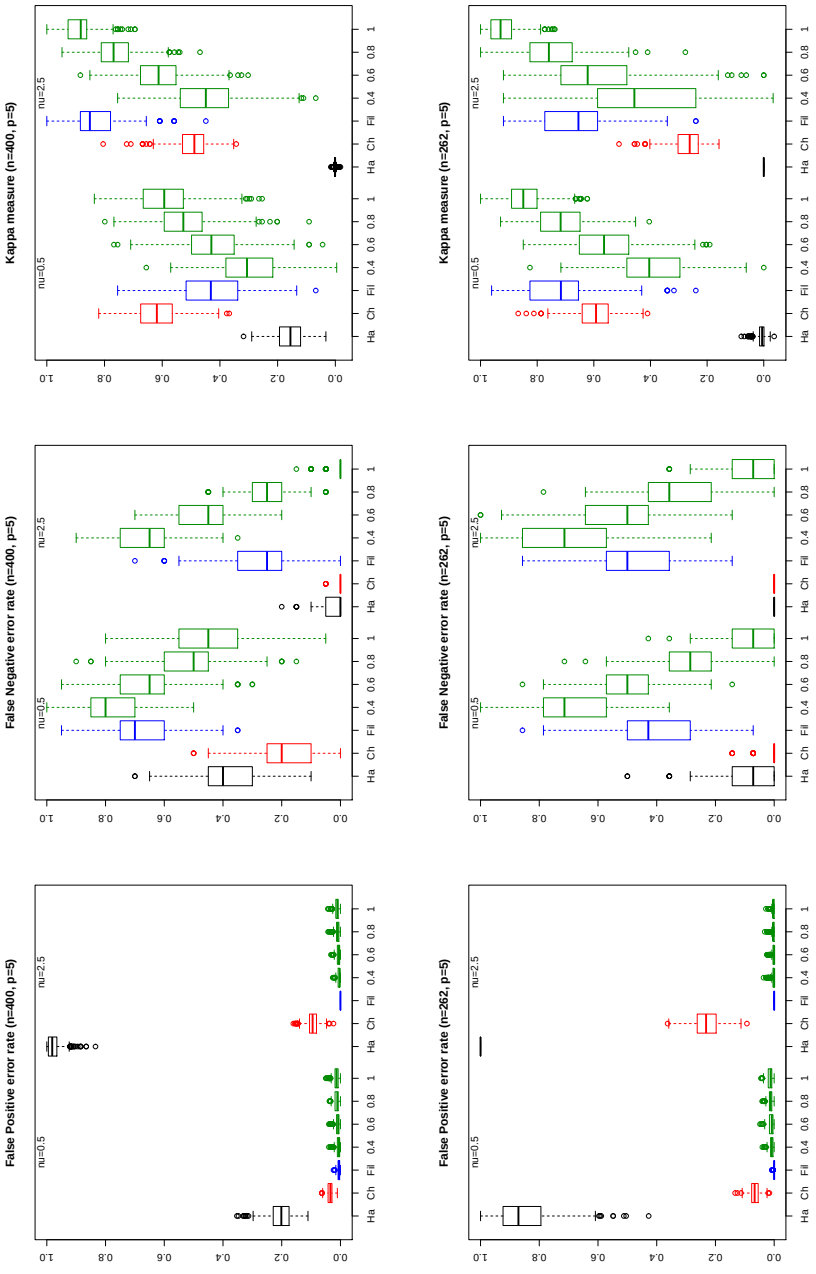


Figure 1.13: False positive (left panels), false negative error rates (middle panels) and Kappa statistics (right panels) for the 500 simulations with $p = 5$ tested by the four different techniques (**Harris et al., 2014** (Ha), **Chen et al., 2008** (Ch), **Filzmoser et al., 2014** (Fil) and **the regularized spatial technique** which is tested for $\beta = 0.4$, 0.6, 0.8 and 1). The upper panels correspond to the regular grid ($n = 400$) while the lower ones are based on the Walloon municipalities ($n = 262$). In each case, two values of ν are tested (0.5 and 2.5).

Table 1.3: Means of false positive, false negative error rates and the Kappa measures for 500 five-dimensional simulations.

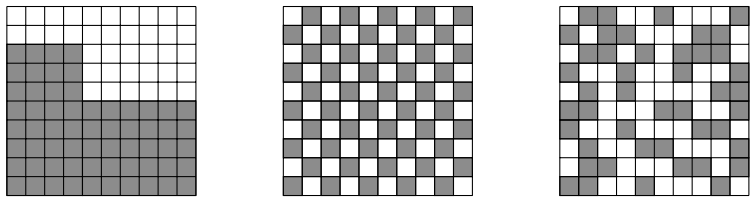
			Harris et al. (2014)	Chen et al. (2008)	Filzmoser et al. (2014)	
Grid	$\nu = 0.5$	FP	0.203	0.036	0.006	
		FN	0.391	0.207	0.675	
		Kappa	0.157	0.618	0.429	
	$\nu = 2.5$	FP	0.976	0.094	0	
		FN	0.017	7×10^{-4}	0.279	
		Kappa	0.001	0.497	0.825	
Wallonia	$\nu = 0.5$	FP	0.851	0.066	2×10^{-4}	
		FN	0.090	0.005	0.416	
		Kappa	0.008	0.604	0.715	
	$\nu = 2.5$	FP	1	0.232	0	
		FN	0	0	0.482	
		Kappa	0	0.269	0.661	
			Regularized spatial technique			
			$\beta = 0.4$	$\beta = 0.6$	$\beta = 0.8$	$\beta = 1$
Grid	$\nu = 0.5$	FP	0.007	0.007	0.007	0.015
		FN	0.781	0.657	0.534	0.435
		Kappa	0.299	0.422	0.524	0.590
	$\nu = 2.5$	FP	0.005	0.005	0.005	0.012
		FN	0.665	0.470	0.247	0.007
		Kappa	0.450	0.617	0.764	0.891
Wallonia	$\nu = 0.5$	FP	0.008	0.008	0.008	0.014
		FN	0.706	0.520	0.307	0.076
		Kappa	0.386	0.555	0.711	0.842
	$\nu = 2.5$	FP	0.003	0.003	0.003	0.004
		FN	0.682	0.527	0.339	0.074
		Kappa	0.430	0.589	0.749	0.926

CHAPTER 2

Spatial autocorrelation: robustness of tests and robust alternatives

2.1 Introduction

The data studied in the geographical sector are usually attached to geographical locations, and these spatial positions usually provide additional information about the underlying analyses. In order to determine if the spatial pattern indeed contains significant information about the data, the concept of spatial autocorrelation is defined. A variable has positive (resp. negative) spatial autocorrelation if close observations have a similar (resp. antagonistic) behaviour on that variable. On the other hand, if no spatial influence is measured, one says that the variable has no spatial autocorrelation. Figure 2.1 is a classic illustration of the three cases as mentioned in the literature dealing with GIS, *geographic information systems* (see for instance Campbell and Shin, 2011, Worboys and Duckham, 2004).



Positive spatial autocorrelation | Negative spatial autocorrelation | No spatial autocorrelation

Figure 2.1: Different schemes for the spatial autocorrelation.

In practice, geographers estimate the spatial autocorrelation using mostly Moran's index I (Moran, 1950), Geary's ratio c (Geary, 1954) or the general index G (Getis and Ord, 1992).

Testing if the data are sufficiently spatially autocorrelated is a current practice in health geography (see Osei and Duker, 2008, Ibrahim et al., 2015, Havard et al., 2009), economics (e.g. Altay and Celebioglu, 2015, Melecky, 2015, Osland et al., 2016 and Wu et al., 2019), biogeography (for instance Fu et al., 2014), geographic profile (see Quick et al., 2019), ...

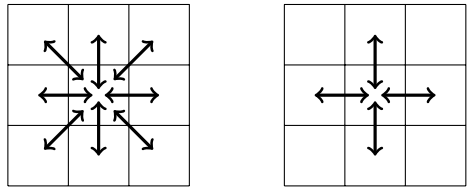
In Section 2.2, the definitions of the three spatial autocorrelation measures are outlined, while Section 2.3 derives the corresponding tests of autocorrelation. The robust issue is developed in Section 2.4, where the lack of robustness is flagrant: a unique contamination can totally change the conclusion of an autocorrelation test. In order to be concise, we consider only Moran's index from then on, the results based on Geary's ratio and Getis and Ord's statistic being available in the Appendix. Then, robust alternatives based on Moran's index I are proposed in Section 2.5 in order to deal with spatial outliers. Finally, a simulation study allows a thorough comparison of the different techniques in Section 2.6. Throughout this chapter, a real dataset is used for illustration. It is for example shown that the robust tests detect positive spatial autocorrelation in these data while the classic tests fail to do so.

2.2 Spatial autocorrelation indexes

Let us consider a spatial process $\{Z(s_i) : s_i \in D\}$ over a fixed and discrete domain D (i.e., we consider *raster* data using GIS terminology, see Worboys and Duckham, 2004). The sample data points are denoted by $\{z_1, \ldots, z_n\}$ and the corresponding spatial locations $\{s_1, \ldots, s_n\}$. In order to measure the autocorrelation, one needs to define the neighbourhood of a spatial location. Different strategies are possible and one usually resorts to the construction of a weighting matrix to formalize the matter. A not necessarily symmetric weighting matrix $W = (w_{ij})_{1 \leq i,j \leq n}$, with zero diagonal, is used here to describe spatial neighbours.

As illustration, two spatial contexts are considered in this chapter: a regular grid $a \times a$ and the irregular domain of Belgian municipalities. The weighting matrix associated with a regular grid is defined using the queen contiguity (neighbouring regions share at least a common point) or the rook contiguity (border of non-zero length). Figure 2.2 is a classic illustration of these contiguity cases (see for instance Figure 1 in Holmberg and Lundevaller, 2015). The weighting matrix associated with the irregular domain is defined by the adjacency matrix based on the 589 Belgian municipalities, i.e., municipalities are neighbours when they share common bound-

aries. More specifically, W is constructed as follows: a value 1 is indicated at the (i, j) position if location s_j is neighbouring s_i, 0 otherwise. These weighting matrices are binary and symmetric. Row-standardized matrices could also be considered. By construction, in the queen configuration, each cell has 3, 5 or 8 neighbours while the size of neighbourhoods is restricted to 2, 3 or 4 neighbours for the rook contiguity. Moreover, the Belgian municipalities have 1 to 16 neighbours.



Queen contiguity Rook contiguity

Figure 2.2: An arrow indicates if two areas are contiguous.

In this chapter, we consider the following notations introduced in Cliff and Ord (1973): the row and column weights $w_{i\bullet} = \sum_{j=1}^{n} w_{ij}$ and $w_{\bullet i} = \sum_{k=1}^{n} w_{ki}$, and the sums of weights $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, $S_1 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(w_{ij} + w_{ji})$ and $S_2 = \sum_{i=1}^{n}(w_{i\bullet} + w_{\bullet i})^2$.

As explained in the introduction, we consider three usual measures of spatial autocorrelation. They are here defined empirically.

Moran's index is a global indicator of spatial autocorrelation and is defined by

$$I(\{z_1, \ldots, z_n\}) = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}. \tag{2.1}$$

where $\bar{z}$ is the sample mean of $\{z_1, \ldots, z_n\}$.

Geary's ratio is based on comparisons between pairs of observations and is determined by

$$c(\{z_1, \ldots, z_n\}) = \frac{n-1}{2S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - z_j)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}. \tag{2.2}$$

The *general Getis and Ord's statistic* measures the concentration or lack of concentration of the sum of values for a positive variable. More precisely, for $z_1, \ldots, z_n \geq 0$,

$$G(\{z_1, \ldots, z_n\}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} z_i z_j}. \tag{2.3}$$

Moran's index is structured so that, under no spatial autocorrelation, the mean value is $-1/(n-1)$. Values of Moran's I larger than $-1/(n-1)$ indicate positive

spatial autocorrelation and values smaller than $-1/(n-1)$ indicate negative spatial autocorrelation (Cliff and Ord, 1973). Geary's ratio has a mean value equal to 1 without spatial autocorrelation and it behaves in the opposite way than Moran's index: a smaller value of c corresponds to positive spatial autocorrelation and a large value of c corresponds to negative spatial autocorrelation (Cliff and Ord, 1973). Moreover, both indexes are affine invariant and to allow the interpretation of the observed values, de Jong et al. (1984) determined their extreme values for a fixed weighting matrix. These extreme values are given by the largest and smallest eigenvalues of a matrix $\widetilde{W}$ defined with respect to the initial weighting matrix W. Getis and Ord's statistic is used to measure the degree of clustering for either high values or low values. Unlike the two other indexes, Getis and Ord's statistic is only scale-invariant and not location-invariant. If the weighting matrix is binary, G statistic ranges from 0 to 1. A value near the mean value $S_0/n(n-1)$ indicates no apparent clustering within the study area, whereas a larger and a lower value indicates clustering of high and low values, respectively. Unfortunately, in presence of both, high and low clusters, the G statistic is not able to detect the spatial aggregation. Therefore, Moran's index or Geary's ratio is more adequate in this situation.

These measures can also be written as ratios of quadratic forms. This notation will be useful, for instance in Section 2.4. Let $\mathbf{z}$ be the n-dimensional column vector of observed values $(z_1, \ldots, z_n)$. Then, the indexes defined in (2.1), (2.2) and (2.3) may be written as:

$$I(\mathbf{z}) = \frac{n}{S_0} \frac{\mathbf{z}'HWH\mathbf{z}}{\mathbf{z}'H\mathbf{z}}, \tag{2.4}$$

$$c(\mathbf{z}) = \frac{n-1}{S_0} \frac{\mathbf{z}'(N-W)\mathbf{z}}{\mathbf{z}'H\mathbf{z}} \tag{2.5}$$

and

$$G(\mathbf{z}) = \frac{\mathbf{z}'W\mathbf{z}}{\mathbf{z}'B\mathbf{z}} \tag{2.6}$$

where N is a diagonal matrix based on W and defined by $N_{ii} = (w_{i\bullet} + w_{\bullet i})/2$; the centring matrix $H = I_n - (1/n)\mathbb{1}\mathbb{1}'$ is an idempotent matrix, $\mathbb{1}$ is the column-vector of n ones, and $B = \mathbb{1}\mathbb{1}' - I_n$ is the square matrix with ones everywhere except on the diagonal which is null.

One straightly observes that, as mentioned in Genton and Ruiz-Gazen (2010), if the weighting matrix is not symmetric, $(W + W')/2$ is symmetric and using $(W + W')/2$ as weighting matrix yields the same value for Moran's index. The same comment holds for the two other indexes. Therefore from now onward, without loss of generality, one can assume that W is symmetric. Despite this simplification, the choice of the definition for the weighting matrix is essential as different matrices can lead to different values of the indexes. For instance, if the number of neighbours is

not constant, the binary and the row-standardized matrices do not give the same values for the indexes. The effect of the weighting matrix on Moran's tests applied on residuals is studied in Anselin and Rey (1991).

The distribution of these indexes were studied in the literature in order to define inference tools on spatial autocorrelation.Two hypothetical cases are studied as cited in Cliff and Ord (1973). The first assumption (noted N for normality) is that the observations $z_1, \ldots, z_n$ are the results of n independent drawings from a normal population. The second hypothesis (noted R for randomisation) does not depend on any underlying distribution. One considers a set of fixed values $\{z_1, \ldots, z_n\}$ which are randomly permuted on the locations $\{s_1, \ldots, s_n\}$ (which gives $n!$ possibilities).

The two first moments of the three indexes are known and given in Table 2.1. Cliff and Ord (1973, p 15-16) proved them for Moran's index and Geary's ratio under both assumptions and Getis and Ord (1992, 1993) did it for their G statistic under the randomisation assumption. We may directly observe that the mean values of the three indexes do not depend on the vector $\mathbf{z}$. Moreover, under normality, the variance of the indexes is also independent on $\mathbf{z}$. The initial vector $\mathbf{z}$ influences only the variances under randomisation by means of its kurtosis $b_2(\mathbf{z})$ for Moran's and Geary's indexes and by means of the three first moments for Getis and Ord's statistic.

Sufficient conditions for the asymptotic normality of I and c are proved under both assumptions (Cliff and Ord, 1973, Sen, 1976). Zhang (2008) proves the asymptotic normality of G under the randomisation assumption, as conjectured in Getis and Ord (1992).

More sophisticated models are also developed in the literature. For instance, Maruyama (2015) proposed a linear transformation of Moran's index in order to obtain values between -1 and 1. Moran's distribution was adapted to linear regression models in Cliff and Ord (1972) (see for instance an application in Richardson et al., 1992) and Kelejian and Prucha (2001) established the limiting distribution of Moran's I for various dependent variable models. Another example is the method of Principal Coordinates analysis of neighbour matrices (PCNM) (or Moran's eigenvector maps) which can also be defined by Moran's index (see for instance Dray et al., 2006 or Murakami and Griffith, 2019). The properties discussed in this chapter can easily be extended to these particular features. Nevertheless, our work focuses only on the general case as presented hereafter.

2.3 Tests for spatial autocorrelation

As mentioned earlier, spatial autocorrelation inference is currently applied in several fields (health geography, economics, biogeography, ...). The null hypothesis

$\mathbb{E}_R[I] =$	$\mathbb{E}_N[I] = \frac{-1}{n-1}$
$V_N[I] =$	$\frac{3S_0^2+n^2S_1-nS_2}{S_0^2(n-1)(n+1)} - \frac{1}{(n-1)^2}$
$V_R[I(\mathbf{z})] =$	$\frac{n[(n^2-3n+3)S_1-nS_2+3S_0^2]-b_2(\mathbf{z})[(n^2-n)S_1-2nS_2+6S_0^2]}{(n-1)(n-2)(n-3)S_0^2} - \frac{1}{(n-1)^2}$

(Cliff and Ord, 1973, p 15)

$\mathbb{E}_R[c] =$	$\mathbb{E}_N[c] = 1$
$V_N[c] =$	$\frac{(2S_1+S_2)(n-1)-4S_0^2}{2(n+1)S_0^2}$
$V_R[c(\mathbf{z})] =$	$\frac{(n-1)(n^2-3n+3)S_1-1/4(n-1)(n^2+3n-6)S_2+(n^2-3)S_0^2}{n(n-2)(n-3)S_0^2} - b_2(\mathbf{z})\frac{(n-1)^2S_1-1/4(n-1)(n^2-n+2)S_2+(n-1)^2S_0^2}{n(n-2)(n-3)S_0^2}$

(Cliff and Ord, 1973, p 15-16)

$\mathbb{E}_R[G] =$	$\frac{S_0}{n(n-1)}$
$V_R[G(\mathbf{z})] =$	$\frac{m_2^2(\mathbf{z})((n^2-3n+3)S_1-nS_2+3S_0^2)-m_4(\mathbf{z})(n(n-1)S_1-2nS_2+6S_0^2)-m_1^2(\mathbf{z})m_2(\mathbf{z})(2nS_1-(n+3)S_2+6S_0^2)}{(m_1^2(\mathbf{z})-m_2(\mathbf{z}))^2n(n-1)(n-2)(n-3)}$ $+\frac{m_1(\mathbf{z})m_3(\mathbf{z})(4(n-1)S_1-2(n+1)S_2+8S_0^2)+m_1^4(\mathbf{z})(S_1-S_2+S_0^2)}{(m_1^2(\mathbf{z})-m_2(\mathbf{z}))^2n(n-1)(n-2)(n-3)} - \frac{S_0^2}{n^2(n-1)^2}$

(Getis and Ord, 1992, 1993)

Table 2.1: Expectations and variances of Moran's index I, Geary's ratio c and Getis and Ord's G under the randomisation and the normality assumptions ($b_2(\mathbf{z})$ denotes the kurtosis of $\mathbf{z}$ and $m_k = \sum_{j=1}^{n} z_j^k$).

of these tests is the lack of spatial autocorrelation. In order to formalize such a hypothesis, we consider the first order spatial autoregressive (SAR) model

$$(Z - \mu) = \rho W(Z - \mu) + \varepsilon \tag{2.7}$$

where Z is the variable under study, ρ is the spatial autoregressive parameter, μ is the constant mean of Z and ε is a disturbance term for which the distribution is known (in the simulations of Section 2.6, we consider Gaussian, Poisson and Bernoulli distributions). This model attempts to explain variation in the variable Z as a linear combination of the spatially lagged observations. In this context, the null hypothesis can be written as $H_0 : \rho = 0$.

Spatial correlation can also be tested on regression residuals using the same scheme as presented in (2.7). More complex spatial dependence tests are also available in the literature. For instance Tiefelsdorf (2002) based his test on saddlepoint approximation and Anselin et al. (1996) uses Lagrange multiplier. Nevertheless, this kind of testing is beyond the scope of our work. We focus here on spatial autocorrelation of any univariate variable.

From now on, we present only the implementation of the tests as available in the software `R` but the interested reader can find information on other software implementations of the three indexes of spatial autocorrelation and their associated tests in Bivand and Wong (2018).

Tests based on asymptotic normality A first naive approach is to use the asymptotic normality of the indexes to test if the spatial autocorrelation is significantly positive, negative or different from zero. These tests described in Cliff and Ord (1973) and in Getis and Ord (1992) are implemented in the `spdep` library of the software `R` under the normality or randomisation assumption (see Bivand and Wong, 2018, for further details). Their scheme is a comparison of the standardized indexes

$$\frac{I(\mathbf{z}) - \mathbb{E}[I]}{\sigma[I(\mathbf{z})]}; \qquad \frac{\mathbb{E}[c] - c(\mathbf{z})}{\sigma[c(\mathbf{z})]}; \qquad \frac{G(\mathbf{z}) - \mathbb{E}[G]}{\sigma[G(\mathbf{z})]} \tag{2.8}$$

with quantiles of the Gaussian distribution. The opposite standardization of Geary's ratio is explained by its opposite behaviour with respect to Moran's index. The test based on Getis and Ord's statistic is only available under randomisation assumption.

Permutation tests An empirical cut-off for the significant values of the indexes can also be defined by means of permutation tests. In this case, indexes are computed for random permutations of $\mathbf{z}$ on the fixed spatial domain in order to establish the rank of the observed statistic in relation to the simulated values. These tests are also

implemented in the `spdep` library of `R` (except for Getis and Ord's statistic). This technique is for instance applied to residuals in Lin et al. (2009). As explained in Smyth and Phipson (2010) and Ernst (2004), the pseudo p-value of this one-sided Monte-Carlo test is given by $(\text{nb}+1)/(\text{nsim}+1)$, where nb is the number of simulated values greater than the observed one.

Dray's test An alternative test based on Moran's index is proposed by Dray (2011). This test decomposes the influence of positive and negative autocorrelation. It is based on the following decomposition of Moran's index:

$$I(\mathbf{z}) = \underbrace{\sum_{I(u_k)<E[I]} I(u_k)cor^2(u_k,\mathbf{z})}_{S_I^-(\mathbf{z})} + \underbrace{\sum_{I(u_k)>E[I]} I(u_k)cor^2(u_k,\mathbf{z})}_{S_I^+(\mathbf{z})} \tag{2.9}$$

where u_k are the orthogonal eigenvectors of the symmetric matrix HWH. Then, to test positive (resp. negative) spatial autocorrelation, one considers S^+ (resp. S^-) and the two-sided test uses both statistics. A pseudo p-value is computed using the Hope (1968)–type permutation test with around 5000 simulations according to Dray (2011).

2.4 Robustness of tests

The term "Robustness" may be loaded with many connotations. Here, we define robustness as the insensitivity to small deviations from assumptions and more precisely, the outlier resistance (Huber, 1981). The robustness of estimators is usually quantified by means of two tools: the breakdown point (Hampel, 1971) and the influence function (Hampel et al., 1986). The breakdown point determines the minimal proportion of contamination needed to break the estimators (explosion or implosion). Genton and Lucas (2003) adapted its definition to the context of dependent data. The influence function quantifies the impact of an infinitesimal contamination on the estimator.

When one deals with inferential statistics, the tools are slightly different as the actual question is how contaminated observations influence the result of the test and no longer the value of the estimator. Influence functions and other robustness measures for inference are developed in the literature, for instance in Lambert (1981), Hampel et al. (1986) and Ronchetti (1997). However, their proposals are based on functionals which correspond to asymptotic values of the statistics under consideration, and this approach is not so natural in our setting. Indeed, as already discussed in Genton and Ruiz-Gazen (2010), we work on finite sets of locations and assuming n increasing

to infinity is usually not realistic. Therefore, we decided to favour empirical tools instead of asymptotic ones.

As suggested in Lambert (1981), the strength of the evidence against the decision to reject H_0, i.e., the effect of an observation on the p-value of the test, can be measured. For example, the empirical influence function of the p-value of a test could be computed. Moreover, Lambert (1981) proposed to derive the influence function on transformed p-values, $P_n := -n^{-1}\log(\text{p-value})$ or $P_n := -\Phi^{-1}(\text{p-value})$. These transformations were initially motivated by the asymptotically log-normal distribution of the p-value under the alternative and the asymptotically normal distribution of the statistics of the test under the null hypothesis (see Lambert, 1981 and Lambert and Hall, 1982 for more details about the conditions). In addition, the influence functions of transformed p-values have the same properties as usual influence functions in the sense that boundedness may be interpreted as a robust property. However, in order to keep an understandable scale, we develop hereafter the empirical influence function based on the p-value even if, by definition, the boundedness of the function is assured and can not be interpreted in terms of robustness in this case.

In the context of spatial data on a fixed design, the influence function should be defined in terms of "replacement" of an observation z_i by the modified value $z_i + \xi$. Therefore, in the contaminated p-value, the vector $\mathbf{z}$ is replaced by the vector $\mathbf{z}+\xi e_i$. By definition,

$$EIF(\xi, i) = \frac{\text{p-value}(\mathbf{z} + \xi e_i) - \text{p-value}(\mathbf{z})}{1/n},$$

where $1/n$ is the proportion of contamination and p-value($\mathbf{z}$) is the p-value of the considered test applied on the vector $\mathbf{z}$. This definition depends on the contaminated location s_i. Moreover, as explained in Genton and Ruiz-Gazen (2010), this tool could be useful to detect influential observations in a dependent dataset. In order to make the reading easier, we develop only Moran's index case. The results based on Geary's ratio and Getis and Ord's statistic are available in the Appendix.

Proposition 2.4.1. *Let $\mathbf{z}$ be the n-dimensional column vector of observed values at location $\{s_1, \ldots, s_n\}$ and W the weighting matrix associated with the domain. The empirical influence function of the p-value of unilateral tests based on asymptotic normality for Moran's index is explicitly given by*

$$EIF(\xi, i; I) = n\left[\Phi\left(-\frac{I(\mathbf{z} + \xi e_i) - \mathbb{E}[I]}{\sigma[I(\mathbf{z} + \xi e_i)]}\right) - \Phi\left(-\frac{I(\mathbf{z}) - \mathbb{E}[I]}{\sigma[I(\mathbf{z})]}\right)\right]$$

where Φ is the cdf of the standard Gaussian distribution, $\mathbb{E}[I] = -1/(n-1)$ and the standard deviation $\sigma[I(\cdot)]$ is explicit under normality and randomisation assumptions (see Table 2.1).

The proof is straightforward by noticing that p-value($\mathbf{z}$) for the unilateral asymptotic test is given by $\Phi\left[-(I(\mathbf{z}) - \mathbb{E}[I])/\sigma[I(\mathbf{z})]\right]$. Moreover, under normality assumption, the variance is constant and explicit (see Table 2.1). Therefore, as ξ tends to infinity, we get

$$EIF(\xi, i; I) \to n\left[\Phi\left(\frac{2\left(nw_{i\bullet} - S_0\right)}{(n-1)S_0\sigma[I]}\right) - \Phi\left(-\frac{I(\mathbf{z}) - \mathbb{E}[I]}{\sigma[I]}\right)\right]. \tag{2.10}$$

The only dependence on i in the limit value of the empirical influence function lies in the row weight $w_{i\bullet}$ of the chosen location s_i. This limit value increases with the row weight $w_{i\bullet}$. If, instead of unilateral tests, we consider bilateral alternatives, the limit value is maximal for a mean row weight, i.e., $w_{i\bullet} = S_0/n$ and decreases as $|w_{i\bullet} - S_0/n|$ increases. Similar observations can be made under randomisation assumption, with the difference that $\sigma[I]$ is replaced with the limit of the standard error obtained under contamination, which can be expressed as $\sqrt{a_1 - a_2(n^2 - 3n + 3)/(n-1)}$. The constants a_1, a_2 are defined according to the expression of the variance of Moran's index under randomisation assumption, i.e., $V_R[I(\mathbf{z})] = a_1 - a_2 b_2(\mathbf{z})$ (see Table 2.1 for details).

Once again, one insists on the importance of the choice of the weighting matrix as it entirely determines the limit of the p-value of a test when there is an outlier. Let us recall that if the weighting matrix is row-standardized, the elements $w_{i\bullet}$ are defined according to their symmetrized version (see discussion page 40). Therefore, they usually can get different values.

As an illustration, we consider the *crude divorce rate*[1] for 1,000 inhabitants in Belgian municipalities in 2017. As discussed in Su et al. (2018), the divorce rates may be spatially autocorrelated in various provinces of China. We may wonder if such a behaviour is also observed in our country. The crude divorce rate in Belgian municipalities is illustrated on Figure 2.3 where we can observe some homogeneous regions, which may lead to some positive spatial autocorrelation. However, Moran's tests fail to reject the null hypothesis of no spatial autocorrelation. Indeed, Moran's index is 0.01 and the p-value of unilateral Moran's test is 0.25 under randomisation assumption and 0.32 under normality assumption. In the graphical representation, we directly observe the particular behaviour of the municipality of Brussels where 3698 divorces were pronounced in 2017 for a population of 176,545 inhabitants, which gives a crude divorce rate of 20.95 per mille when the other municipalities vary from 0 to 4 per mille (this huge rate is explained by the fact that the divorces of foreign marriages are registered in Brussels, and this corresponds to 3,674 such divorces in 2017). In order to have an understandable scale, the empirical function is represented up to a multiplicative factor in order to compare the difference between the p-values.

1. Source: StatBel, the Belgian Statistical office.

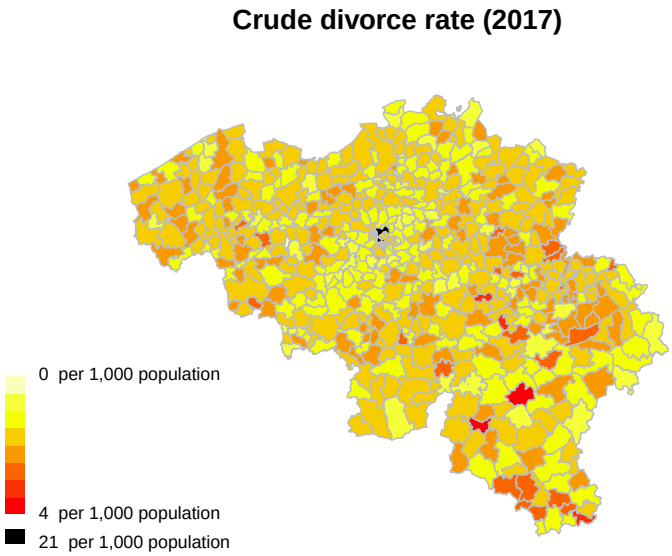


Figure 2.3: Distribution of the crude divorce rate for 1,000 inhabitants in Belgian municipalities (Source: StatBel).

Figures 2.4 and 2.5 illustrate the impact of a unique contamination on the divorce rate example. This graph is called a *hair-plot* (Genton and Ruiz-Gazen, 2010) and each hair corresponds to a different contaminated location. In Figure 2.4, we represent the influence of the municipality of Brussels in red and the influence of its neighbours in blue for Moran's test. All the other municipalities correspond to black curves. A small contamination ξ on the neighbours of Brussels has a larger impact on the p-value than any other location.

Figure 2.5 illustrates the impact of a unique contamination ξ on the p-value of the test, for large values of ξ and for the two tests (under normality and randomisation assumptions). The initial p-values of both tests are represented by a horizontal red line. As illustrated, the contaminated p-value can reach almost any probability for well chosen contaminated location s_i and contaminated value ξ. Moreover, the colors make it possible to distinguish the different values of $w_{i\bullet}$ which correspond to different limit values. It illustrates the monotonically increasing behaviour of

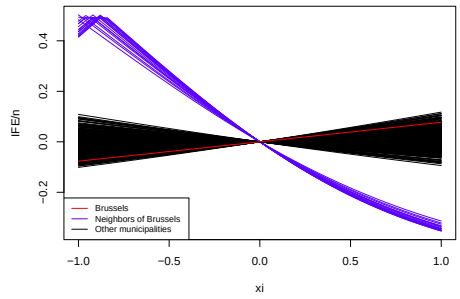


Figure 2.4: Variation of the p-value for a unique contamination using Moran's asymptotic test under normality assumption. Each hair corresponds to a different contaminated location.

empirical influence functions as the row weight increases. The hair-plot of p-values associated with bilateral tests (not shown) shows that the empirical influence function decreases as the row weight goes away from the mean value S_0/n.

As mentioned earlier, the choice of weighting matrix is crucial for the spatial autocorrelation as the results of the test may differ. Indeed, if the weighting matrix was defined by the 6 nearest neighbours, using the official data set of crude divorce rate, we would obtain a Moran's index of 0.04 and would conclude that there is positive spatial autocorrelation (p-value is 0.008 under randomisation and 0.049 under normality assumption).

In complement to the empirical influence function, we can determine the resistance of a test which is the analogous definition of the breakdown point of an estimator. Ylvisaker (1977) defined *the resistance to acceptance (rejection)* of a test as the smallest portion m/n of the data that must be corrupted to guarantee the acceptance (rejection) of the null hypothesis. Coakley and Hettmansperger (1992) and Hettmansperger and McKean (2010) refer to similar concepts as *acceptance breakdown* and *rejection breakdown*. As mentioned in Genton (1998b) in the spatial context or in Ma and Genton (2000) for the temporal context, the locations of the perturbed data become important. Therefore, the resistance is defined according to the most unfavourable configurations of perturbation. Furthermore, Genton (1998b) and Ma and Genton (2000) notice that this definition is local, in the sense that it may be valid only for fixed neighbourhoods, a.k.a. a fixed weighting matrix.

We proved the following results about the resistance to acceptance and rejection of Moran's tests.

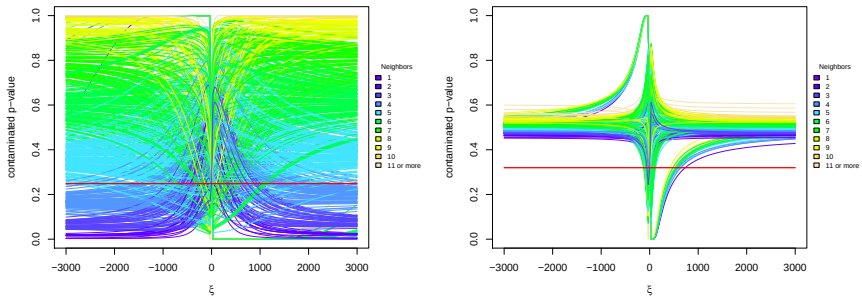


Figure 2.5: Contaminated p-value for a unique contamination using asymptotic Moran's tests. Tests are based on Moran under randomisation assumption (left panel) and normality assumption (right panel). The limit of the hair i (corresponding to i-th contaminated location) depends on the correspond $w_{i\bullet}$ value (different colours).

Proposition 2.4.2. *Let* $\mathbf{z}$ *be the observed values at locations* $\{s_1, \ldots, s_n\}$ *and* W *the corresponding weighting matrix. The resistance to acceptance of both asymptotic tests based on Moran's index is* $1/n$.

The proof (see the Appendix) is based on an argument of Genton and Ruiz-Gazen (2010). In that paper, simple tools are developed in order to visualize influential observations in the context of dependent data. In the spatial context, they suggest to contaminate, as done in the empirical influence function, with the vector $\mathbf{z}^* = \mathbf{z} + \xi e_i$. Then, following their argument, the influence of such a contamination on Moran's index is measured. Same results based on other tests and Geary's ratio and Getis and Ord's statistic are given in the Appendix.

The derivation of the resistance to rejection is more challenging because it depends on the choice of weighting matrix and the level α of the test. The proof is provided in the Appendix.

Proposition 2.4.3. *Let* $\mathbf{z}$ *be the observed values at locations* $\{s_1, \ldots, s_n\}$ *and* W *the corresponding weighting matrix. The resistance to rejection is* m/n *where* m *is the size of the smallest subset* $A \subseteq \{1, \ldots, n\}$ *which satisfies*

$$\frac{n^2 w_A - 2nm(w_A + w_B) + m^2 S_0}{S_0 m(n-m)} \geq \frac{-1}{n-1} + \sigma_N[I] z_{1-\alpha} \tag{2.11}$$

for tests based on normality assumption or which satisfies

$$\frac{n^2 w_A - 2mn(w_A + w_B) + m^2 S_0}{S_0 m(n-m)} > \frac{-1}{n-1} + \sqrt{a_1 - a_2 \frac{n^2 - 3nm + 3m}{(n-m)m}} z_{1-\alpha} \quad (2.12)$$

for tests based on randomisation assumption. We denote $w_A = \sum_{i,j \in A} w_{ij}$ *and* $w_B = \sum_{i \in A} \sum_{j \notin A} w_{ij}$. *The constants* a_1 *and* a_2 *are defined as previously by the weighting matrix (see page 46 or Table 2.1 for details).*

Quite naturally, the resistance of a test is linked to the breakdown point of the considered statistics (mentioned for instance in Capéraà and Guillem, 1997). As explained in Genton and Ruiz-Gazen (2010), the asymptotic breakdown point of Moran's index is $1/n$ because Moran's index may break down to $-1/(n-1)$, the expected value of Moran's index under the null hypothesis of independence, with one extreme contamination. Then, it is obvious that the resistance to acceptance is equal to $1/n$. It means that a unique large value is enough to always fail to reject the null hypothesis.

As an illustration, we go back to the crude divorce rate in Belgium. The municipality of Brussels illustrates the lack of resistance to acceptance as only one municipality can drastically modify the result of this test. Indeed, if we replace the value of Brussels with the "true" value (24 local divorces), Moran's index reaches 0.14, which corresponds to a p-value smaller than 0.0001 (under both assumptions). On the other hand, on the pre-mentioned domains, for a level of 5%, the resistance to rejection is $2/n$ for Moran's asymptotic tests (as long as $a \geq 4$ for the regular grid). The same results is obtained using row-standardized weighting matrices. Indeed, Moran's condition (2.11) is never verified for $m = 1$ but if $m = 2$ and A is a subset of two neighbours, (2.11) is verified for $a \geq 4$ on a grid and is verified under the irregular Belgian setting. Therefore, two neighbours with large observed values are enough to always reject the null hypothesis, whatever the value corresponding to the other $n - 2$ observations.

2.5 Robust versions of Moran's tests

The previous section illustrates the lack of robustness of the classic tests. One could argue that, in practice, the classic tests are often replaced with a permutation test whose sensibility to contamination might be better. Nevertheless, this conclusion is not correct. Indeed, even if the expression of the empirical influence function of permutation tests cannot be explicitly given as for asymptotic tests, it is easy to observe that for a unique contamination that is large enough, the empirical influence function

for Moran's permutation test is only linked to the rank of $(S_0 - 2nw_{i\bullet})/(S_0(n-1))$ in the set of values $(S_0 - 2nw_{j\bullet})/(S_0(n-1))$ for permuted locations $j = \tau(s_i)$. The contaminated p-value of the permutation tests for Moran's index is represented in Figure 2.6 for the crude divorce rate in Belgium. As for asymptotic tests, the contaminated p-value can take almost any value for well-chosen s_i and ξ. Moreover, as detailed in Appendix A.3, the resistance of the permutation test is $1/n$.

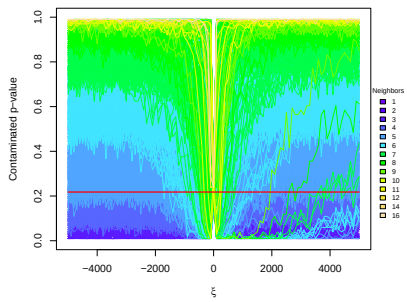


Figure 2.6: Contaminated p-value for a unique contamination using permutation test based on Moran. Each hair corresponds to a different contaminated location. The limit of the hair i depends on the corresponding $w_{i\bullet}$ value.

Therefore, robust alternatives are needed. In order to robustify these tests, we propose to replace the usual index with a robust version without changing the procedures. This follows the "plug-in" principle frequently advocated in robust statistics. If the asymptotic normality is no longer valid, we may still use the permutation test or a test based on a bootstrapped cut-off. Several paths are considered here. A first naive approach is to compute the "rank Moran index" obtained by replacing, in the definition of the index, the observations with their ranks. A second idea is to use the definition of Moran's index as a regression slope and to estimate that slope by means of a robust regression.

2.5.1 Moran index based on ranks

Let R_i denote the rank of z_i in $\{z_1, \ldots, z_n\}$ and I_r denote the rank Moran index, i.e.,

$$I_r(\{R_1, \ldots, R_n\}) = \frac{12}{S_0(n^2-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(R_i - \frac{n+1}{2}\right)\left(R_j - \frac{n+1}{2}\right). \quad (2.13)$$

The interpretation of the rank Moran index is the same as the classic Moran's index, the range of values is also defined according to the eigenvalues of a transformed weighting matrix (see de Jong et al., 1984). It is convenient to rewrite (2.13) in a simpler form, after modifying the numbering of the spatial locations according to the order of observed values in order to get $R_i = i$ for the location s_i. This modification only implies that the new weighting matrix W^r is defined up to a permutation of rows and columns. In this context, if $\mathbf{R} = (1, 2 \ldots, n)'$,

$$\begin{aligned} I_r(\mathbf{R}) &= \frac{12}{S_0(n^2-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^r \left(i - \frac{n+1}{2}\right)\left(j - \frac{n+1}{2}\right) \\ &= \frac{12}{S_0(n^2-1)} \left(\mathbf{R} - \frac{n+1}{2}\mathbb{1}\right)' W^r \left(\mathbf{R} - \frac{n+1}{2}\mathbb{1}\right). \end{aligned}$$

Normality can no longer be assumed for the ranks but the randomisation assumption and the permutation test are still valid. As mentioned in Cliff and Ord (1973, p. 142), I_r keeps the same expectation as I, i.e.,

$$\mathbb{E}[I_r] = \frac{-1}{n-1}$$

and, if ties are absent, the mean rank is $(n+1)/2$ and $b_2(\mathbf{R}) = (9n^2 - 21)/(5n^2 - 5)$. Therefore, the rank Moran index has the following variance which correspond to the variance $V_R[I(\mathbf{R})]$ in Table 2.1:

$$V_R[I_r] = \frac{n(n-1)(6+5n)S_1 - (5n+7)(nS_2 - 3S_0^2)}{5(n-1)^2(n+1)S_0^2} - \frac{1}{(n-1)^2}.$$

Some insight on the spatial autocorrelation can be deduced from the general shape of the reordered weighting matrix. If consecutive ranks are associated with neighbours, the weights are mostly on the diagonal blocks of the matrix. The left panel of Figure 2.7 illustrates a situation based on the Belgian context which is clearly spatially autocorrelated (rank Moran index is 0.92 while the mean value under the null hypothesis is -0.0017 and the extreme values are -0.67 and 1.16). On the other hand, an "average" rank Moran index corresponds to a larger dispersion of the weights. Such a situation is illustrated on the right panel using a regular grid 25×25 and the queen contiguity (the rank Moran is -0.023 when the mean value under the null hypothesis is -0.0016 and the extreme values are -0.52 and 1.03).

Robustifying a statistical techniques by using ranks instead of the observations is a common practice in robust statistics, e.g. Spearman correlation, multivariate Oja's ranks (Oja, 1999). Such a strategy indeed leads to a better resistance to

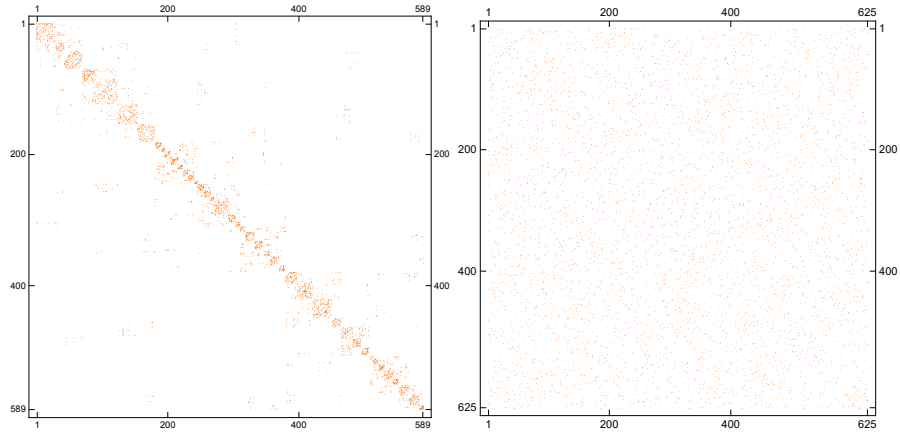


Figure 2.7: Reordered weighting matrix to obtain a large (left panel - Belgian context) or "average" (right panel - regular grid) rank Moran index.

contamination. Let us come back to the examples considered in Section 2.4. On the original data, the rank Moran index is 0.140, while on the modified data set, we get 0.147. Knowing that the rank Moran index ranges from -0.67 to 1.16, this example illustrates the fact that a single outlying observation cannot change drastically the result of the test. This can be formalized by means of the empirical influence function on the p-values. Let us first introduce some notations. W.l.o.g. we may assume that $z_1 \leq \cdots \leq z_n$ (even if it means that the rows and columns of W are reordered). Therefore, the initial rank vector $\mathbf{R}$ is $(1, \ldots, n)$. In order to compute the influence function, the value observed at a given location s_i needs to be perturbed. Obviously, such a perturbation induces the modification of several ranks. The observation at location s_i can increase such that its rank i becomes $i + k$ and the next k ranks decrease by one for k in $\{1, \ldots, n - i\}$. Alternatively, the observation at location s_i could be reduced such that the i-th rank decreases to $i - k$ and the k previous ranks increase by one. In the first case, the new rank vector can be written as $\mathbf{R}^* = \mathbf{R} + \sum_{j \in A}(e_i - e_j)$ for $1 \leq 1 < n$, $1 \leq k \leq n - i$ and $A = \{i + 1, \ldots, i + k\}$. In the second case, it can be written as $\mathbf{R}^* = \mathbf{R} - \sum_{j \in A}(e_i - e_j)$ for $1 < k \leq n$, $1 \leq k < i$ and $A = \{i - 1, \ldots, i - k\}$. Proposition 2.5.1 yields the difference in the rank index when perturbing a single observation.

Proposition 2.5.1. *Let $z_1 \leq \cdots \leq z_n$ be the ordered observations and $\mathbf{R}$ the associated rank vector $(1, \ldots, n)'$. If $\mathbf{R}^* = \mathbf{R} \pm \sum_{j \in A}(e_i - e_j)$ is the contaminated rank*

vector obtained by the perturbation of the location s_i,

$$I_r(\mathbf{R}^*) - I_r(\mathbf{R}) \tag{2.14}$$
$$= \frac{24}{S_0(n^2-1)} \left(\sum_{j\in A} \sum_{\ell \in A} \left(\frac{1}{2} w_{j\ell} - w_{i\ell} \right) \right) \pm \sum_{j=1}^{n} \left(j - \frac{n+1}{2} \right) \left(\sum_{\ell \in A} (w_{ij} - w_{\ell j}) \right).$$

The proof is trivial by a simple rewriting of the rank Moran index. The first term in the right hand side can be seen as a difference over the perturbed weights and the second term is some average rank over the modified neighbourhoods. From that proposition, it is straightforward to derive the influence function of the p-value.

Proposition 2.5.2. *Using the same notations as in Proposition 2.5.1, we have*

$$\begin{aligned} EIF(\xi, i; I_r) &= n \left[\Phi \left(\frac{I_r(\mathbf{z}) - E[I_r]}{\sigma[I_r]} \right) - \Phi \left(\frac{I_r(\mathbf{z} + \xi e_i) - E[I_r]}{\sigma[I_r]} \right) \right] \\ &= n \left[\Phi \left(\frac{I_r(\mathbf{z}) - E[I_r]}{\sigma[I_r]} \right) - \Phi \left(\frac{I_r(\mathbf{z}) - E[I_r]}{\sigma[I_r]} + \frac{e}{\sigma[I_r]} \right) \right] \end{aligned} \tag{2.15}$$

where e is the difference between Moran's two indexes $Ir(\mathbf{z})$ and $Ir(\mathbf{z} + \xi e_i)$.

For a fixed weighting matrix W, the equation (2.14) gives us directly the difference e which can be expressed as a function of i and k. This difference only depends on the spatial context via the location s_i and the weight of the k collateral damaged ranks. Therefore, due to the shape of the cdf of the Gaussian distribution, the impact on the p-value is limited, especially if the initial p-value is close to zero or one. For instance, in the Belgian context, the empirical influence function is constantly equal to zero using the rank vector used in the left panel of Figure 2.7, i.e. in a highly spatially autocorrelated case. Using randomly assigned ranks, the contaminated vector induces small modifications in the p-value but they do not change the conclusion of the test. An example is illustrated in Figure 2.8 where we represent the contaminated p-value with respect to the number of contaminated other ranks (negative values correspond to negative contaminations). Each hair corresponds to a specific contaminated location. In this example, the initial rank Moran index is 0.0004 and the initial p-value is 0.47 under randomisation assumption.

In Section 2.4, additional robustness measures were introduced in order to determine the resistance to acceptance and rejection. However, these characteristics are extremely dependent on the weighting matrix. Indeed, as already stressed, the resistance is defined using the most unfavourable configurations. Therefore, as in Ma and Genton (2000) (temporal context) or in Genton (1998b) (spatial context), we would need to find the most unfavourable configuration of perturbed data for a fixed weighting matrix W. In the mentioned papers, specific neighbourhoods are

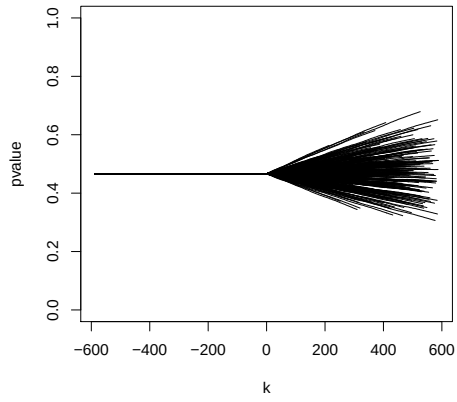


Figure 2.8: Contaminated p-value of the rank Moran test in the Belgian context for a non spatially autocorrelated vector.

defined using the distance-based nearest neighbours (the temporal context can be seen as a uni-dimensional spatial grid). On these specific neighbourhoods, even with robust methods, the resistance decreases steadily as the size of neighbourhoods rises. One expects to make similar comments for any definition of neighbourhoods. For instance, it is expected to observe a smaller resistance for the queen contiguity than the rook contiguity on the same regular grid. As illustration, we may consider the simplistic example of a grid 2×3 which gives a resistance to rejection of 3/6 for the rook contiguity and 2/6 for the queen contiguity.

Nevertheless, as detailed in Genton (1998b), it is generally hard to compute the maximal number of perturbations. The expression (2.14) is useful in order to reach some local extrema for the rank Moran. From any starting configuration, we may define chains of contaminations which increase (resp. decrease) the rank Moran in order to reach the cut-off value. The resistance to rejection and acceptance are at most the lengths of the smallest chain obtained from the most unfavourable configurations. Unfortunately, our statistic is not linear, then the result derived in Capéraà and Guillem (1997, Proposition 2.3) can not be applied here and we only get an upper bound for the resistance instead of an equality. Therefore, numeric approximations are performed to determine upper bounds of the resistance to acceptance and to rejection. We consider here the alternative hypothesis associated with the presence of positive spatial autocorrelation. As previously discussed, the other alternatives can easily be adapted. In this configuration, unfavourable situations to reject the

null hypothesis are when there is negative spatial autocorrelation and unfavourable configurations to accept the null hypothesis are when the rank is spatially positively autocorrelated. Rank vectors based on the vectors associated with the extreme values for Moran's index (de Jong et al., 1984) can be used as starting points. However, in presence of ties, several choices of rank vectors are possible and may lead to different values. Numerical computations based on 100 different starting points allow to estimate the resistance to rejection and to acceptance to be of order $a/n = 1/\sqrt{n}$ for the regular grid (rook or queen contiguity). Moreover, as expected, the resistance to rejection is larger than the resistance to acceptance and the queen contiguity less robust than rook contiguity. On the Belgian context, the resistance to rejection is at most 14/589 and the resistance to acceptance 6/589.

2.5.2 Moran index using robust regression

As mentioned in Anselin (1996), Moran's index can be defined as the slope of a bivariate linear regression of the spatially lagged variable on the original variable. More precisely, let the centred lag vector of z around the location s_j be denoted by

$$\tilde{\mathbf{z}} = (\tilde{z}_1, \ldots, \tilde{z}_n)'$$

where each component is the weighted sum of the centred vector over the considered neighbourhood, i.e.,

$$\tilde{z}_j = \sum_{i=1}^{n} w_{ij}(z_i - \bar{z}).$$

The vector can be rewritten as $\tilde{\mathbf{z}} = WH\mathbf{z}$, where $H = I_n - 1/n\mathbb{1}\mathbb{1}'$. Then, the mean of this vector is $\bar{\tilde{z}} = 1/n \sum_{i,j} w_{ij}(z_i - \bar{z})$. Therefore, Moran's index can be rewritten as

$$I(\mathbf{z}) = \frac{n}{S_0} \frac{\sum_{i=1}^{n}(z_i - \bar{z})(\tilde{z}_i - \bar{\tilde{z}})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

which is the least squares estimated slope of the linear regression of $\tilde{\mathbf{z}}$ over $\mathbf{z}$ multiplied by the factor n/S_0. If the weighting matrix is row-standardized, Moran's index is directly given by the estimated slope of the LS linear regression of the lagged vector $WH\mathbf{z}$ over $\mathbf{z}$.

The scatterplot of the spatially lagged observations with respect to the initial ones is called the *Moran scatterplot* (Anselin, 1996). Each point corresponds to a location and such a construction allows to detect outliers with respect to the central tendency. This graphical tool is useful to analyse the local impact of each observation. Such local association is measured using local indicators of spatial statistics (LISA) such as local Moran and local Geary (Anselin, 1995) or G_i statistic (Getis and Ord, 1992).

These local indicators would make another interesting topic of study which is not developed here (see for instance the work of Lee, 2009). Figure 2.9 illustrates Moran scatterplot for a simulated example on a regular grid 10×10 with rook contiguity. In this example, five observations have been contaminated. We observe the particular behaviour of the contaminated observations (red points) and the impact on their neighbours (blue points). This example illustrates the fact that outliers in the vector $\mathbf{z}$ can induce outlying observations with respect to both axes in the Moran scatterplot as it impacts their neighbours as well.

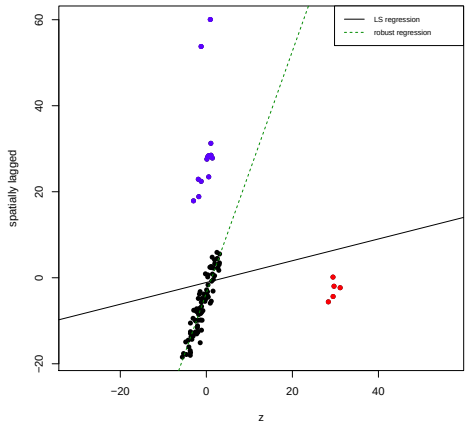


Figure 2.9: Moran scatterplot on a simulated example based on a regular grid of size 10×10 with rook contiguity.

An intuitive way to robustify Moran's index is to use an estimation of the regression parameter which is robust against outliers and bad leverage points. In the literature, many robust regressions have been introduced (see a review of classic methods in Hampel et al., 1986 or, more recently in Yu and Yao, 2017). In this context, we suggest to favour a robust and efficient regression estimation. Several methods satisfy these properties and are already implemented in `R` packages [2], especially the Least trimmed squares (LTS – Rousseeuw, 1985), the S-estimator (Rousseeuw and Yohai, 1984), the MM-estimator (Yohai, 1987) and the robust and efficient weighted least squared estimator (REWLSE – Gervini and Yohai, 2002). By construction, these techniques have a high breakdown point and are highly efficient. Moreover, Yu

2. Implementations of robust regressions are available in the `R` packages `robustbase`, `robust` and `MASS`. For instance, LTS and S-estimator with `lqs` (`MASS`), MM-estimator with `rlm` (`MASS`) and REWLSE using `lmRob` (`robust`).

and Yao (2017) pointed out that, based on their simulation study and despite their robust properties, LTS and S-estimates are less efficient than MM and REWLSE estimates. Indeed, we observe a larger variability in the estimations of the robustly estimated Moran's index based on LTS and S-estimations than the others.

In the simulated example of Figure 2.9, all these techniques give a similar estimation of the slope which is equivalent to the classic estimation on the uncontaminated vector. Therefore, only one robust regression is represented in Figure 2.9, which corresponds to the robustly estimated Moran's index $\hat{a}n/S_0 = 0.78$ instead of 0.07 for the classic Moran's index.

To test whether the robustly estimated Moran's index induces a large spatial autocorrelation, if the asymptotic normality is plausible, we may plug the estimated parameter into the asymptotic tests and the permutation test. To justify the plug-in procedure, we may cite Gervini and Yohai (2002) who prove that REWLSE is asymptotically equivalent to the least square estimator. Moreover, based on simulated examples, we observe that the asymptotic distribution of the robustly estimated Moran's index is acceptable for MM and REWLSE estimations. However, under randomisation assumption, the normality is rejected for LTS and S-estimations. Therefore, we focus hereafter on tests based on MM and REWLSE estimations for asymptotic tests. Alternative test based on bootstrap cut-off could be defined but this proposition is not developed hereafter.

Due to the robust properties of the considered regression techniques, the perturbation at a location does not drastically change the estimation of the regression parameters (the expression of influence functions for MM and REWLSE estimations are available in their initial papers, i.e. respectively Yohai, 1987 and Gervini and Yohai, 2002). Therefore the impact on the p-value of a test is small and the empirical influence function on p-value is close to zero almost everywhere. Moreover, the robust regressions mentioned hereabove may reach the optimal breakdown point of 50%. Therefore, the resistance of the tests is defined by the number of contaminations which induce at most $n/2$ outliers in the Moran scatterplot. The change of m observations z_i modifies m x-coordinate in the scatterplot and the y-coordinate of all points. Nevertheless, only the neighbours of the modified points are highly perturbed, as illustrated in Figure 2.9 for the divorce rate in Belgium. Therefore, even if the regression method has the largest breakdown point, the resistance of the associated test is always lower than 50% due to the spatial context.

More precisely, the minimal resistance is defined by m/n where m is the size of the smallest subset A for which the number of their neighbours outside A is larger than $n/2 - m$. Again, the resistance is defined with respect to the weighting matrix. For the Belgian municipalities, the minimal resistance is $35/589$ instead of $1/589$ or $2/589$ for the classic asymptotic tests. For a regular grid, the resistance is $\lceil n/18 \rceil / n \approx 1/18$

for the queen contiguity and $\lceil n/10 \rceil / n \approx 1/10$ for the rook contiguity (see details in Appendix A.4).

2.6 Simulation study

In the previous sections, we discussed the lack of robustness of the classic tests of spatial autocorrelation and proposed some robust alternatives to tests based on Moran's index. Now that their robustness has been discussed, we need to verify that the gain in robustness does not overstep the efficiency of the robust alternatives with respect to the classic methods. Therefore, some simulations are performed to compare the power of the classic and robust tests. The simulations are conducted on uncontaminated datasets to allow the comparison between classic tests and robust alternatives. The power of classic tests based on Moran's index is already studied in the literature in the context of linear regression (see for instance Anselin and Rey, 1991, Bivand et al., 2009 or Ou et al., 2015). However, to the best of our knowledge, there is no comparison of power in the general settings.

The considered tests included in the simulation study are presented in Table 2.2. For each estimation of Moran's index, we may use tests based on normality assumption (N), randomisation assumption (R) or permutation tests (perm.).

Moran estimation		Assumptions		
Classic I		N	R	Perm.
Dray				Perm.
Rank Moran			R	Perm.
Robust regression	LTS			Perm.
	S-estimator			Perm.
	MM-estimator [3]	N	R	Perm.
	REWLSE	N	R	Perm.

Table 2.2: Considered tests for spatial autocorrelation.

2.6.1 Simulation setting

The simulation set-up follows the scheme introduced in Holmberg and Lundevaller (2015) (in their homoscedastic case). More precisely, the first order spatial autoregressive model (2.7) is used on regular square grids of different sizes (100, 400, 900 or 1600 areas) and on the irregular domain of Belgian municipalities. The weighting

3. The results are identical whatever the scale estimator (MAD or Huber) and the ψ-function (Huber, Hampel or bisquare function).

matrices are defined as discussed in the examples on page 38. Different values for the spatial correlation coefficient are used, $\rho = 0, 0.1, 0.2$ or 0.3. The vector ε is generated according to a Gaussian, Poisson or Bernoulli distribution. As in Holmberg and Lundevaller (2015), we compare the proportion of rejected null hypotheses of the tests based on 5000 replicates of each setting for a significant level of 5%.

The simulation setting of Chapter 1 using the Matérn model, could also be easily adapted to the univariate case to carry out these simulations. However, we decided to consider here the simplest model, i.e., the first order autoregressive model.

2.6.2 Results

The results of the simulations are useful to analyse the level and the power of each test. A correct test should have 5% of rejection when the spatial correlation coefficient ρ is zero. If the correlation parameter increases, the power of the tests increases as well. Below follows a description of the simulation results for Gaussian, Poisson and Bernoulli observations. The results are detailed in the Gaussian case. For the Poisson and Bernoulli cases, only condensed results for rook contiguity are presented as the main findings from the normal case are similar for the other layout.

Gaussian case The proportions of rejection are displayed in Tables 2.3 and 2.4 for the regular grid using rook and queen contiguity; Table 2.5 lays out the results based on the Belgian municipalities domain.

On a regular grid, all the tests hold their significance level around 0.05. Moreover, under the null hypothesis, the p-values of each test should be uniformly distributed (see for instance Murdoch et al., 2008). In Figure 2.10, one can observe the uniform distribution for the asymptotic Moran's test and the robust test based on MM-regression.

As the spatial correlation coefficient ρ increases, the power of all tests increases. Moreover, the permutation tests based on Dray proposal, LTS and S estimators are slightly less powerful than the others which all have similar results. Therefore, Figure 2.11(a) only illustrates the power of the asymptotic Moran's test under normality with respect to the spatial correlation coefficient, the size of the grid and the contiguity. As expected, the power increases with the sample size and the spatial autocorrelation is more often detected using rook than queen contiguity. Similar conclusions hold as well under the irregular setting of Belgian municipalities (see Table 2.5 for details).

Results for Poisson and Bernoulli cases Table 2.6 shows the proportion of rejected null hypotheses for the tests using respectively Poisson and Bernoulli simulations for a regular grid $a \times a$ using rook contiguity (the conclusions of other settings

are similar to the Gaussian case). Of course, in these settings, tests under normality assumption do not apply. For Poisson, all tests hold reasonable level while the power of rank tests clearly outperforms the other tests' and Dray's test always has the lowest power. Figure 2.11(b) compares the power of the asymptotic Moran's test for any configuration (sample size and contiguity) and the power of the rank test (similar results in all configurations).

In the Bernoulli case, robust tests perform rather similarly and beat the classic asymptotic Moran's test. However, the REWLSE algorithm does not always converge to a solution, around 1200 simulations over 5000 fail to converge in each mentioned case (results with a symbol $*$). Figure 2.11(c) allows us to observe this situation based on a regular grid 20×20.

2.7 Conclusion

The aim of this chapter was to propose robust alternatives to tests of spatial autocorrelation based on Moran's index using either the rank vector or a robust regression. Moreover, these tests work well for detecting spatial autocorrelation especially under non-Gaussian distributions where they outperform the classic asymptotic tools while, under normality, all of them yield similar conclusions. In addition, all tests have more difficulties keeping their power for queen than rook layout. Therefore, as the number of neighbours increases, it is more difficult to rule on the problematic of spatial autocorrelation.

We also illustrate the usefulness of robust alternatives using the real dataset of divorce rate in Belgium. The classic tests conclude to the lack of spatial autocorrelation when the robust ones find strong evidence of positive spatial autocorrelation.

Grid	Normality tests			Randomisation tests				Permutations tests						
Gaussian (rook) ρ	I	MM	REWLSE	I	rank	MM	REWLSE	I	Dray	rank	LTS	S	MM	REWLSE
10×10														
0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.06	0.06	0.06	0.06	0.06
0.1	0.17	0.18	0.18	0.17	0.16	0.18	0.18	0.19	0.15	0.18	0.11	0.14	0.19	0.19
0.2	0.42	0.42	0.42	0.42	0.40	0.42	0.41	0.45	0.35	0.42	0.21	0.27	0.44	0.42
0.3	0.72	0.72	0.71	0.72	0.68	0.72	0.71	0.74	0.60	0.70	0.32	0.46	0.72	0.71
20×20														
0	0.05	0.06	0.06	0.05	0.05	0.06	0.06	0.06	0.05	0.06	0.06	0.06	0.06	0.06
0.1	0.43	0.42	0.42	0.43	0.40	0.42	0.42	0.45	0.35	0.43	0.16	0.25	0.43	0.42
0.2	0.90	0.90	0.89	0.90	0.87	0.90	0.89	0.90	0.80	0.88	1.00	1.00	0.90	0.89
0.3	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.98	0.99	1.00	1.00	1.00	1.00
30×30														
0	0.05	0.06	0.06	0.05	0.05	0.06	0.06	0.06	0.06	0.06	1.00	1.00	0.06	0.07
0.1	0.71	0.70	0.70	0.71	0.67	0.70	0.70	0.72	0.58	0.68	1.00	1.00	0.71	0.70
0.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00
0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
40×40														
0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.04	0.06	1.00	1.00	0.06	0.06
0.1	0.89	0.89	0.89	0.89	0.86	0.89	0.89	0.90	0.79	0.87	1.00	1.00	0.89	0.89
0.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 2.3: Proportion of rejected null hypotheses for the robust tests and Moran's tests from the 5000 replicates of each setting in the simulation study (Gaussian case with rook contiguity on a grid).

Grid	Normality tests			Randomisation tests				Permutations tests						
Gaussian (queen) ρ	I	MM	REWLSE	I	rank	MM	REWLSE	I	Dray	rank	LTS	S	MM	REWLSE
10×10														
0	0.05	0.06	0.06	0.05	0.06	0.06	0.06	0.06	0.05	0.06	0.06	0.06	0.06	0.06
0.1	0.16	0.17	0.16	0.16	0.15	0.17	0.16	0.16	0.13	0.15	0.10	0.12	0.16	0.16
0.2	0.35	0.35	0.34	0.35	0.33	0.35	0.34	0.34	0.29	0.32	0.17	0.22	0.33	0.32
0.3	0.58	0.58	0.55	0.58	0.55	0.58	0.55	0.58	0.50	0.55	0.26	0.36	0.56	0.54
20×20														
0	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.06	0.06	0.06	0.06	0.06
0.1	0.29	0.30	0.29	0.29	0.29	0.30	0.29	0.31	0.26	0.30	0.15	0.20	0.29	0.29
0.2	0.72	0.71	0.70	0.72	0.69	0.71	0.70	0.72	0.64	0.69	0.28	0.44	0.70	0.69
0.3	0.96	0.95	0.95	0.96	0.94	0.95	0.95	0.96	0.92	0.95	0.47	0.70	0.95	0.95
30×30														
0	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.07	0.06
0.1	0.49	0.49	0.49	0.49	0.47	0.49	0.49	0.50	0.43	0.48	0.19	0.28	0.50	0.49
0.2	0.94	0.94	0.93	0.94	0.92	0.94	0.93	0.94	0.90	0.92	0.41	0.64	0.93	0.93
0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.68	0.91	1.00	1.00
40×40														
0	0.05	0.05	0.06	0.05	0.05	0.05	0.06	0.06	0.05	0.06	0.06	0.06	0.06	0.06
0.1	0.69	0.69	0.68	0.69	0.66	0.69	0.68	0.70	0.62	0.67	0.22	0.34	0.68	0.67
0.2	1.00	0.99	1.00	1.00	0.99	0.99	1.00	1.00	0.99	0.99	0.54	0.79	0.99	0.99
0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.84	0.98	1.00	1.00

Table 2.4: Proportion of rejected null hypotheses for the robust tests and Moran's tests from the 5000 replicates of each setting in the simulation study (Gaussian case with queen contiguity on a grid).

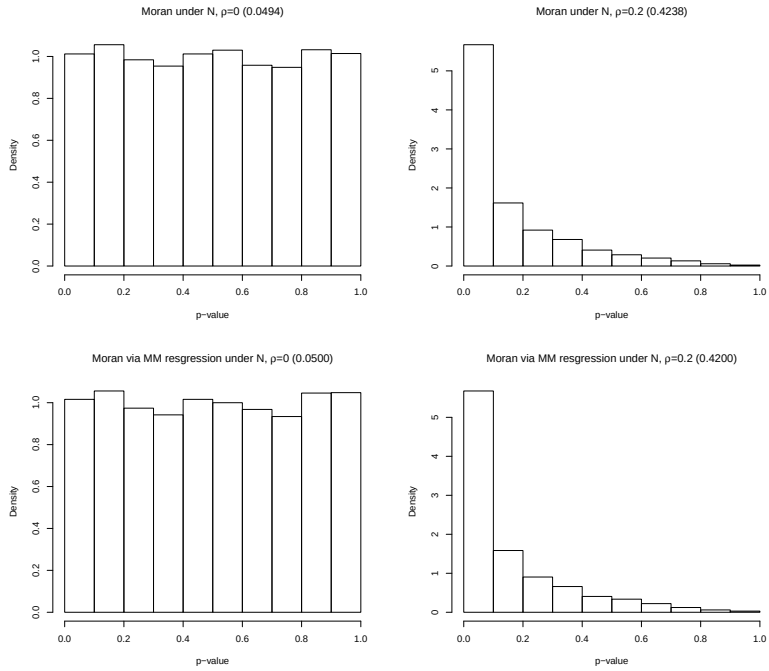
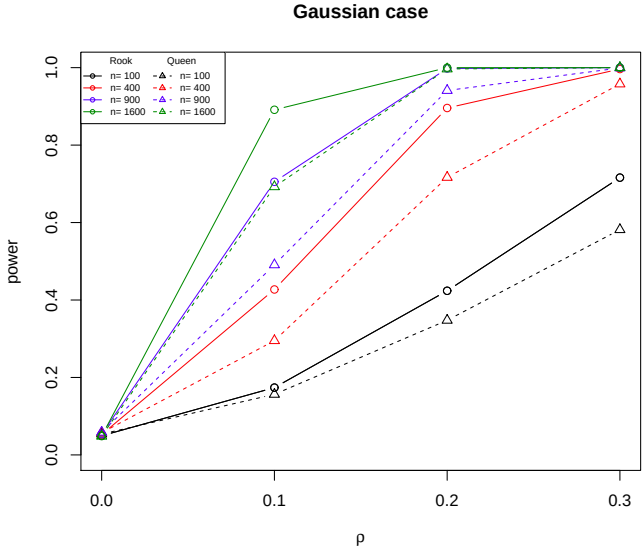


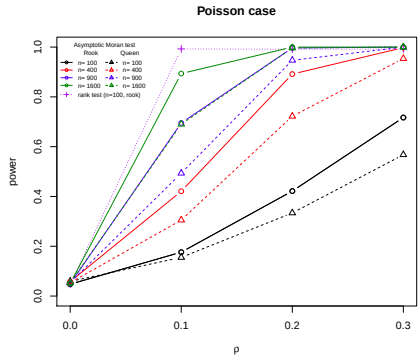
Figure 2.10: Histograms of the p-values from the simulation study with a 10×10 lattice and rook contiguity using Gaussian distribution. The number in parentheses in the title of each plot is the proportion of rejected null hypotheses. The left panels correspond to data generated under the null hypothesis ($\rho = 0$) while the right ones are based on $\rho = 0.2$. The illustrated tests are classic Moran under normality assumption (upper panels) and robust Moran via MM-estimator (lower panels).

Belgium	Normality tests			Randomisation tests				Permutations tests				
ρ	I	MM	REWLSE	I	rank	MM	REWLSE	I	Dray	rank	MM	REWLSE
Gaussian												
0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.04	0.05	0.06	0.06
0.01	0.21	0.20	0.20	0.21	0.19	0.20	0.20	0.22	0.18	0.21	0.21	0.21
0.02	0.52	0.51	0.51	0.52	0.48	0.51	0.51	0.53	0.46	0.49	0.52	0.51
0.03	0.81	0.80	0.79	0.81	0.78	0.80	0.79	0.82	0.74	0.78	0.80	0.79
0.04	0.96	0.95	0.95	0.96	0.94	0.95	0.95	0.96	0.92	0.94	0.95	0.95
0.05	1.00	1.00	0.99	1.00	0.99	1.00	0.99	0.99	0.99	0.99	0.99	0.99
Poisson												
0				0.06	0.06	0.06	0.06	0.06	0.05	0.06	0.06	0.06
0.01				0.21	1.00	0.20	0.20	0.23	0.19	1.00	0.22	0.22
0.02				0.52	1.00	0.50	0.49	0.53	0.46	1.00	0.51	0.50
0.03				0.81	1.00	0.79	0.78	0.81	0.74	1.00	0.80	0.79
0.04				0.96	1.00	0.95	0.94	0.95	0.92	1.00	0.95	0.94
0.05				0.99	1.00	0.99	0.99	0.99	0.99	1.00	0.99	0.99
Bernoulli												
0				0.05	0.05	0.06	0.06	0.06	0.05	0.06	0.05	0.06
0.01				0.21	1.00	0.22	0.21	0.23	0.19	1.00	0.21	0.21
0.02				0.51	1.00	0.51	0.51	0.52	0.45	1.00	0.50	0.50
0.03				0.82	1.00	0.80	0.80	0.82	0.74	1.00	0.80	0.79
0.04				0.96	1.00	0.96	0.95	0.96	0.93	1.00	0.95	0.95
0.05				1.00	1.00	0.99	0.99	1.00	0.99	1.00	0.99	0.99

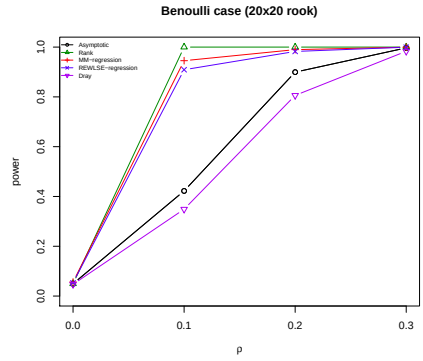
Table 2.5: Proportion of rejected null hypotheses for the robust tests and Moran's tests from the 5000 replicates of each setting in the simulation study. The tests based on normality assumption are only applied on Gaussian observations.



(a) Gaussian case



(b) Poisson case



(c) Bernoulli case

Figure 2.11: Power curves for Moran's tests using simulations on a regular grid.

		Randomisation tests				Permutations tests				
	Grid - rook ρ	I	rank	MM	REWLSE	I	Dray	rank	MM	REWLSE
Poisson	10×10									
	0	0.05	0.05	0.05	0.05	0.06	0.05	0.06	0.06	0.05
	0.1	0.18	0.99	0.18	0.17	0.19	0.15	0.99	0.19	0.19
	0.2	0.42	0.99	0.42	0.41	0.44	0.34	0.99	0.44	0.43
	0.3	0.72	0.99	0.71	0.70	0.73	0.61	0.99	0.72	0.70
	20×20									
	0	0.06	0.06	0.06	0.06	0.07	0.05	0.07	0.07	0.07
	0.1	0.42	1.00	0.42	0.40	0.44	0.34	1.00	0.43	0.43
	0.2	0.89	1.00	0.88	0.87	0.89	0.79	1.00	0.89	0.88
	0.3	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00
	30×30									
	0	0.05	0.05	0.05	0.04	0.06	0.05	0.06	0.06	0.05
	0.1	0.70	1.00	0.68	0.66	0.71	0.58	1.00	0.70	0.70
	0.2	1.00	1.00	1.00	0.99	0.99	0.98	1.00	1.00	1.00
	0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	40×40									
	0	0.05	0.05	0.05	0.04	0.06	0.05	0.06	0.06	0.06
	0.1	0.89	1.00	0.89	0.88	0.90	0.78	1.00	0.89	0.90
	0.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Bernoulli	10×10									
	0	0.06	0.06	0.06	0.07	0.07	0.05	0.07	0.06	0.06
	0.1	0.17	1.00	0.83	0.74*	0.19	0.15	1.00	0.83	0.80
	0.2	0.43	1.00	0.89	0.83*	0.45	0.36	1.00	0.89	0.87
	0.3	0.71	1.00	0.94	0.91*	0.73	0.61	1.00	0.94	0.92
	20×20									
	0	0.05	0.05	0.06	0.05	0.06	0.05	0.06	0.06	0.06
	0.1	0.42	1.00	0.95	0.91*	0.45	0.35	1.00	0.95	0.93
	0.2	0.90	1.00	0.99	0.98*	0.90	0.81	1.00	0.99	0.99
	0.3	1.00	1.00	1.00	1.00*	1.00	0.98	1.00	1.00	1.00
	30×30									
	0	0.05	0.05	0.06	0.05	0.06	0.05	0.06	0.06	0.06
	0.1	0.71	1.00	0.98	0.97*	0.72	0.58	1.00	0.98	0.98
	0.2	1.00	1.00	1.00	1.00*	1.00	0.98	1.00	1.00	1.00
	0.3	1.00	1.00	1.00	1.00*	1.00	1.00	1.00	1.00	1.00
	40×40									
	0	0.05	0.05	0.06	0.05	0.06	0.06	0.06	0.06	0.06
	0.1	0.90	1.00	1.00	0.99*	0.90	0.80	1.00	1.00	0.99
	0.2	1.00	1.00	1.00	1.00*	1.00	1.00	1.00	1.00	1.00
	0.3	1.00	1.00	1.00	1.00*	1.00	1.00	1.00	1.00	1.00

Table 2.6: Proportion of rejected null hypotheses for the robust tests and Moran's tests from the 5000 replicates of each setting in the simulation study (Poisson and Bernoulli case with rook contiguity on a grid). The symbol * means that all simulations did not converge to a solution (around 1200 simulations over 5000 did not converge).

A Appendix

A.1 Contaminated indexes

The following Lemmas are useful to deduce the empirical influence functions of p-values and the resistance of tests. We detail here how the indexes behave under additive contaminations.

Lemma 2.1.1. *Let A be a m-subset of $\{1,\dots,n\}$. We consider an additive contamination $\xi \in \mathbb{R}$ at m locations $\{s_i : i \in A\}$. We denote $\mathbf{z}^* = \mathbf{z} + \xi \sum_{i\in A} e_i$ the contaminated vector of observed values. Moran's index and Geary's ratio of the contaminated vector $\mathbf{z}^*$ are given by*

$$I(\mathbf{z}^*) = \frac{n}{S_0}\frac{P(\xi)}{Q(\xi)} \text{ and } c(\mathbf{z}^*) = \frac{n-1}{S_0}\frac{R(\xi)}{Q(\xi)}$$

where $P(\xi)$, $Q(\xi)$ and $R(\xi)$ are quadratic polynomials in ξ. More precisely,

$$\begin{cases} P(\xi) = & \xi^2\left(w_A - 2\frac{m}{n}(w_A+w_B) + \frac{m^2}{n^2}S_0\right) + 2\xi\sum_{j=1}^{n}\left(\sum_{i\in A} w_{ij} - \frac{m}{n}w_{j\bullet}\right)(z_j-\bar{z}) \\ & +\mathbf{z}'HWH\mathbf{z} \\ Q(\xi) = & \xi^2\frac{m(n-m)}{n} + 2\xi\sum_{i\in A}(z_i-\bar{z}) + \mathbf{z}'H\mathbf{z} \\ R(\xi) = & \xi^2 w_B + 2\xi\sum_{i\in A}\sum_{j=1}^{n} w_{ij}(z_i-z_j) + \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} w_{jk}(z_j-z_k)^2. \end{cases}$$

where $w_A = \sum_{i,j\in A} w_{ij}$ and $w_B = \sum_{i\in A}\sum_{j\notin A} w_{ij}$. Moreover, it follows that

$$I(\mathbf{z}^*) \to \frac{n^2 w_A - 2mn(w_A+w_B) + m^2 S_0}{S_0 m(n-m)} \text{ and } c(\mathbf{z}^*) \to \frac{n(n-1)w_B}{S_0 m(n-m)}$$

if $\xi \to \pm\infty$.

Proof. The proof is direct knowing that the weighting matrix W is symmetric and has a zero-diagonal. □

Remark 2.1.2. *When the contamination is large enough, the limit values only depend on the weighting matrix and the chosen locations while it is independent of the initial vector $\mathbf{z}$.*

Remark 2.1.3. *If there is a unique contamination, i.e. $\mathbf{z}^* = \mathbf{z} + \xi e_i$, Moran's index and Geary's ratio of the contaminated vector $\mathbf{z}^*$ are still defined by the ratio of quadratic polynomials in ξ (in this case, $m = 1$, $w_A = w_{ii} = 0$ and $w_B = w_{i\bullet}$). In particular, $I(\mathbf{z}^*)$ and $c(\mathbf{z}^*)$ converge respectively to $(S_0 - 2nw_{i\bullet})/(S_0(n-1))$ and $nw_{i\bullet}/S_0$ when $|\xi|$ increases.*

Lemma 2.1.4. *Under the same conditions as above, Getis and Ord's statistic of the contaminated vector* $\mathbf{z}^* = \mathbf{z} + \xi \sum_{i\in A} e_i$ *is a ratio of two quadratic polynomials in* ξ*. More precisely,*

$$G(\mathbf{z}^*) = \frac{w_A\xi^2 + 2\xi \sum_{i\in A}\sum_{k=1}^{n} w_{ik}z_k + \mathbf{z}'W\mathbf{z}}{m(m-1)\xi^2 + 2\xi\left(mn\bar{z} - \sum_{i\in A} z_i\right) + \mathbf{z}'B\mathbf{z}}$$

where $w_A = \sum_{i,j\in A} w_{ij}$*.*

Proof of Lemma 2.1.4. The proof is direct knowing that W and B are symmetric and $B_{ij} = 1$ outside the diagonal. □

Remark 2.1.5. *If a unique location is contaminated,* $G(\mathbf{z}^*)$ *is reduced to a ratio of first degree polynomials in* ξ *(*$m = 1$*,* $w_A = w_{ii} = 0$*). If we consider two contaminated locations which are not neighbours (*$m = 2$ *and* $w_A = w_{ij} + w_{ji} = 0$*), then the limit of* $G(\mathbf{z}^*)$ *is zero if* $\xi \to \pm\infty$*.*

A.2 Empirical influence functions of the classic tests

Proposition 2.1.6. *Let* $\mathbf{z}$ *be the n-dimensional column vector of observed values at location* $\{s_1, \ldots, s_n\}$ *and* W *the weighting matrix associated with the domain. The empirical influence function of the p-value of unilateral tests based on asymptotic normality for Geary's ratio is explicitly given by*

$$EIF(\xi, i; c) = n\left[\Phi\left(-\frac{\mathbb{E}[c] - c(\mathbf{z} + \xi e_i)}{\sigma[c(\mathbf{z} + \xi e_i)]}\right) - \Phi\left(-\frac{\mathbb{E}[c] - c(\mathbf{z})}{\sigma[c(\mathbf{z})]}\right)\right]$$

where Φ *is the cdf of the Gaussian distribution,* $\mathbb{E}[c] = 1$ *and the standard deviation* $\sigma[c(\mathbf{z})]$ *is explicit under normality and randomisation assumptions (see Table 2.1).*

The proof is straightforward by noticing that p-value($\mathbf{z}$) for the unilateral asymptotic test is defined by $\Phi\left[-(\mathbb{E}[c] - c(\mathbf{z}))/\sigma[c(\mathbf{z})]\right]$. Moreover, under normality assumption, the variance is constant and explicit (see Table 2.1). Therefore, as ξ tends to infinity, we get

$$EIF(\xi, i; c) \to n\left[\Phi\left(\frac{nw_{i\bullet} - S_0}{S_0\sigma[c]}\right) - \Phi\left(-\frac{\mathbb{E}[c] - c(\mathbf{z})}{\sigma[c]}\right)\right]. \tag{2.16}$$

The influence functions associated with Geary's ratio are represented Figure 2.12 for the asymptotic tests and the permutation test.

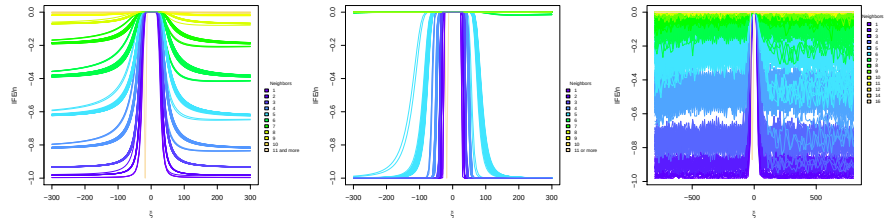


Figure 2.12: Contaminated p-value for a unique contamination using tests based on Geary's ratio using asymptotic test under randomisation assumption (left panel) and normality assumption (middle panel) or permutation tests (right panel).

Proposition 2.1.7. *Let* $\mathbf{z}$ *be the* n*-dimensional column vector of observed values at location* $\{s_1, \ldots, s_n\}$ *and* W *the weighting matrix associated with the domain. The empirical influence function of the p-value of unilateral tests based on asymptotic normality for Getis and Ord's statistic is explicitly given by*

$$EIF(\xi, i; G) = n\left[\Phi\left(-\frac{G(\mathbf{z}+\xi e_i) - \mathbb{E}[G]}{\sigma[G(\mathbf{z}+\xi e_i)]}\right) - \Phi\left(-\frac{G(\mathbf{z}) - \mathbb{E}[G]}{\sigma[G(\mathbf{z})]}\right)\right]$$

where Φ *is the cdf of the Gaussian distribution,* $\mathbb{E}(G) = S_0/(n(n-1))$ *and the standard deviation* $\sigma[G(\mathbf{z})]$ *is explicit under randomisation assumption (see Table 2.1).*

The proof is straightforward by noticing that p-value($\mathbf{z}$) for the unilateral asymptotic test is defined by $\Phi\left[-(G(\mathbf{z}) - \mathbb{E}[G])/\sigma[G(\mathbf{z})]\right]$. Moreover, as ξ tends to infinity, we get

$$EIF(\xi, i; G) \to n\left[\Phi\left(-\frac{\frac{\sum_{k=1}^{n} x_{ik}z_k}{\sum_{j\neq i} z_j} - \frac{S_0}{n(n-1)}}{\sqrt{a(i,\mathbf{z})/b(i,\mathbf{z})}}\right) - \Phi\left(-\frac{G(\mathbf{z}) - \mathbb{E}[G]}{\sigma[G(\mathbf{z})]}\right)\right] \tag{2.17}$$

where $a(i,\mathbf{z}) = 2(S_2 - nS_1)z_i^2 + 2(2S_1 - S_2)m_1(\mathbf{z})z_i + (2(n-1)S_1 - S_2)m_2(\mathbf{z}) + (S_2 - 2S_1)m_1^2(\mathbf{z})$ and $b(i,\mathbf{z}) = n(n-1)(n-2)(4m_1^2(\mathbf{z}) - 8m_1(\mathbf{z})z_i + 4z_i^2)$. We observe that, unlike the two other indexes, the limit value of the contaminated p-value still depends on the initial vector $\mathbf{z}$.

The influence functions associated with Getis and Ord's statistic are represented Figure 2.13 for the asymptotic test under randomisation and the permutation test.

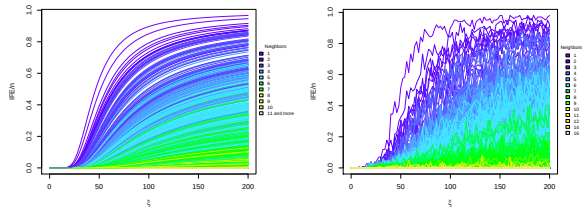


Figure 2.13: Contaminated p-value for a unique contamination using tests based on Getis and Ord's statistic using the asymptotic test under randomisation assumption (left panel) or permutation test (right panel).

A.3 Resistance of the classic tests

Resistance to acceptance for asymptotic tests

Proof of Proposition 2.4.2 - Moran. Let us first show that Moran's tests may fail to reject the null hypothesis as soon as one sufficiently extreme value is observed on a location corresponding to an row weight satisfying $w_{i\bullet} \geq S_0/n$. If it exists, we may choose a location with an "average" row weight, i.e., $w_{i\bullet} \approx S_0/n$. If the weighting matrix is row-standardized, each location can play the role as, by definition, $w_{i\bullet} = 1$ for all $i \in \{1, \ldots, n\}$.

Firstly, it has been shown that the expectation and variance of the statistic I is constant under the *normality assumption*. The corresponding test rejects the null hypothesis as soon as the ratio $(I(\mathbf{z}) - E[I])/\sigma_N[I]$ is larger than the Gaussian quantile $z_{1-\alpha}$.

Consider now a corrupted data set $\mathbf{z}^*$ corresponding to $m = 1$ contaminated observation located at s_i, for an arbitrary i, i.e., $z_j^* = z_j$ for all $j \neq i$ and $z_i^* = z_i + \xi$ with $\xi \in \mathbb{R}$. By Lemma 2.1.1,

$$\lim_{\xi \to \infty} \frac{I(\mathbf{z}^*) - E[I]}{\sigma_N[I]} = \frac{2(S_0 - n w_{i\bullet})}{(n-1) S_0 \sigma_N[I]}.$$

If the location is chosen such that $w_{i\bullet} \geq S_0/n$, then the limiting value of the statistic is negative or nil, and for ξ large enough, it will always be smaller than $z_{1-\alpha}$, leading to the non-rejection of H_0, whatever the values of $z_1, \ldots, z_n$. This proves that the resistance to acceptance is $1/n$.

Under *randomisation assumption*, the variance of Moran's index is modified. In this case, the variance depends on the observed values $\mathbf{z}$ by means of the kurtosis $b_2(\mathbf{z}) = \mu_4(\mathbf{z})/\mu_2^2(\mathbf{z})$ and can be written $a_1 - a_2 b_2(\mathbf{z})$ for a_1 and a_2 some constants defined by the weighting matrix (see Table 2.1 for details). The kurtosis of the

contaminated data set $\mathbf{z}^* = \mathbf{z} + \xi e_i \sum_{i\in A} e_i$ can be expressed as the ratio of two fourth-degree polynomials in ξ. More precisely, the numerator of $b_2(\mathbf{z}^*)$ is

$$\begin{aligned}\mu_4(\mathbf{z}^*) =& \frac{\xi^4}{n^4} m(n-m)(n^2-3nm+3m^2) + \frac{4\xi^3}{n^3}(n^2-3nm+3m^2)\sum_{i\in A}(z_i-\bar{z}) \\ &+ \frac{6\xi^2}{n^2}\left(m\mu_2(\mathbf{z}) + (n-2m)\sum_{i\in A}(z_i-\bar{z})^2\right) + \frac{4\xi}{n}\left(\sum_{i\in A}(z_i-\bar{z})^3 - m\mu_3(\mathbf{z})\right) \\ &+ \mu_4(\mathbf{z})\end{aligned}$$

and its denominator is

$$\begin{aligned}\mu_2^2(\mathbf{z}^*) =& \frac{\xi^4}{n^4}(n-m)^2m^2 + \frac{4\xi^3}{n^3}m(n-m)\sum_{i\in A}(z_i-\bar{z}) \\ &+ \frac{2\xi^2}{n^2}\left(m(n-m)\mu_2(\mathbf{z}) + 2\left(\sum_{i\in A}(z_i-\bar{z})\right)^2\right) \\ &+ \frac{4\xi}{n}\mu_2(\mathbf{z})\sum_{i\in A}(z_i-\bar{z}) + \mu_2^2(\mathbf{z}).\end{aligned}$$

Therefore, for $m = 1$ and $\mathbf{z}^* = z + \xi e_i$, the limit of the statistic is given by

$$\lim_{\xi\to\infty}\frac{I(\mathbf{z}^*) - E[I]}{\sigma_R[I(\mathbf{z}^*)]} = \frac{2(S_0 - nw_{i\bullet})}{(n-1)S_0\sqrt{a_1 - a_2\frac{n^2-3n+3}{n-1}}}.$$

Then, similarly to the normality assumption, we are now able to conclude that the resistance to acceptance is $1/n$ (contaminating a location with a total weight smaller or equal to the mean weight is enough).

The conclusion is similar under randomisation assumption as the variance of Geary's ratio can also be written as $a_1 - a_2 b_2(\mathbf{z})$ for a_1 and a_2 some constants defined by the weighting matrix. □

Proposition 2.1.8. *Let* $\mathbf{z}$ *be the observed values at locations* $\{s_1, \ldots, s_n\}$ *and* W *the corresponding weighting matrix. The resistance to acceptance of the asymptotic tests based on Geary's ratio are* $1/n$.

Proof. The proof for Geary's ratio is similar to Moran's case. Indeed, under the normality assumption, for $\mathbf{z}^* = z + \xi e_i$, Lemma 2.1.1 allows us to write

$$\lim_{\xi\to\infty}\frac{E[c] - c(\mathbf{z}^*)}{\sigma_N[c]} = \frac{S_0 - nw_{i\bullet}}{S_0\sigma_N[c]}.$$

If the location is chosen such that $w_{i\bullet} \geq S_0/n$, then the limiting value of the statistic is negative or nil, and for ξ large enough, it will always be smaller than $z_{1-\alpha}$, leading

to the non-rejection of H_0, whatever the values of $z_1, \ldots, z_n$. This proves that the resistance to acceptance is $1/n$. □

Proposition 2.1.9. *Let* $\mathbf{z}$ *be the observed values at locations* $\{s_1, \ldots, s_n\}$ *and* W *the corresponding weighting matrix. The resistance to acceptance of the asymptotic test based on Getis and Ord's statistic is* $1/n$.

Proof. We fail to reject the null hypothesis for a vector $\mathbf{z}$ if

$$\frac{G(\mathbf{z}) - E_R[G]}{\sigma_R[G(\mathbf{z})]} < z_{1-\alpha}.$$

By Lemma 2.1.4 for $m = 1$,

$$G(\mathbf{z} + \xi e_i) = \frac{2\xi \sum_{k=1}^{n} w_{ik} z_k + \mathbf{z}' W \mathbf{z}}{2\xi \sum_{k \neq i} z_k + \mathbf{z}' B \mathbf{z}}.$$

If there exist a location s_i such that $\sum_{k=1}^{n} w_{ik} z_k = S_0/(n(n-1)) \sum_{k \neq i} z_k$, contaminate this location with a large ξ is enough to always fail to reject the null hypothesis as

$$\lim_{\xi \to \infty} G(\mathbf{z} + \xi e_i) = \frac{S_0}{n(n-1)} = E[G].$$

If the equality is not verified for a location s_i, the following contamination ξ does the job:

$$\xi = \frac{S_0 z' B z - n(n-1) z' W z}{2\left(n(n-1) \sum_{k=1}^{n} w_{ik} z_k - S_0 \sum_{k \neq i} z_k\right)}.$$

This is enough to prove that the resistance to acceptance is $1/n$ for Getis and Ord's statistic as an adequate contamination on any location is enough to always fail to reject the null hypothesis. □

Resistance to rejection for asymptotic tests

Proof of Proposition 2.4.3 - Moran. The proof is immediate by Lemma 2.1.1 and the definition of resistance to rejection. Indeed, under normality assumption, if there exist a subset A of $\{1, \ldots, n\}$ such that

$$\lim_{\xi \to \infty} I(\mathbf{z}^*) \geq \frac{-1}{n-1} + \sigma_N(I) z_{1-\alpha} \tag{2.18}$$

for all values of $\mathbf{z}$, then m observations suffice to produce rejection. Similarly, under randomisation assumption, the limit of the statistic is given by

$$\lim_{\xi \to \infty} \frac{I(\mathbf{z}^*) - E[I]}{\sigma_R[I(\mathbf{z}^*)]} = \frac{\frac{n^2 w_A - 2mn(w_A + w_B) + m^2 S_0}{S_0 m(n-m)} + \frac{1}{n-1}}{\sqrt{a_1 - a_2 \frac{n^2 - 3nm + 3m}{(n-m)m}}}.$$

□

Proposition 2.1.10. *Let* $\mathbf{z}$ *be the observed values at locations* $\{s_1, \ldots, s_n\}$ *and* W *the corresponding weighting matrix. The resistance to rejection of asymptotic tests based on Geary's ratio is* m/n *where* m *is the size of the smallest subset* $A \subseteq \{1, \ldots, n\}$ *which satisfies*

$$\frac{n(n-1)w_B}{S_0 m(n-m)} \leq 1 - \sigma_N[c] z_{1-\alpha} \tag{2.19}$$

for tests based on normality assumption or which satisfies

$$\frac{S_0 m(n-m) - n(n-1)w_B}{S_0 m(n-m)\sqrt{a_1' - a_2' \frac{n^2-3nm+3m}{(n-m)m}}} \geq z_{1-\alpha} \tag{2.20}$$

for tests based on randomisation assumption. We denote $w_B = \sum_{i \in A} \sum_{j \notin A} w_{ij}$. *The constants* a_1' *and* a_2' *are defined according to the expression of the variance of Geary's ratio under randomisation, i.e.,* $V_R[c(\mathbf{z})] = a_1' - a_2' b_2(\mathbf{z})$ *(see Table 2.1 for details).*

Proof. Similarly to Moran's case, the proof is direct by Lemma 2.1.1. □

Proposition 2.1.11. *Let* $\mathbf{z}$ *be the observed values at locations* $\{s_1, \ldots, s_n\}$ *and* W *the corresponding weighting matrix. The resistance to rejection of the asymptotic test based on Getis and Ord's statistic is* m/n *where* $m > 1$ *is the smallest size of a subset* $A \subseteq \{1, \ldots, n\}$ *which satisfies*

$$\frac{w_A}{m(m-1)} - \frac{S_0}{n(n-1)} \geq z_{1-\alpha} \sqrt{\frac{m_{[4]} S_0^2 + (n-m)_{[2]} S_1 + (n-m)(m-2) S_2}{n_{[4]} m^2 (1-m)^2} - \frac{S_0^2}{n^2 (n-1)^2}}$$

where $w_A = \sum_{i,j \in A} w_{ij}$ *and* $a_{[k]} = a(a-1)\ldots(a-k+1)$.

Proof. If there exists a subset A of m locations such that

$$\lim_{\xi \to \infty} \frac{G(\mathbf{z}^*) - \mathbb{E}[G]}{\sigma_R[G(\mathbf{z}^*)]} \geq z_{1-\alpha} \tag{2.21}$$

for all values of $\mathbf{z}$, then m observations suffice to produce rejection.

If $m = 1$, the variance based on the contaminated vector $\mathbf{z}^* = \mathbf{z} + \xi e_i$ can be rewritten as $\frac{P(\xi)}{Q(\xi)} - \frac{S_0^2}{n^2(n-1)^2}$ where $P(\xi)$ and $Q(\xi)$ are quadratic polynomials in ξ. If we write $P(\xi) = p_2(\mathbf{z})\xi^2 + p_1(\mathbf{z})\xi + p_0(\mathbf{z})$ and $Q(\xi) = q_2(\mathbf{z})\xi^2 + q_1(\mathbf{z})\xi + q_0(\mathbf{z})$, then

$$\lim_{\xi \to \infty} \frac{G(\mathbf{z} + \xi e_i) - E_R[G]}{\sigma_R[G(\mathbf{z} + \xi e_i)]} = \frac{\frac{\sum_{j=1}^n w_{ij} z_j}{\sum_{j \neq i} z_j} - \frac{S_0}{n(n-1)}}{\sqrt{\frac{p_2(\mathbf{z})}{q_2(\mathbf{z})} - \frac{S_0^2}{n^2(n-1)^2}}}$$

which still depends on the initial vector $\mathbf{z}$. Therefore, the condition (2.21) is not satisfied for any vector $\mathbf{z}$ and the resistance to rejection is larger than $1/n$. If $m > 1$, the variance based on the contaminated vector $\mathbf{z}^* = \mathbf{z} + \xi \sum_{i \in A} e_i$ can be rewritten as $\frac{P(\xi)}{Q(\xi)} - \frac{S_0^2}{n^2(n-1)^2}$ where $P(\xi)$ and $Q(\xi)$ are polynomials of degree 4 in ξ. In that case, by Lemma 2.1.4, the condition (2.21) is equivalent to

$$\frac{w_A}{m(m-1)} - \frac{S_0}{n(n-1)}$$
$$\geq z_{1-\alpha} \sqrt{\frac{m_{[4]} S_0^2 + (n-m)_{[2]} S_1 + (n-m)(m-2) S_2}{n_{[4]} m^2 (m-1)^2} - \frac{S_0^2}{m^2 (n-1)^2}}$$

which is enough to conclude. □

Permutation test

Under the contamination $\mathbf{z}^* = \mathbf{z} + \xi e_i$, Moran's index and Geary's ratio respectively converge to $(S_0 - 2nw_{i\bullet})((n-1)S_0)^{-1}$ and $nw_{i\bullet}/S_0$. Under a permutation τ of $\{1, \ldots, n\}$, the corresponding Moran's index and Geary's ratio will respectively converge to $(S_0 - 2nw_{\tau(i)\bullet})/((n-1)S_0)$ and $nw_{\tau(i)\bullet}/S_0$. Therefore, for a large enough contamination ξ, the indexes based on contaminated permuted datasets are independent of the initial values $z_1, \ldots, z_n$. Then, the rank and the pseudo p-value given by the permutation test only depend on the permutation τ and the weighting matrix. The resistance to acceptance and the resistance to rejection is $1/n$.

Dray test (Dray, 2011)

This test is based on Moran's index which is decomposed into two parts $S_I^+(\mathbf{z})$ and $S_I^-(\mathbf{z})$ as in Equation (2.9). The pseudo p-value is calculated as for permutation test. We easily observe that the eigenvectors u_k of the matrix HWH do not depend on the values of $\mathbf{z}$.

Under the contamination $\mathbf{z}^* = \mathbf{z} + \xi \sum_{i \in A} e_i$, the contribution of positive spatial autocorrelation $S_I^+(\mathbf{z}^*)$ and of negative spatial autocorrelation $S_I^-(\mathbf{z}^*)$ respectively converge to

$$\sum_{I(u_k) > E[I]} I(u_k) \frac{\sum_{i \in A} ((u_k)_i - \bar{u_k})^2}{m(n-m) s_{u_k}^2} \quad \text{and} \quad \sum_{I(u_k) < E[I]} I(u_k) \frac{\sum_{i \in A} ((u_k)_i - \bar{u_k})^2}{m(n-m) s_{u_k}^2}$$

which depend only on the contaminated locations s_i and the eigen-decomposition of HWH. Moreover, for any permuted version of $\mathbf{z}^*$, the limits of $S_I^+(\tau(\mathbf{z}^*))$ and $S_I^-(\tau(\mathbf{z}^*))$ for $\xi \to \infty$, are the same limits as above. Therefore, like for permutation test, the pseudo p-value only depends on the permutation τ and the weighting matrix, which leads to a resistance to acceptance and to rejection equals to $1/n$.

A.4 Resistance of robust tests on regular grid

In the queen contiguity, each cell has 3, 5, or 8 neighbours. With the cell it self, each contamination will modify at most 9 cells (3×3). Therefore, if we consider m neighbours of maximal size without overlap, one needs $m \geq n/18$ to recover half of the observations. It is easy to deduce that there are $(\lfloor a/3 \rfloor)^2$ disjoint blocks of 3 cells by 3 in a grid $a \times a$. Moreover, $(\lfloor a/3 \rfloor)^2 > a^2/18$ for any $a > 5$. For the first cases, one can directly observe than one contamination is enough for $a = 3$ or 4 and two contaminations as in Figure 2.14 is enough for $a = 5$. This is enough to prove the assertion.

In the rook contiguity, each cell has 2, 3 or 4 neighbours. With the cell it self, each contamination will modify at most 5 cells. Therefore, if we consider m neighbours of maximal size without overlap, one needs $m \geq n/10$ to recover half of the observations. We find an iterative selection of locations to contaminate which prove that $m = \lceil n/10 \rceil$ is enough. The iterative selection is represented Figure 2.15 where we conserve the lower left corner and successively add lines and columns at the right and top of the grid.

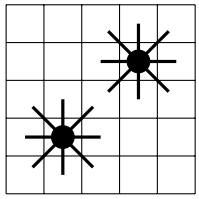


Figure 2.14: Selection of the subset A for queen contiguity to cover half of the grid 5×5.

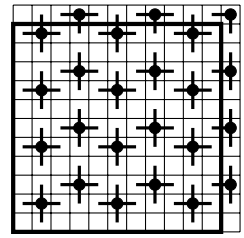


Figure 2.15: Iterative selection of the subset A for rook contiguity to cover half of the grid.

General conclusion and perspectives

Some conclusions have already been outlined at the end of each chapter. To conclude the first part of the doctoral thesis, we would like to point out here some guidelines for new directions for future research around the developed thematics.

In complement to Chapter 1, a non parametric detection technique could be emphasized using depth functions. The notion of depth was introduced by Zuo and Serfling (2000a) and has been successfully applied in many different contexts such as the detection of outliers (e.g., Chen et al., 2008, Dang and Serfling, 2010), discrimination (see for instance Cui et al., 2008), etc. Roughly speaking, the empirical depth of an observation characterizes how central it is with respect to the other data points. An observation lying close to the "center" of the data should have the highest depth while all observations lying at the "border" of the data cloud should correspond to the lowest possible depth. In the univariate case, the median is clearly the deepest point in the data, the depth of the other observations being smaller and smaller as the distance to the median increases. Several depth functions have been suggested in the literature. We may mention some common statistical depths such as half-space depth (Tukey, 1975), simplicial depth (Liu, 1990), projection depth (Zuo and Serfling, 2000a,b) and regression depth (Rousseeuw and Hubert, 1999). As mentioned in Chapter 1, a global outlier is an observation which lies at the border of the data (or even further away) and should therefore correspond to a low depth with respect to the center of the data cloud. Local outliers tend to have a low depth with respect to their neighbours. To measure how deep a neighbour is from an observation, the concept of symmetrized neighbourhood as introduced by Paindaveine and Van Bever (2013) may be useful. Therefore, the central observation x_i, the deepest point, may be compared to its neighbours. In the symmetrized dataset, if the proportion of its neighbours which are too far according to other observations is larger than a fixed value, the observation x_i is considered as a local outlier. This suggested methodology still needs to be completed in order to be efficient, due to, among other issues, the computational cost of depth functions in high dimensions (see for instance Zuo, 2019).

Moreover, a non parametric way to define the most autocorrelated neighbourhoods needs to be developed.

Chapter 2 proposed robust alternatives to tests based on Moran's index. Robust alternative may also be constructed using Geary's ratio or Getis and Ord's statistic, either using rank values or more sophisticated constructions. Piegorsch and Bailer (2005, p 298) observed that Geary's ratio can be defined as the ratio between an empirical variogram and the empirical variance. On the other hand, as developed in Genton (1998b), the variogram is estimated using an empirical covariance over $\mathcal{V}_n$, the sample of differences for all pairs of locations at fixed distance. An idea would be to combine those two results using a robust estimation of the variogram and the variance. For instance, Genton (1998a) (suggestion followed by Lark (2008) and Miranda and de Miranda (2011)) proposes estimating the variogram using the robust scale estimator Q_n introduced by Rousseeuw and Croux (1993). Therefore, an estimation of Geary's ratio could be given by the ratio of Q_n estimated over $\mathcal{V}_n$ and Q_n estimated over the observed vector. The efficiency and robust properties of this estimation could be studied. Then, similarly to Moran's index, this robust version of Geary's ratio could be plugged into asymptotic or permutation tests.

As mentioned on page 56, the global spatial autocorrelation can be decomposed into local associations for each observation. These local indicators are also well used in the literature as diagnostic tools (Droesbeke et al., 2006) to interpret spatial clusters or to detect spatial outliers (see for instance McGrath and Zhang, 2003, Fu et al., 2014, Zhang et al., 2008). The robust properties of these indicators could also be studied. Moreover, current research is still in progress in order to extend these indicators to the multivariate setting. For instance, Anselin (2019) defines a multivariate indicator as the sum of the individual local statistics for each variable.

Finally, additional research could be performed in order to determine the exact resistance of tests based on rank Moran's index. Some links with graph theory could be emphasized. Indeed, Dijkstra's algorithm could be used to optimize the search of minimal path between vertices of a graph. In this case, each vertex corresponds to a rank vector and there is an edge if a rank vector is obtained by the contamination of the other vector. However, to the best of our knowledge, the complexity of such a problem does not allow to use it easily as the best complexity is of order (number of edges + number of vertices) log(number of vertices), i.e. of order $n! \log(n!)$.

PART II

Stein's method

Motivation: binomiality

The developments of the previous chapter can be generalized to the multivariate case by considering multiple testing. Generally, we consider repeated testing of null hypotheses $(H_{0,i})_{i=1,\dots,n_p}$ on non-necessarily-independent data sets $\mathbf{x}_i = (x_{i1}, \dots, x_{in_s})$ for $i = 1, \dots, n_p$, each decision being made on the basis of a statistic $t_i = t_i(x_{i1}, \dots, x_{in_s})$. We introduce the random variables

$$\mathbf{X} = \begin{pmatrix} X_{11} & \dots & X_{1n_s} \\ \vdots & \ddots & \vdots \\ X_{n_p1} & \dots & X_{n_pn_s} \end{pmatrix} \text{ and } \mathbf{T} = \begin{pmatrix} T_1 \\ \vdots \\ T_{n_p} \end{pmatrix}$$

from which $\mathbf{x}$ (and therefore $\mathbf{t}$) are sampled; under our assumption the columns $\mathbf{X}_{\cdot j}$, $j = 1, \dots, n_s$ are independent but not necessarily the lines $\mathbf{X}_{i\cdot}$, $i = 1, \dots, n_p$. Similarly, $\mathbf{T}$ does not necessarily have independent components. The initial question that occurred to us was to determine the impact of the dependence in multiple testing and how it could be measured.

It is well-known that, under $H_{0,i}$, a p-value based on an independent sample $\mathbf{x}_i$ is a realization of a uniform random variable on $[0, 1]$ (see for instance the discussion in Section 2.6.2 or Murdoch et al., 2008). Suppose that each T_i is a p-value for some test. If the different tests are independent then $\mathbf{T}$ is uniformly distributed on $[0, 1]^{n_p}$, an illustration is provided on the left panel of Figure 2.16. On the other hand, if the lines of $\mathbf{X}$ are not independent, then although the margins of $\mathbf{T}$ remain uniform, the overall distribution is not and is unknown (see more developments in Wang, 2014). This situation is represented on the right panel of Figure 2.16.

One of the most studied issues of multiple testing settings is the so-called False Discovery Rate (FDR, for short) which is defined as the expected proportion of errors among the rejected hypotheses (Benjamini and Hochberg, 1995); any statistical procedure on the whole data set $\mathbf{X}$ must always be devised in such a way that the procedures each *individually* have correct level and power but also *overall* guarantee that FDR $\leq \alpha$ for some α small. The founding references in this regard are Benjamini

Figure 2.16: Scatterplot of p-values based on tests constructed from independent lines (left panel) and dependent lines (right panel). The marginal distributions are represented by histograms.

and Hochberg (1995), Benjamini and Yekutieli (2001) (the first dealing with the case where the T_i are independent and the second with the case where they are dependent). However, their proposals are omnibus techniques without specificity of the dependence structure in the date.

Under specific assumptions on the data and the tests one can also devise adhoc more powerful procedures e.g. by adapting the tests t_i to take the FDR into account (as in Cai and Liu, 2016). In order to obtain a global information on the entire data set, Leek and Storey (2008) transforms the data $\mathbf{X}$ to remove the inter-line dependence. The FDR-based approaches all end up in combining one-dimensional summary statistics in order to obtain a general conclusion.

These considerations lead to rephrase the initial question using simpler objects as the indicator of rejection for each test. In this case, we want to study the distribution of $\sum_{i=1}^{n} X_i$ when $X_1, \ldots, X_n$ are not-necessarily-independent and identically distributed Bernoulli random variables. If the random variables are independent and identically distributed, their sum is obviously distributed as a Binomial; without the identical distribution, we find the so called *Poisson binomial distribution*. The k-runs is a particular case of dependence between indicators even if they are identically distributed. Explicit laws could also be written as a sum of correlated indicators, see for instance the beta binomial and hypergeometric distributions.

This problematic was the initial motivation of the second part of the thesis, where we study discrete, and more generally, univariate distributions by means of Stein's method. Therefore, we firstly focus on Stein's method applied to discrete distributions. This interest leads us to other questions which are related to this method, such

as covariance identities. First order inequalities are obtained using mostly the inverse Stein operators when infinite expansions are mainly obtained using a probabilistic Lagrange identity. Then, we use our formalism to provide new representations of solutions to Stein equations. This leads to the study of Stein factors and distances between distributions. The topic of my research has clearly changed over time. Indeed, at the end, we mainly studied Stein's method and we did not get the chance to go back to the initial problematic, as mentioned at the end of the thesis.

CHAPTER 3

Stein differentiation

1 Introduction

Stein's method consists in a collection of techniques for distributional approximation that was originally developed for normal approximation in Stein (1972) and for Poisson approximation in Chen (1975); for expositions see the books Stein (1986), Barbour and Chen (2005a,b), Chen et al. (2011), Nourdin and Peccati (2012) and the review papers Reinert (2004), Ross (2011), Chatterjee (2014a). Outside the Gaussian and Poisson frameworks, there exist several non-equivalent general theories allowing to setup Stein's method for many probability distributions, of which we single out the papers Chatterjee and Shao (2011), Döbler (2015), Upadhye et al. (2017), Xu (2019), Chen et al. (2018) for univariate distributions under analytical assumptions, Arras and Houdré (2019a,b) for infinitely divisible distributions, Barbour et al. (2018) for discrete multivariate distributions, and Mackey and Gorham (2016), Gorham et al. (2019), Gorham and Mackey (2017) as well as Fang et al. (2018) for multivariate densities under diffusive assumptions.

The backbone of the present work consists in the approach from Ley et al. (2017a,b), Reinert et al. (2018). Before introducing these results, we fix the notations. Let $\mathcal{X} \in \mathcal{B}(\mathbb{R})$ and equip it with some σ-algebra $\mathcal{A}$ and σ-finite measure μ. Let X be a random variable on $\mathcal{X}$, with induced probability measure $\mathbb{P}^X$ which is absolutely continuous with respect to μ; we denote by p the corresponding probability density, and its support by $\mathcal{S}(p) = \{x \in \mathcal{X} : p(x) > 0\}$. As usual, $L^1(p)$ is the collection of all real valued functions f such that $\mathbb{E}|f(X)| < \infty$. We sometimes call the expectation under p the p-mean. Although we could in principle keep the discussion to come very general, in order to make the chapter more concrete and readable we

shall restrict our attention to distributions satisfying the following Assumption.

Assumption A. The measure μ is either the counting measure on $\mathcal{X} = \mathbb{Z}$ or the Lebesgue measure on $\mathcal{X} = \mathbb{R}$. If μ is the counting measure then there exist $a<b \in \mathbb{Z} \cup \{-\infty, \infty\}$ such that $\mathcal{S}(p) = [a, b] \cap \mathbb{Z}$. If μ is the Lebesgue measure then there exist $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$ such that $\mathcal{S}(p)^{\circ} =]a, b[$ and $\overline{\mathcal{S}(p)} = [a, b]$. Moreover, the measure μ is not point mass.

Here not allowing point mass much simplifies the presentation. Stein's method for point mass is available in Reinert (1995).

Let $\ell \in \{-1, 0, 1\}$. In the sequel we shall restrict our attention to the following three derivative-type operators:

$$\Delta^{\ell} f(x) \quad = \begin{cases} f'(x), & \text{if } \ell = 0; \\ (f(x+\ell) - f(x))/\ell & \text{if } \ell \in \{-1, +1\}, \end{cases}$$

with $f'(x)$ the weak derivative defined Lebesgue almost everywhere, $\Delta^{+1} (\equiv \Delta^{+})$ the classical forward difference and $\Delta^{-1} (\equiv \Delta^{-})$ the classical backward difference. Whenever $\ell = 0$ we take μ as the Lebesgue measure and speak of the *continuous case*; whenever $\ell \in \{-1, 1\}$ we take μ as the counting measure and speak of the *discrete case*. There are two choices of derivatives in the discrete case, only one in the continuous case. We let $\mathrm{dom}(\Delta^{\ell})$ denote the collection of functions $f : \mathbb{R} \to \mathbb{R}$ such that $\Delta^{\ell} f(x)$ exists and is finite μ-almost surely. In the case $\ell = 0$, this corresponds to all absolutely continuous functions; in the case $\ell = \pm 1$ the domain is the collection of all functions on $\mathbb{Z}$. For ease of reference we note that, if $f \in \mathrm{dom}(\Delta^{\ell})$ is such that $\Delta^{\ell} f \, \mathbb{I}[a, b] \in L^1(\mu)$ then, for all c, d such that $a \le c \le d \le b$ we have

$$\int_c^d \Delta^{\ell} f(x) \, \mu(\mathrm{d}x) = \begin{cases} \int_c^d f'(x) \mathrm{d}x = f(d) - f(c) & \text{if } \ell = 0 \\ \sum_{j=c}^d \Delta^{-} f(x) = f(d) - f(c-1) & \text{if } \ell = -1 \\ \sum_{j=c}^d \Delta^{+} f(x) = f(d+1) - f(c) & \text{if } \ell = +1 \end{cases}$$

which we summarize as

$$\int_c^d \Delta^{\ell} f(x) \mu(\mathrm{d}x) = f(d + a_{\ell}) - f(c - b_{\ell}) \tag{3.1}$$

where

$$a_{\ell} = \mathbb{I}[\ell = 1] \text{ and } b_{\ell} = \mathbb{I}[\ell = -1]. \tag{3.2}$$

We stress the fact that the values at c, d are understood as limits if either is infinite.

2 Stein operators and Stein equations

Our first definitions come from Ley et al. (2017b). We first define $\mathrm{dom}(p, \Delta^\ell)$ as the collection of $f : \mathbb{R} \to \mathbb{R}$ such that $f\,p \in \mathrm{dom}(\Delta^\ell)$.

Definition 3.2.1 (Canonical Stein operators)**.** *Let $f \in \mathrm{dom}(p, \Delta^\ell)$ and consider the linear operator $f \mapsto \mathcal{T}_p^\ell f$ defined as*

$$\mathcal{T}_p^\ell f(x) = \frac{\Delta^\ell(f(x)p(x))}{p(x)}$$

for all $x \in \mathcal{S}(p)$ and as $\mathcal{T}_p^\ell f(x) = 0$ for $x \notin \mathcal{S}(p)$. The operator $\mathcal{T}_p^\ell$ is called the canonical (ℓ-)Stein operator of p. The cases $\ell = 1$ and $\ell = -1$ provide the forward and backward Stein operators, denoted by $\mathcal{T}_p^+$ and $\mathcal{T}_p^-$, respectively; the case $\ell = 0$ provides the differential Stein operator denoted by $\mathcal{T}_p$.

To describe the domain and the range of $\mathcal{T}_p^\ell$ we introduce the following sets of functions:

$$\begin{aligned}
\mathcal{F}^{(0)}(p) &= \left\{ f \in L^1(p) : \mathbb{E}[f(X)] = 0 \right\}; \\
\mathcal{F}_\ell^{(1)}(p) &= \big\{ f \in \mathrm{dom}(p, \Delta^\ell) : \Delta^\ell(fp)\mathbb{I}[\mathcal{S}(p)] \in L^1(\mu) \\
&\qquad \text{and } \int_{\mathcal{S}(p)} \Delta^\ell(fp)(x)\,\mu(\mathrm{d}x) = \mathbb{E}\left[\mathcal{T}_p^\ell f(X)\right] = 0 \Big\}.
\end{aligned}$$

We draw the reader's attention to the fact that the second condition in the definition of $\mathcal{F}_\ell^{(1)}(p)$ can be rewritten as

$$f(b + a_\ell)p(b + a_\ell) = f(a - b_\ell)p(a - b_\ell).$$

The particular choice of the constant function $f(x) = 1$ leads to the score function. This function plays a crucial role in the sequel, notably in Chapter 6.

Definition 3.2.2 (The score function)**.** *The (ℓ-)score function of probability density p is the function*

$$\rho_p^\ell(x) = \mathcal{T}_p^\ell 1(x) = \frac{\Delta^\ell p(x)}{p(x)}.$$

Example 3.2.3. *If $p(x) = \phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$ is the standard Gaussian density, then $\ell = 0$ and $\mathcal{T}_p^0 f(x) = f'(x) - xf(x)$ is the classical operator for this distribution. The set $\mathcal{F}_0^{(1)}(\phi)$ consists of all functions such $f'(x) - xf(x) \in L^1(\phi)$ and $\lim_{x\to-\infty} f(x)\phi(x) = \lim_{x\to\infty} f(x)\phi(x)$.*

If $p(x) = p_\lambda(x) = e^{-\lambda}\lambda^x/x!$ *is the Poisson density on* $\mathbb{N}$ *then* $\ell \in \{-1, 1\}$*;* $\mathcal{T}_p^+ f(x) = \lambda f(x+1)/(x+1) - f(x)$ *for all* $x \in \mathbb{N}$ *and 0 otherwise, and* $\mathcal{T}_p^- f(x) = f(x) - xf(x-1)/\lambda$ *for all* $x \in \mathbb{N}$ *and 0 otherwise. Both are equivalent, up to scaling, to the classical operator* $\lambda f(x+1) - xf(x)$ *from Chen (1975) for this distribution. The set* $\mathcal{F}_+^{(1)}(p_\lambda)$ *consists of all functions such that* $\lambda f(x+1)/(x+1) - f(x) \in L^1(p_\lambda)$ *and* $\lim_{x\to\infty} f(x)p_\lambda(x) = f(0)p_\lambda(0)$*. Similarly,* $\mathcal{F}_-^{(1)}(p_\lambda)$ *consists of all functions such that* $f(x) - xf(x-1)/\lambda \in L^1(p_\lambda)$ *and* $\lim_{x\to\infty} f(x)p_\lambda(x) = 0$*.*

The next lemma, which follows immediately from the definition of $\mathcal{T}_p^\ell f$ and of the different sets of functions, shows why $\mathcal{F}_\ell^{(1)}(p)$ is called the *canonical Stein class.*

Lemma 3.2.4 (Canonical Stein class)**.** *For* $f \in \mathcal{F}_\ell^{(1)}(p)$*,* $\mathcal{T}_p^\ell f \in \mathcal{F}^{(0)}(p)$ *.*

Crucially for the results in this work, for all $f \in \mathrm{dom}(\Delta^\ell)$, $g \in \mathrm{dom}(\Delta^{-\ell})$ such that $f(\cdot)g(\cdot - \ell) \in \mathrm{dom}(\Delta^\ell)$ the operators Δ^ℓ satisfy the product rule

$$\Delta^\ell\big(f(x)g(x-\ell)\big) = (\Delta^\ell f(x))g(x) + f(x)\Delta^{-\ell}g(x) \tag{3.3}$$

for all $\ell \in \{-1, 0, 1\}$. This product rule leads to an integration by parts (IBP) formula (a.k.a. Abel-type summation formula) as follows.

Lemma 3.2.5 (Stein IBP formula - version 1)**.** *For all* $f \in \mathrm{dom}(p, \Delta^\ell)$ *and* $g \in \mathrm{dom}(\Delta^{-\ell})$ *such that (i)* $f(\cdot)g(\cdot - \ell) \in \mathcal{F}_\ell^{(1)}(p)$ *and (ii)* $f(\cdot)\Delta^{-\ell}g(\cdot) \in L^1(p)$ *we have*

$$\mathbb{E}\left[(\mathcal{T}_p^\ell f(X))g(X)\right] = -\mathbb{E}\left[f(X)\Delta^{-\ell}g(X)\right]. \tag{3.4}$$

Proof. Under the stated assumptions, we can apply (3.3) to get

$$\mathcal{T}_p^\ell\big(f(x)g(x-\ell)\big) = (\mathcal{T}_p^\ell f(x))g(x) + f(x)(\Delta^{-\ell}g(x)) \tag{3.5}$$

for all $x \in \mathcal{S}(p)$. Condition (i) in the statement guarantees that the left hand side (l.h.s.) of (3.5) has mean 0, while condition (ii) guarantees that we can separate the expectation of the sum on the right hand side (r.h.s.) into the sum of the individual expectations. □

A natural interpretation of (3.4) is that operator $\mathcal{T}_p^\ell$ is, in some sense to be made precise, the *skew-adjoint* operator to $\Delta^{-\ell}$ with respect to the scalar product $\langle f, g\rangle = \mathbb{E}\left[f(X)g(X)\right]$; this provides a supplementary justification to the use of the terminology "canonical" for operator $\mathcal{T}_p^\ell$. We discuss a consequence of this interpretation in Section 3. The conditions under which Lemma 3.2.5 holds are all but transparent. We clarify these assumptions in Section 4. For more details on Stein class and operators, we refer to Ley et al. (2017b) for the construction in an abstract setting and Ley et al. (2017a) for the construction in the continuous setting (i.e. $\ell = 0$) Multivariate extensions are developed in Reinert et al. (2018).

The fundamental stepping stone for our theory is an inverse of the canonical operator $\mathcal{T}_p^\ell$ provided in the next definition.

Definition 3.2.6 (Canonical pseudo inverse Stein operator)**.** *Let $\ell \in \{-1, 0, 1\}$ and recall the notations a_ℓ, b_ℓ from (3.2). The canonical pseudo-inverse Stein operator $\mathcal{L}_p^\ell$ for the operator $\mathcal{T}_p^\ell$ is defined, for $h \in L^1(p)$, as*

$$\mathcal{L}_p^\ell h(x) = \frac{1}{p(x)} \int_a^{x-a_\ell} (h(u) - \mathbb{E}[h(X)])p(u)\mu(\mathrm{d}u) \tag{3.6}$$

$$= \frac{1}{p(x)} \int_{x+b_\ell}^{b} (\mathbb{E}[h(X)] - h(u))p(u)\mu(\mathrm{d}u) \tag{3.7}$$

for all $x \in \mathcal{S}(p)$ and $\mathcal{L}_p^\ell h(x) = 0$ for all $x \notin \mathcal{S}(p)$.

Equality between the expressions in (3.6) and in (3.7) is justified because $h \in L^1(p)$ so that the integral of $h(\cdot) - \mathbb{E}[h(X)]$ over the whole support cancels out. For ease of reference we detail $\mathcal{L}_p^\ell$ in the three cases that interest us:

$$\begin{aligned}
\mathcal{L}_p^0 h(x) &= \frac{1}{p(x)} \int_a^x (h(u) - \mathbb{E}[h(X)])p(u)\mathrm{d}u \\
&= \frac{1}{p(x)} \int_x^b (\mathbb{E}[h(X)] - h(u))p(u)\mathrm{d}u && (\ell = 0) \\
\mathcal{L}_p^- h(x) &= \frac{1}{p(x)} \sum_{j=a}^{x} (h(j) - \mathbb{E}[h(X)])p(j) \\
&= \frac{1}{p(x)} \sum_{j=x+1}^{b} (\mathbb{E}[h(X)] - h(j))p(j) && (\ell = -1) \\
\mathcal{L}_p^+ h(x) &= \frac{1}{p(x)} \sum_{j=a}^{x-1} (h(j) - \mathbb{E}[h(X)])p(j) \\
&= \frac{1}{p(x)} \sum_{j=x}^{b} (\mathbb{E}[h(X)] - h(j))p(j) && (\ell = 1).
\end{aligned}$$

Note that $\mathcal{L}_p^- h(b) = 0$ but $\mathcal{L}_p^- h(a) = h(a) - \mathbb{E}[h(X)]$ and, conversely, $\mathcal{L}_p^+ h(a) = 0$ but $\mathcal{L}_p^+ h(b) = \mathbb{E}[h(X)] - h(b)$. The denomination *pseudo-inverse-Stein operator* for $\mathcal{L}_p^\ell$ is justified by the following lemma whose proof is immediate.

Lemma 3.2.7. *For any $h \in L^1(p)$, $\mathcal{L}_p^\ell h \in \mathcal{F}_\ell^{(1)}(p)$. Moreover, (i) for all $h \in L^1(p)$ we have $\mathcal{T}_p^\ell \mathcal{L}_p^\ell h(x) = h(x) - \mathbb{E}[h(X)]$ at all $x \in \mathcal{S}(p)$ and (ii) for all $f \in \mathcal{F}_\ell^{(1)}(p)$ we have $\mathcal{L}_p^\ell \mathcal{T}_p^\ell f(x) = f(x) - f(a^+ - b_\ell)p(a^+ - b_\ell)/p(x) = f(x) - f(b^- + a_\ell)p(b^- + a_\ell)/p(x)$ at all $x \in \mathcal{S}(p)$. Operator $\mathcal{L}_p^\ell$ is invertible (with inverse $\mathcal{T}_p^\ell$) on the subclass of functions in $\mathcal{F}^{(0)}(p) \cap \mathcal{F}^{(1)}(p)$ which, moreover, satisfy $f(b^- + a_\ell)p(b^- + a_\ell) = f(a^+ - b_\ell)p(a^+ - b_\ell) = 0$.*

Starting from (3.5) we postulate the next definition.

Definition 3.2.8 (Standardizations of the canonical operator)**.** *Fix $\ell \in \{-1, 0, 1\}$ and $\eta \in L^1(p)$. The η-standardized Stein operator is*

$$\begin{aligned}\mathcal{A}_p^{\ell,\eta} g(x) &= \mathcal{T}_p^{\ell}\big(\mathcal{L}_p^{\ell}\eta(\cdot)g(\cdot-\ell)\big)(x) \\ &= \big(\eta(x) - \mathbb{E}[\eta(X)]\big)g(x) + \mathcal{L}_p^{\ell}\eta(x)\big(\Delta^{-\ell}g(x)\big) \end{aligned} \tag{3.8}$$

acting on the collection $\mathcal{F}(\mathcal{A}_p^{\ell,\eta})$ of test functions g such that $\mathcal{L}_p^{\ell}\eta(\cdot)g(\cdot-\ell) \in \mathcal{F}_{\ell}^{(1)}(p)$ and $(\mathcal{L}_p^{\ell}\eta)\Delta^{-\ell}g \in L^1(p)$.

Example 3.2.9. *If $p = \phi$ is the standard normal density and $\eta(x) = x$ then $\mathcal{L}_{\phi}^{0}\eta(x) = -1$. More generally if $\eta(x) = H_k(x) = (-1)^k e^{x^2/2}(d^k/dx^k)e^{-x^2/2}$ is the kth Hermite polynomial so that $H_{k+1}(x) = xH_k(x) - H_k'(x)$ then $\mathcal{L}_{\eta}^{0}\eta(x) = -H_{k-1}(x)$. This leads to the family of standardized Stein operators $\mathcal{A}_{\phi}^{0,k}g(x) = H_k(x)g(x) - H_{k-1}(x)g'(x)$ already considered e.g. in Goldstein and Reinert (2005).*

If $p = p_{\lambda}$ and $\eta(x) = x$ then $\mathcal{L}_p^{+}\eta(x) = -x$ and $\mathcal{L}_p^{-}\eta(x) = -\lambda$. This leads to the standardized operators $\mathcal{A}^{+}g(x) = (x-\lambda)g(x) - x\Delta^{-}g(x) = \lambda g(x) - xg(x-1)$ and $\mathcal{A}^{-}g(x) = (x-\lambda)g(x) - \lambda\Delta^{+}g(x) = -\lambda g(x+1) + xg(x)$; both are equivalent to the classical operator $\lambda g(x+1) - xg(x)$ first identified by Chen (1975). Similarly as for the Gaussian one could introduce the appropriate family of orthogonal polynomials (here the Charlier polynomials) and propose an entire family of operators; we refer to Goldstein and Reinert (2005) for an overview.

Remark 3.2.10. *The conditions appearing in the definition of $\mathcal{F}(\mathcal{A}_p^{\ell,\eta})$ are tailored to ensure that all identities and manipulations follow immediately. For instance, the requirement that $\mathcal{L}_p^{\ell}\eta(\cdot)g(\cdot-\ell) \in \mathcal{F}_{\ell}^{(1)}(p)$ in the definition of $\mathcal{F}(\mathcal{A}_p^{\ell,\eta})$ guarantees that the resulting functions $\mathcal{A}_p^{\ell,\eta}g(x)$ have p-mean 0 and the condition $(\mathcal{L}_p^{\ell}\eta)\Delta^{-\ell}g \in L^1(p)$ guarantees that the expectations of the individual summands on the r.h.s. of* (3.8) *exist. Again, our assumptions are not transparent; we discuss them in detail in Section 4.*

The final ingredient for Stein differentiation is the *Stein equation*:

Definition 3.2.11 (Stein equation)**.** *Fix $\ell \in \{-1, 0, 1\}$ and $\eta \in L^1(p)$. For $h \in L^1(p)$, the $\mathcal{A}_p^{\ell,\eta}$-Stein equation for h is the functional equation $\mathcal{A}_p^{\ell,\eta}g(x) = h(x) - \mathbb{E}[h(X)], x \in \mathcal{S}(p)$, i.e.*

$$\big(\eta(x) - \mathbb{E}[\eta(X)]\big)g(x) + \mathcal{L}_p^{\ell}\eta(x)\big(\Delta^{-\ell}g(x)\big) = h(x) - \mathbb{E}[h(X)],\ x \in \mathcal{S}(p). \tag{3.9}$$

A solution to the Stein equation is any function $g \in \mathcal{F}(\mathcal{A}_p^{\ell,\eta})$ which satisfies (3.9) *for all $x \in \mathcal{S}(p)$.*

Our notations lead immediately to the next result.

Lemma 3.2.12 (Solution to the Stein equation)**.** *Fix $\eta \in L^1(p)$. The Stein equation (3.9) for $h \in L^1(p)$ is solved by*

$$g_h^{p,\ell,\eta}(x) = \frac{\mathcal{L}_p^\ell h(x+\ell)}{\mathcal{L}_p^\ell \eta(x+\ell)} \tag{3.10}$$

with the convention that $g_h^{p,\ell,\eta}(x) = 0$ for all $x + \ell$ outside of $\mathcal{S}(p)$.

Proof. With $g = g_h^{p,\ell,\eta}$,

$$\begin{aligned}\mathcal{A}_p^{\ell,\eta} g(x) &= \mathcal{T}_p^\ell\big(\mathcal{L}_p^\ell \eta(\cdot) g(\cdot - \ell)\big)(x) = \mathcal{T}_p^\ell\left(\mathcal{L}_p^\ell \eta(\cdot)\frac{\mathcal{L}_p^\ell h(\cdot)}{\mathcal{L}_p^\ell \eta(\cdot)}\right)(x) = \mathcal{T}_p^\ell\left(\mathcal{L}_p^\ell h(\cdot)\right)(x) \\ &= h(x) - \mathbb{E}[h(X)]\end{aligned}$$

using Lemma 3.2.7 for the last step. Hence (3.9) is satisfied for all $x \in \mathcal{S}(p)$. Since, by construction, $g \in \mathcal{F}(\mathcal{A}_p^{\ell,\eta})$, the claim follows. □

When the context is clear then we drop the superscripts and the subscript in g of (3.10). Before proceeding we provide two examples. The notation Id refers to the identity function $x \to \mathrm{Id}(x) = x$.

Example 3.2.13 (Binomial distribution)**.** *Let $0 < \theta < 1$ and $p(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ be the binomial density with parameters (n, θ) and $\mathcal{S}(p) = [0, n] \cap \mathbb{N}$; assume that $0 < \theta < 1$. Stein's method for the binomial distribution was first developed in Ehm (1991) using Δ^-; see also Soon (1996), Holmes (2004).*

Picking $\ell = 1$, the class $\mathcal{F}_+^{(1)}(p)$ consists of functions $f : \mathbb{Z} \to \mathbb{R}$ which are bounded on $\mathcal{S}(p)$ and $f(0) = 0$. Fixing $\eta(x) = x - n\theta$ gives $\mathcal{L}_{\mathrm{bin}(n,\theta)}^+ \eta(x) = -(1-\theta)x$ leading to

$$\mathcal{A}_{\mathrm{bin}(n,\theta)}^{+,\mathrm{Id}} g(x) = (x - n\theta)g(x) - (1-\theta)x\Delta^- g(x) \tag{3.11}$$

with corresponding class $\mathcal{F}\left(\mathcal{A}_{\mathrm{bin}(n,\theta)}^{+,\mathrm{Id}}\right)$ which contains all functions $g : \mathbb{Z} \to \mathbb{R}$. The solution to the $\mathcal{A}_{\mathrm{bin}(n,\theta)}^{+,\mathrm{Id}}$-Stein equation (see (3.9)) is

$$g^+(x) = \frac{-1}{(1-\theta)(x+1)p(x+1)}\sum_{j=0}^{x}(h(j) - \mathbb{E}[h(X)])p(j) \text{ for all } 0 \le x \le n-1$$

and $g^+(n) = 0$.

Picking $\ell = -1$, the class $\mathcal{F}_-^{(1)}(p)$ consists of functions $f : \mathbb{Z} \to \mathbb{R}$ which are bounded on $\mathcal{S}(p)$ and such that $f(n) = 0$. Again fixing $\eta(x) = x - n\theta$ gives $\mathcal{L}^-_{\mathrm{bin}(n,\theta)}\eta(x) = -\theta(n-x)$ leading to

$$\mathcal{A}^{-,\mathrm{Id}}_{\mathrm{bin}(n,\theta)}g(x) = (x-n\theta)g(x) - \theta(n-x)\Delta^+ g(x) \tag{3.12}$$

acting on the same class as (3.11). The solution to the $\mathcal{A}^{-,\mathrm{Id}}_{\mathrm{bin}(n,\theta)}$-Stein equation is

$$g^-(x) = \frac{-1}{\theta(n-(x-1))p(x-1)} \sum_{j=0}^{x-1} (h(j) - \mathbb{E}[h(X)])p(j) \text{ for all } 1 \le x \le n$$

and $g^-(0) = 0$. The function $-g^-$ is studied in Ehm (1991) where bounds on $\|\Delta^- g^-\|$ are provided (see equation (10) in that paper); see also Section 5 where bounds on $\|g^-\|$ are provided.

Example 3.2.14 (Beta distribution)**.** *Let $p(x) = x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha,\beta)$ be the beta density with parameters (α,β) and $\mathcal{S}(p) = (0,1)$. Stein's method for the beta distribution was developed in Goldstein and Reinert (2013), Döbler (2015) using the Stein operator $\mathcal{A}f(x) = x(1-x)f'(x) + (\alpha(1-x) - \beta x)f(x)$. In our notations, we have $\ell = 0$ and $\mathcal{F}_0^{(1)}(p)$ consists of functions $f : \mathbb{R} \to \mathbb{R}$ such that $f(0^+)p(0^+) = f(1^-)p(1^-)$ and $|(fp)'|$ is Lebesgue integrable on $[0,1]$. Fixing $\eta(x) = x - \frac{\alpha}{\alpha+\beta}$ gives $\mathcal{L}_{\mathrm{beta}(\alpha,\beta)}\eta(x) = -x(1-x)/(\alpha+\beta)$ leading to the operator*

$$\mathcal{A}^{\mathrm{Id}}_{\mathrm{Beta}(\alpha,\beta)}g(x) = \Big(x - \frac{\alpha}{\alpha+\beta}\Big)g(x) - \frac{x(1-x)}{\alpha+\beta}g'(x)$$

with domain $\mathcal{F}\left(\mathcal{A}^{\mathrm{Id}}_{\mathrm{Beta}(\alpha,\beta)}\right)$ the set of differentiable functions $g : \mathbb{R} \to \mathbb{R}$ such that $x(1-x)g(x) \in \mathcal{F}_0^{(1)}(p)$ and $x(1-x)g'(x) \in L^1(p)$. The solution to the $\mathcal{A}^{\mathrm{Id}}_{\mathrm{Beta}(\alpha,\beta)}$-Stein equation is

$$g(x) = \frac{-(\alpha+\beta)}{x(1-x)p(x)} \int_0^x (h(u) - \mathbb{E}[h(X)])p(u)du, \quad x \in (0,1).$$

The operator $\mathcal{A}^{\mathrm{Id}}_{\mathrm{Beta}(\alpha,\beta)}f$ is, up to multiplication by $\alpha+\beta$, the classical Stein operator $\mathcal{A}f$ for the beta density, see Goldstein and Reinert (2013), Döbler (2015) for details and bounds on solutions and their derivatives. See also Section 5 where bounds on $\|g\|$ are provided.

In order to propose a more general example, we recall the concept of a Stein kernel, here extended to continuous and discrete distributions alike.

Definition 3.2.15 (The Stein kernel)**.** *Let $X \sim p$ have finite mean. The (ℓ-)Stein kernel of X (or of p) is the function*

$$\tau_p^\ell(x) = -\mathcal{L}_p^\ell(\mathrm{Id})(x).$$

Metonymously, we refer to the random variable $\tau_p^\ell(X)$ as the (ℓ-)Stein kernel of X.

Remark 3.2.16. *The function $\tau_p^\ell(\cdot)$ is studied in detail for $\ell = 0$ in Stein (1986, Lecture VI). This function is particularly useful for Pearson (and discrete Pearson a.k.a. Ord) distributions which are characterized by the fact that their Stein kernel τ_p^ℓ is a second degree polynomial, see Example 4.2.8. For more on this topic, we also refer to Chapter 5 as well as Courtade et al. (2019), Fathi (2019, 2018) wherein important contributions to the theory of Stein kernels are provided in a multivariate setting.*

The next example gives some (ℓ-)Stein kernels, exploiting the fact that if the mean of X is ν, then $\mathcal{L}_p^\ell(\mathrm{Id})(x) = \mathcal{L}_p^\ell(\mathrm{Id} - \nu)(x)$.

Example 3.2.17. *If $X \sim \mathrm{Bin}(n, \theta)$ then using $\eta(x) = x - n\theta$, Example 3.2.13 gives $\tau^+_{\mathrm{bin}(n,\theta)}(x) = (1-\theta)x$ and $\tau^-_{\mathrm{bin}(n,\theta)}(x) = \theta(n-x)$. If $X \sim \mathrm{Beta}(\alpha, \beta)$ then Example 3.2.14 with $\eta(x) = x - \frac{\alpha}{\alpha+\beta}$ gives $\tau^0_{\mathrm{Beta}(\alpha,\beta)}(x) = x(1-x)/(\alpha+\beta)$.*

Example 3.2.18 (A general example)**.** *Let p satisfy Assumption A and suppose that it has finite mean ν. Fixing $\eta(x) = x - \nu$, operator* (3.8) *becomes*

$$\mathcal{A}_p^{\tau_p^\ell} g(x) = \big(x - \nu\big) g(x) - \tau_p^\ell(x) \Delta^{-\ell} g(x)$$

with corresponding class $\mathcal{F}\big(\mathcal{A}_p^{\tau_p^\ell}\big)$ which contains all functions $g : \mathbb{R} \to \mathbb{R}$ such that $\tau_p^\ell(\cdot) g(\cdot - \ell) \in \mathcal{F}_\ell^{(1)}(p)$ and $\tau_p^\ell \Delta^{-\ell} g \in L^1(p)$. Again, we stress that such conditions are clarified in Section 4. Using Lemma 3.2.12, the solution to the $\mathcal{A}_p^{\tau_p}$ Stein equation is

$$g_{\mathrm{Id}}^{p,\ell,h}(x) = \frac{-\mathcal{L}_p^\ell h(x+\ell)}{\tau_p^\ell(x+\ell)}.$$

Bounds on $\|g\|$ are provided in Section 4. Stein's method based on $\mathcal{A}_p^{\tau_p^\ell}$ is already available in several important subcases, e.g. in Schoutens (2001), Kusuoka and Tudor (2012), Döbler (2015) for continuous distributions.

The construction is tailored to ensure that all operators have mean 0 over the entire classes of functions on which they are defined. We immediately deduce the following family of Stein integration by parts formulas:

Lemma 3.2.19 (Stein IBP formula - version 2)**.** *Let $X \sim p$. Then*

$$\mathbb{E}\left[-\{\mathcal{L}_p^\ell f(X)\}\Delta^{-\ell} g(X)\right] = \mathbb{E}\left[(f(X) - \mathbb{E}[f(X)])g(X)\right] \tag{3.13}$$

for all $f \in L^1(p)$, $g \in \mathrm{dom}(\Delta^{-\ell})$ such that $\mathcal{L}_p^\ell f(\cdot) g(\cdot - \ell) \in \mathcal{F}_\ell^{(1)}(p)$ and $\mathcal{L}_p^\ell f \Delta^{-\ell} g \in L^1(p)$.

Proof. Identity (3.13) follows directly from the Stein product rule in Ley et al. (2017b, Theorem 3.24) or by using the fact that expectations of the operators in (3.8) are equal to 0. □

For our future developments it is important to note that in the formulation of Lemma 3.2.19 the test functions f and g do not play a symmetric role. If $g \in L^1(p)$ then the right hand side of (3.13) is the covariance $\mathrm{Cov}(f(X), g(X))$. We shall use this heavily in our future developments. Similarly as for Lemma 3.2.5, the conditions under which Lemma 3.2.19 applies are not transparent in their present form. In Section 4 various explicit sets of conditions are provided under which the IBP (3.13) is applicable.

3 Representations of the inverse Stein operator

This section contains the first main results, namely probabilistic representations for the inverse Stein operator (see Definition 3.2.6). Such representations are extremely useful for manipulations of the operators. We start with a simple rewriting of $\mathcal{L}_p^\ell$. Given $\ell \in \{-1, 0, 1\}$, recall the notation $a_\ell = \mathbb{I}[\ell = 1]$ and define

$$\chi^\ell(x, y) = \mathbb{I}[x \le y - a_\ell]. \tag{3.14}$$

Such generalized indicator functions particularize, in the three cases that interest us, to $\chi^0(x, y) = \mathbb{I}[x \le y]$ $\big(\ell = 0\big)$, $\chi^-(x, y) = \mathbb{I}[x \le y]$ $\big(\ell = -1\big)$ and $\chi^+(x, y) = \mathbb{I}[x < y]$ $\big(\ell = 1\big)$. Their properties lead to some form of "calculus" which shall be useful in the sequel.

Lemma 3.3.1 (Chi calculation rules)**.** *The function $\chi^\ell(x, y)$ is non-increasing in x and non-decreasing in y. For all x, y we have*

$$\chi^\ell(x, y) + \chi^{-\ell}(y, x) = 1 + \mathbb{I}[\ell = 0]\mathbb{I}[x = y]. \tag{3.15}$$

Moreover,

$$\chi^\ell(u, y)\chi^\ell(v, y) = \chi^\ell(\max(u, v), y) \text{ and } \chi^\ell(x, u)\chi^\ell(x, v) = \chi^\ell(x, \min(u, v)). \tag{3.16}$$

Let p with support $\mathcal{S}(p)$ satisfy Assumption A. Then for any $f \in L^1(p)$ it is easy to check from the definition 3.2.6 that

$$\begin{aligned} \mathcal{L}_p^\ell f(x) &= \frac{1}{p(x)} \mathbb{E}\left[\chi^\ell(X, x)\big(f(X) - \mathbb{E}[f(X)]\big)\right] \\ &= \frac{1}{p(x)} \mathbb{E}\left[\big(\chi^\ell(X, x) - \mathbb{E}[\chi^\ell(X, x)]\big)\big(f(X) - \mathbb{E}[f(X)]\big)\right]. \end{aligned} \tag{3.17}$$

Next, define

$$\Phi_p^\ell(u, x, v) = \frac{\chi^\ell(u, x)\chi^{-\ell}(x, v)}{p(x)} \tag{3.18}$$

for all $x \in \mathcal{S}(p)$ and 0 elsewhere. This function is used in the following representation formula for the Stein inverse operator:

Lemma 3.3.2 (Representation formula I)**.** *Let X, X_1, X_2 be independent copies of $X \sim p$ with support $\mathcal{S}(p)$. Then, for all $f \in L^1(p)$ we have*

$$-\mathcal{L}_p^\ell f(x) = \mathbb{E}\left[(f(X_2) - f(X_1))\Phi_p^\ell(X_1, x, X_2)\right]. \tag{3.19}$$

Proof. The $L^1(p)$ condition on f suffices for the expectation on the r.h.s. of (3.19) to be finite for all $x \in \mathcal{S}(p)$. Suppose without loss of generality that $\mathbb{E}[f(X)] = 0$. Using that X_1, X_2 are i.i.d., we reap

$$\begin{aligned}
&\mathbb{E}\big[(f(X_2) - f(X_1))\chi^\ell(X_1, x)\chi^{-\ell}(x, X_2)\big] \\
&= \mathbb{E}\big[\chi^\ell(X_1, x)\big]\mathbb{E}[f(X_2)\chi^{-\ell}(x, X_2)] - \mathbb{E}[f(X_1)\chi^\ell(X_1, x)]\mathbb{E}\big[\chi^{-\ell}(x, X_2)\big] \\
&= \mathbb{E}\big[\chi^\ell(X_1, x)\big]\mathbb{E}\big[f(X_2)\big(1 - \chi^\ell(X_2, x)\big)\big] - \mathbb{E}\big[f(X_1)\chi^\ell(X_1, x)\big]\mathbb{E}[\chi^{-\ell}(x, X_2)] \\
&= -\mathbb{E}\big[f(X)\chi^\ell(X, x)\big]\big(\mathbb{E}[\chi^\ell(X, x)] + \mathbb{E}[\chi^{-\ell}(x, X)]\big),
\end{aligned}$$

where in the third line we used the fact that $\mathbb{E}[f(X)\mathbb{I}[\ell = 0]\mathbb{I}[X = x]] = 0$ under the stated assumptions. For the same reasons, we have $\mathbb{E}[\chi^\ell(X, x) + \chi^{-\ell}(x, X)] = 1$ for all $x \in \mathcal{X}$ and all $\ell \in \{-1, 0, 1\}$. The conclusion follows by recalling (3.17). □

The function defined in (3.18) allows to perform "probabilistic integration" as follows: if $f \in \mathrm{dom}(\Delta^{-\ell})$ is such that $(\Delta^{-\ell} f)$ is integrable on $[x_1, x_2] \cap \mathcal{S}(p)$ then

$$f(x_2) - f(x_1) = \mathbb{E}\left[\Phi_p^\ell(x_1, X, x_2)\Delta^{-\ell} f(X)\right] = \begin{cases} \int_{x_1}^{x_2} f'(u)\mathrm{d}u & (\ell = 0) \\ \sum_{j=x_1}^{x_2 - 1} \Delta^+ f(j) & (\ell = -1) \\ \sum_{j=x_1+1}^{x_2} \Delta^- f(j) & (\ell = 1) \end{cases} \tag{3.20}$$

for all $x_1 < x_2 \in \mathcal{S}(p)$. If, furthermore, $f \in L^1(p)$ then (by a conditioning argument)

$$\mathbb{E}[(f(X_2) - f(X_1))\mathbb{I}[X_1 < X_2]] = \mathbb{E}\left[\Phi_p^\ell(X_1, X, X_2)\Delta^{-\ell} f(X)\right].$$

Equation (3.20) leads to the next representation formula for the inverse Stein operator.

Lemma 3.3.3 (Representation formula II)**.** *Let $X \sim p$. Define the kernel K_p on $\mathcal{S}(p) \times \mathcal{S}(p)$ by*

$$K_p^\ell(x, x') = \mathbb{E}[\chi^\ell(X, x)\chi^\ell(X, x')] - \mathbb{E}[\chi^\ell(X, x)]\mathbb{E}\big[\chi^\ell(X, x')\big].$$

Then $K_p^\ell(x, x')$ is symmetric and positive. Moreover, for all $f \in \mathrm{dom}(\Delta^{-\ell})$ such that $f \in L^1(p)$ we have,

$$-\mathcal{L}_p^\ell f(x) = \mathbb{E}\left[\frac{K_p^\ell(X, x)}{p(X)p(x)}\Delta^{-\ell} f(X)\right]. \tag{3.21}$$

Proof. Symmetry of K_p^ℓ is immediate. To see that it is positive, applying first (3.16) and then (3.15),

$$\begin{aligned} K_p^\ell(x,x') &= \mathbb{E}\left[\chi^\ell(X,\min(x,x'))\right]\left(1-\mathbb{E}\left[\chi^\ell(X,\max(x,x'))\right]\right) \\ &= \mathbb{E}\left[\chi^\ell(X,\min(x,x'))\right]\mathbb{E}\left[\chi^{-\ell}(\max(x,x'),X)\right] \end{aligned} \tag{3.22}$$

which is necessarily positive. To prove (3.21), we insert (3.20) into (3.19), to obtain

$$\begin{aligned} -\mathcal{L}_p^\ell f(x) &= \mathbb{E}\left[\Delta^{-\ell} f(X')\Phi_p^\ell(X_1,X',X_2)\Phi_p^\ell(X_1,x,X_2)\right] \\ &= \mathbb{E}\left[\Delta^{-\ell} f(X')\mathbb{E}\left[\Phi_p^\ell(X_1,X',X_2)\Phi_p^\ell(X_1,x,X_2)\,|\,X'\right]\right]. \end{aligned}$$

For all $x,x' \in \mathcal{S}(p)$, by (3.16),

$$\begin{aligned} &\mathbb{E}\left[\Phi_p^\ell(X_1,x,X_2)\Phi_p^\ell(X_1,x',X_2)\right] \\ &= \frac{1}{p(x)p(x')}\mathbb{E}[\chi^\ell(X,x)\chi^\ell(X,x')]\mathbb{E}[\chi^{-\ell}(x,X)\chi^{-\ell}(x',X)] \\ &= \frac{1}{p(x)p(x')}\left(\mathbb{E}\left[\chi^\ell(X,\min(x,x'))\right]\mathbb{E}\left[\chi^{-\ell}(\max(x,x'),X)\right]\right). \end{aligned}$$

Using (3.22), we recognize the kernel $K_p^\ell(x,x')$ in the numerator, and identity (3.21) follows. □

Example 3.3.4. *Representations* (3.19) *and* (3.21) *can easily be applied to obtain representations for the Stein kernel* $\tau_p^\ell(x)$*:*

$$\tau_p^\ell(x) = -\mathcal{L}_p^\ell(\mathrm{Id})(x) = \mathbb{E}\left[(X_2-X_1)\Phi_p^\ell(X_1,x,X_2)\right] = \mathbb{E}\left[\frac{K_p^\ell(X,x)}{p(X)p(x)}\right].$$

In particular the Stein kernel is positive on $\mathcal{S}(p)$*.*

Identity (3.19) seems to be new, although it is present in non-explicit form in Chatterjee and Shao (2011, Equation (4.16)). Representation (3.21) is, in the continuous $\ell = 0$ case, already available in Saumard (2019). The kernel $K_p^\ell(x,x')$ is a classical object in the theory of covariance representations and inequalities; an early appearance is attributed by Rao (2006) to Höffding (1940) (see Höffding, 2012, pp 57–109 for an English translation). The perhaps not very surprising extension to the discrete case is, to the best of our knowledge, new.

As a first consequence of our set-up, (3.21) applied to the function $f(x) = \mathcal{T}_p^\ell 1(x)$ immediately gives the following proposition.

Proposition 3.3.5 (Menz-Otto formula)**.** *Suppose that the constant function* 1 *belongs to* $\mathcal{F}_\ell^{(1)}(p)$*, that* $-\Delta^{-\ell}\mathcal{T}_p^\ell 1(x) > 0$ *for almost all* $x \in \mathcal{S}(p)$ *and* $\Delta^{-\ell}(\mathcal{T}_p^\ell 1) \in L^1(\mu)$*. Then, for every* $x' \in \mathcal{S}(p)$*, the function*

$$p_{x'}^\star(x) = \frac{K_p^\ell(x,x')}{p(x')}\Big(- \Delta^{-\ell}\mathcal{T}_p^\ell 1(x)\Big) \tag{3.23}$$

is a density on $\mathcal{S}(p)$ *with respect to* μ*.*

Proof. From (3.21) with $f(x) = \mathcal{T}_p^\ell 1(x) = \Delta^\ell(p(x))/p(x)$ and $\mathcal{L}_p^\ell f(x) = 1$,

$$\begin{aligned} 1 &= -\mathbb{E}\left[\frac{K_p^\ell(X,x)}{p(X)p(x)}\Delta^{-\ell}\mathcal{T}_p^\ell 1(X)\right] = \int_a^b \frac{K_p^\ell(u,x)}{p(u)p(x)}\left(-\Delta^{-\ell}\mathcal{T}_p^\ell 1(u)\right)p(u)\mu(du) \\ &= \int_a^b \frac{K_p^\ell(u,x)}{p(x)}\left(-\Delta^{-\ell}\mathcal{T}_p^\ell 1(u)\right)\mu(du) = \int_a^b p_{x'}^\star(x)\mu(dx). \end{aligned}$$

Since $0 \le K_p^\ell(x,x') \le 1$, the integral exists because $-\Delta^{-\ell}\mathcal{T}_p^\ell 1(x)\frac{K_p^\ell(x,x)}{p(x)} \in L^1(p)$ and, by assumption, $-\Delta^{-\ell}\mathcal{T}_p^\ell 1(x) > 0$. Hence the assertion follows. □

Remark 3.3.6. *If* $K_p^\ell(x,x)/p(x)$ *is bounded, then the assumptions in Proposition 3.3.5 are satisfied as soon as* $-\Delta^{-\ell}\mathcal{T}_p^\ell 1 \in L^1(\mu)$*. The proposition thus applies when* $\ell = 0$ *and* $p(x) = \exp(-H(x))$ *with* H *a strictly convex function such that* $\lim_{x\to\pm\infty} H'(x) = 0$*. This puts us in the context studied by Menz and Otto (2013) and formula* (3.23) *is equivalent to their Equation (14); we return to this in Chapter 4 (Corollary 4.2.2).*

The next proposition gives some properties of $K_p^\ell(x,x')$.

Proposition 3.3.7. *(i)* $K_p^\ell(x,x') \le K_p^\ell(\min(x,x'),\min(x,x'))$ *for all* $x,x' \in \mathcal{S}(p)$*. (ii) If* $\mathbb{E}[\chi^\ell(X,x)]/p(x)$ *is non decreasing, then the function* $x \mapsto K_p^\ell(x,x')/p(x)$ *is non-decreasing for* $x < x'$*. (iii) If* $\mathbb{E}[\chi^{-\ell}(x,X)]/p(x)$ *is non increasing, then the function* $x \mapsto K_p^\ell(x,x')/p(x)$ *is non-increasing for* $x > x'$*.*

Proof. To see (i), we start from (3.22),

$$K_p^\ell(x,x') = \mathbb{E}\left[\chi^\ell(X,\min(x,x'))\right]\mathbb{E}\left[\chi^{-\ell}(\max(x,x'),X)\right]$$

and by Lemma 3.3.1, $\chi^{-\ell}(u,x)$ is non-increasing in u:

$$\chi^{-\ell}(\max(x,x'),X) \le \chi^{-\ell}(\min(x,x'),X).$$

We deduce that

$$K_p^\ell(x,x') \le \mathbb{E}\left[\chi^\ell(X,\min(x,x'))\right]\mathbb{E}\left[\chi^{-\ell}(\min(x,x'),X)\right].$$

Assertion (i) follows by reverting the argument, as $(\chi^\ell(y,z)^2 = \chi^\ell(y,z)$. To see (ii), assume that $\mathbb{E}[\chi^\ell(X,x)]/p(x)$ is non-decreasing. Then with (3.22), for $x < x'$,

$$\begin{aligned}\frac{1}{p(x)}K_p^\ell(x,x') &= \frac{1}{p(x)}\mathbb{E}\left[\chi^\ell(X,\min(x,x'))\right]\mathbb{E}\left[\chi^{-\ell}(\max(x,x'),X)\right]\\ &= \left(\frac{1}{p(x)}\mathbb{E}\left[\chi^\ell(X,x)\right]\right)\mathbb{E}\left[\chi^{-\ell}(x',X)\right];\end{aligned}$$

the second factor is a constant, and the first factor is assumed to be non-decreasing. Hence the assertion follows.

For (iii), assume that $\mathbb{E}[\chi^{-\ell}(x,X)]/p(x)$ is non-increasing; then similarly as above, for $x > x'$,

$$\begin{aligned}\frac{1}{p(x)}K_p^\ell(x,x') &= \frac{1}{p(x)}\mathbb{E}\left[\chi^\ell(X,\min(x,x'))\right]\mathbb{E}\left[\chi^{-\ell}(\max(x,x'),X)\right]\\ &= \left(\mathbb{E}\left[\chi^\ell(X,x')\right]\right)\left(\frac{1}{p(x)}\mathbb{E}\left[\chi^{-\ell}(x,X)\right]\right);\end{aligned}$$

the first factor is constant, and the second factor is non increasing. Hence the assertion follows. □

Figures 3.1 and 3.2 display the functions $x \mapsto K_p^\ell(x,x')/p(x)$ (for various values of x') and $x \mapsto K_p^\ell(x,x)/p(x)$ for the standard normal and several choices of the parameters in beta, gamma, binomial, Poisson and hypergeometric distributions (for the discrete distributions, only the case $\ell = -1$ is represented).

Example 3.3.8. *The following facts are easy to prove:*

1. *If $p(x)$ is the standard normal distribution then $K_p^0(x,x)/p(x)$ behaves as $1/|x|$ for large $|x|$, see Figure 3.2(a).*
2. *If $p(x)$ is gamma then $K_p^0(x,x)/p(x)$ behaves as a constant for large x, see Figure 3.2(e).*
3. *The function $x \mapsto K_p^0(x,x)/p(x)$ is not in $L^1(p)$ for p a Cauchy distribution.*
4. *If p is strictly-log concave then $K_p^\ell(x,x)/p(x)$ is bounded.*

4 Sufficient conditions and integrability

As anticipated, we now study the conditions under which the IBP Lemmas 3.2.5 and 3.2.19 hold. We start by the decryption of the conditions for Lemma 3.2.5. Recall the notations a_ℓ and b_ℓ from (3.2). Furthermore if $\ell = 0$ we write $f(a^+) = \lim_{x\to a, x>a} f(x)$ and $f(b^-) = \lim_{x\to b, x<b} f(x)$. In the case that $a = -\infty$ or $b = \infty$, for $\ell \in \{-1,0,1\}$, we write $f(-\infty^+) = \lim_{x\to-\infty} f(x)$ and $f(\infty^-) = \lim_{x\to\infty} f(x)$.

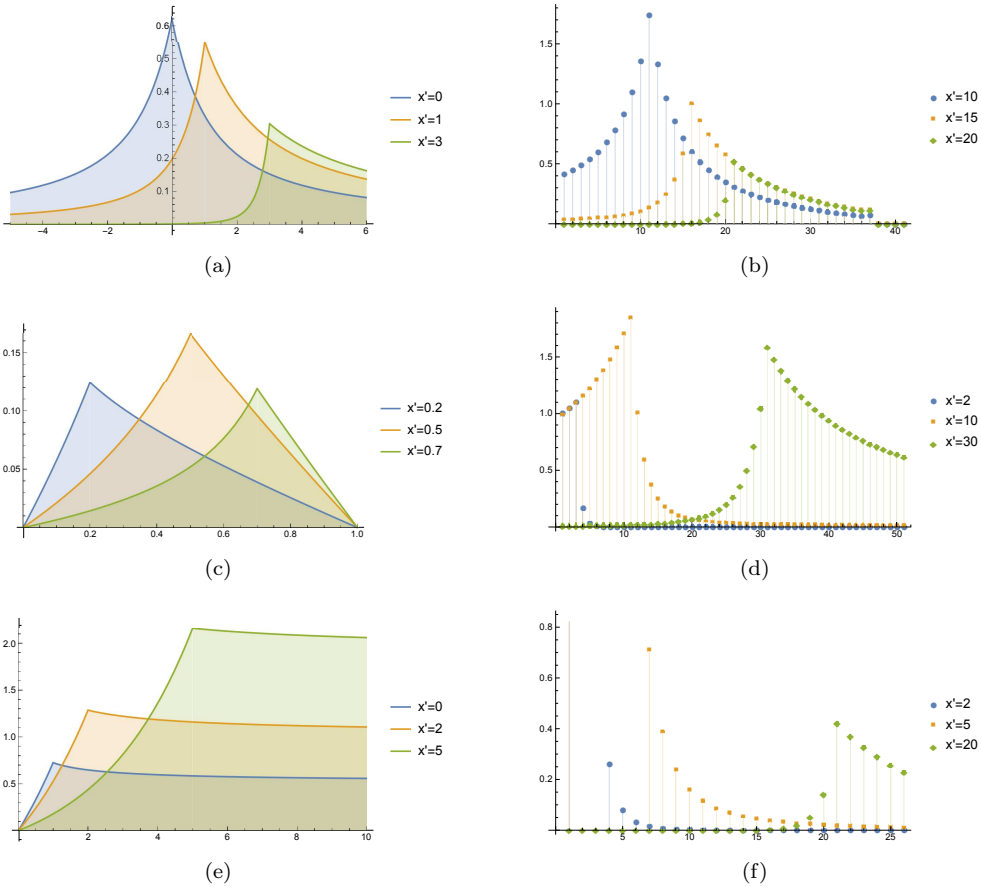


Figure 3.1: The functions $x \mapsto K_p^\ell(x, x')/p(x)$ for different (fixed) values of x' and p the standard normal distribution (Figure 3.1(a)); beta distribution with parameters 1.3 and 2.4 (Figure 3.1(c)); gamma distribution with parameters 1.3 and 2.4 (Figure 3.1(e)); binomial distribution with parameters $(50, 0.2)$ (Figure 3.1(b)); Poisson distribution with parameter 20 (Figure 3.1(d)); hypergeometric distribution with parameters 100, 50 and 500 (Figure 3.1(f)).

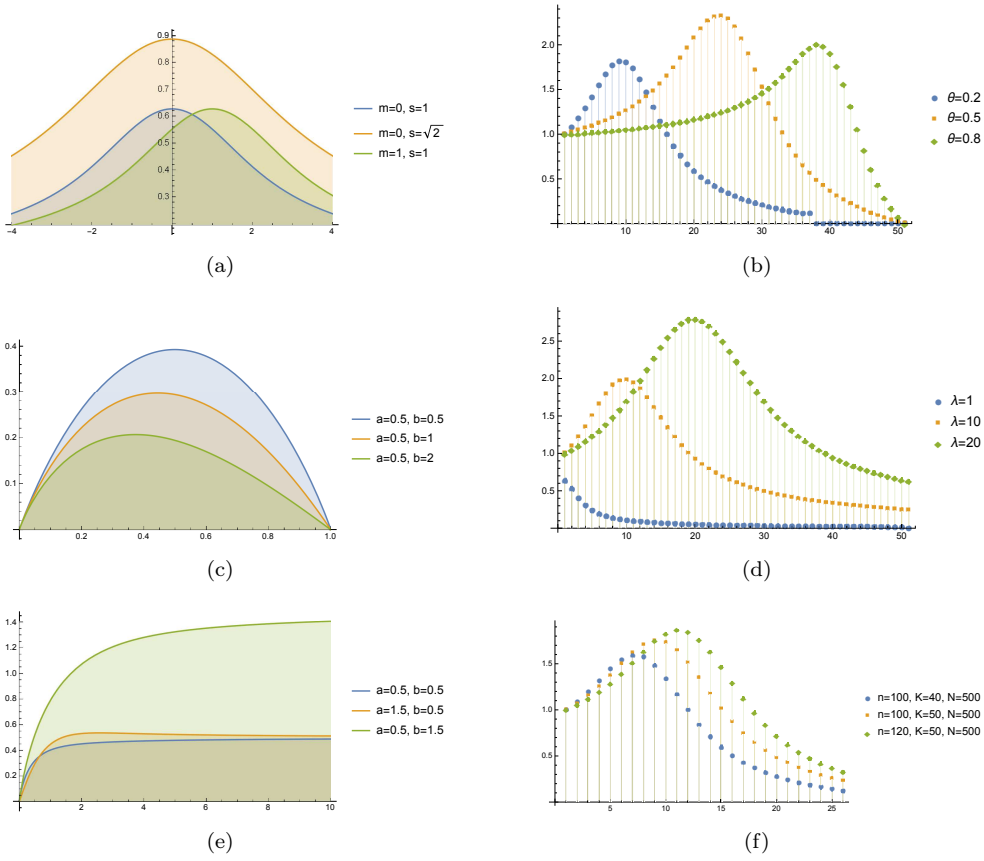


Figure 3.2: The functions $x \mapsto K_p^\ell(x,x)/p(x)$ for different parameter values p the standard normal distribution (Figure 3.2(a)); beta distribution (Figure 3.2(c)); gamma distribution (Figure 3.2(e)); binomial distribution with parameters (Figure 3.2(b)); Poisson distribution (Figure 3.2(d)); hypergeometric distribution (Figure 3.2(f)).

To simplify notation, if $\ell \in \{-1, 1\}$ and $a \neq -\infty$, we write $f(a^+) = f(a)$, and similarly, if $b \neq \infty, f(b^-) = f(b)$.

Proposition 3.4.1 (Sufficient conditions for IBP – version 1)**.** *Let $f \in \mathrm{dom}(p, \Delta^\ell)$ and $g \in \mathrm{dom}(\Delta^{-\ell})$. In order for* (3.4) *to hold it suffices that they jointly satisfy the following conditions*

$$(\mathcal{T}_p^\ell f)g \text{ and } f(\Delta^{-\ell} g) \in L^1(p) \tag{3.24}$$

$$f(b^- + a_\ell)g(b^- + a_\ell - \ell)p(b^- + a_\ell) = f(a^+ - b_\ell)g(a^+ - b_\ell - \ell)p(a^+ - b_\ell). \tag{3.25}$$

For ease of future reference, we spell out (3.25) in the three cases that interest us:

$$\begin{cases} f(b^-)g(b^-)p(b^-) = f(a^+)g(a^+)p(a^+) & \ell = 0 \\ f(b^-)g(b^- + 1)p(b^-) = 0 & \ell = -1 \\ f(a^+)g(a^+ - 1)p(a^+) = 0 & \ell = 1. \end{cases}$$

Proof. In order for (3.4) to hold we need (i) $f(\cdot)g(\cdot - \ell) \in \mathcal{F}_\ell^{(1)}(p)$ and (ii) $f\Delta^{-\ell} g \in L^1(p)$. Condition (ii) is satisfied under (3.24). By definition of $\mathcal{F}_\ell^{(1)}(p)$, condition (i) holds if the following three conditions apply: (iA) $f(\cdot)g(\cdot - \ell) \in \mathrm{dom}(p, \Delta^\ell)$, (iB) $\Delta^\ell\big(p(\cdot)f(\cdot)g(\cdot - \ell)\big)\mathbb{I}[\mathcal{S}(p)] \in L^1(\mu)$ and (iC) $\mathbb{E}\left[\mathcal{T}_p^\ell f(X)g(X - \ell)\right] = 0$. The proof hinges on product formula (3.3) which yields:

$$\Delta^\ell\big(p(x)f(x)g(x - \ell)\big) = \Delta^\ell\big(p(x)f(x)\big)g(x) + p(x)f(x)\Delta^{-\ell} g(x)$$

for all $x \in \mathcal{S}(p)$. In light of this, condition (iA) is implied by the requirement that $f \in \mathrm{dom}(p, \Delta^\ell)$ and $g \in \mathrm{dom}(\Delta^{-\ell})$. Similarly, because condition (iB) is equivalent to $(p(\cdot))^{-1}\Delta^\ell\big(p(\cdot)f(\cdot)g(\cdot - \ell)\big) \in L^1(p)$, we see that it is guaranteed by (3.24). Finally, applying (3.1), (iC) follows from (3.25). Hence Condition (i) holds under the stated assumptions. □

We now derive a set of (almost) necessary and sufficient conditions under which (3.13) holds.

Proposition 3.4.2 (Sufficient conditions for IBP – version 2)**.** *Let $g \in \mathrm{dom}(\Delta^{-\ell})$. In order for* (3.13) *to hold, it is necessary and sufficient that they jointly satisfy the three following conditions:*

$$f, g \text{ and } fg \in L^1(p), \tag{3.26}$$

$$\mathcal{L}_p^\ell f(\Delta^{-\ell} g) \in L^1(p), \tag{3.27}$$

$$\begin{aligned} &\mathcal{L}_p^\ell f(b^- + a_\ell)g(b^- + a_\ell - \ell)p(b^- + a_\ell) \\ &\qquad = \mathcal{L}_p^\ell f(a^+ - b_\ell)g(a^+ - b_\ell - \ell)p(a^+ - b_\ell). \end{aligned} \tag{3.28}$$

Proof. In order for (3.13) to hold, it is necessary and sufficient that (i) $f \in L^1(p)$, $g \in \mathrm{dom}(\Delta^{-\ell})$, (ii) $\big(\mathcal{L}_p^\ell f(\cdot)\big)g(\cdot - \ell)$ and (iii) $\mathcal{L}_p^\ell f(\Delta^{-\ell} g) \in L^1(p)$. Conditions (i) and (iii) are stated explicitly and all that remains is to check that (ii) is equivalent to the stated assumptions. As before, we recall that (ii) is equivalent to (iiA) $\big(\mathcal{L}_p^\ell f(\cdot)\big)g(\cdot - \ell) \in \mathrm{dom}(p, \Delta^\ell)$; (iiB) $\Delta^\ell\Big(p(\cdot)\big(\mathcal{L}_p^\ell f(\cdot)\big)g(\cdot - \ell)\Big)/p(\cdot) \in L^1(p)$; (iiC) $\mathbb{E}\left[\mathcal{T}_p^\ell\left(\big(\mathcal{L}_p^\ell f(\cdot)\big)g(\cdot - \ell)\right)(X)\right] = 0$. As in the proof of Proposition 3.4.1, the result hinges on the product rule (3.3) which now reads

$$\begin{aligned}\Delta^\ell\left(p(x)\big(\mathcal{L}_p^\ell f(x)\big)g(x-\ell)\right) &= \left(\Delta^\ell\big(p(x)\mathcal{L}_p^\ell f(x)\big)\right)g(x) + p(x)\mathcal{L}_p^\ell f(x)\Delta^{-\ell} g(x)\\ &= (f(x) - \mathbb{E}[f(X)])g(x) + p(x)\mathcal{L}_p^\ell f(x)\Delta^{-\ell} g(x).\end{aligned}$$

Hence condition (iiA) holds solely under the assumption that $g \in \mathrm{dom}(\Delta^{-\ell})$, (iiA) holds under (3.26) and (3.27). Finally, (3.28) guarantees that (iiC) is satisfied. □

Requirement (3.26) is natural and condition (3.28) is mild as it is satisfied as soon as g and/or f are well behaved at the edges of the support. Condition (3.27) (which is already stated in the original statement of Lemma 3.2.19) is harder to fathom. In order to make it even more readable, and facilitate the connexion with the literature, we specialise the conditions further in our next result.

Proposition 3.4.3. *Let f, g and $fg \in L^1(p)$. If $g \in \mathrm{dom}(\Delta^{-\ell})$ is of bounded variation and satisfies the following two conditions:*

1. $g(a^+ - b_\ell - \ell)\mathbb{P}(X \le a^+ - a_\ell - b_\ell) = 0$ *and* $g(b^- + a_\ell - \ell)\mathbb{P}(X \ge b^- + a_\ell + b_\ell) = 0$
2. $g(a^+ - b_\ell - \ell)\mathbb{E}[|f(X)|\chi^\ell(X \le a^+ - b_\ell)] = 0$ *and* $g(b^- + a_\ell - \ell)\mathbb{E}[|f(X)|\chi^{-\ell}(b^- + a_\ell, X)] = 0$,

then (3.28) *holds. In particular if f is bounded or in $L^2(p)$, then the condition 2 above is implied by condition 1.*

Proof. We want to apply Proposition 3.4.2; hence we check each condition in Proposition 3.4.2 separately. By assumption, (3.26) is satisfied and $g \in \mathrm{dom}(\Delta^{-\ell})$.

- For Assumption (3.27): First suppose that g is monotone increasing. It is to show that $\mathcal{L}_p^\ell f(\Delta^{-\ell} g) \in L^1(p)$. As $f \in L^1(p)$ is assumed, we can use (3.19) to get

$$\begin{aligned}\mathbb{E}\left[\left|\mathcal{L}_p^\ell f(X)\right|\left|\Delta^{-\ell} g(X)\right|\right] &= \mathbb{E}\left[\left|\mathcal{L}_p^\ell f(X)\right|\Delta^{-\ell} g(X)\right]\\ &\le \mathbb{E}\left[|f(X_2) - f(X_1)|\,\Phi_p^\ell(X_1, X, X_2)\Delta^{-\ell} g(X)\right]\\ &\le \mathbb{E}\left[|f(X_2) - f(X_1)|\,\mathbb{E}\left[\Phi_p^\ell(X_1, X, X_2)\Delta^{-\ell} g(X)\,|\,X_1, X_2\right]\right]\\ &= \mathbb{E}\left[|f(X_2) - f(X_1)|\,(g(X_2) - g(X_1))\,\mathbb{I}[X_1 < X_2]\right]\end{aligned}$$

where we used the first identity in (3.20) in the last line. This last expression is necessarily finite because f, g and fg are in $L^1(p)$. The general conclusion follows from the fact that any function of bounded variation is the difference between two monotone functions; the triangle inequality thus yielding the claim.

- For Assumption (3.28): Since $f \in L^1(p)$, we can apply (3.19) and the definition of Φ_p^ℓ to obtain

$$-p(x)\mathcal{L}_p^\ell f(x) = \mathbb{E}[f(X)\chi^{-\ell}(x,X)]\mathbb{E}[\chi^\ell(X,x)] - \mathbb{E}[f(X)\chi^\ell(X,x)]\mathbb{E}[\chi^{-\ell}(x,X)].$$

Then

$$\begin{aligned}
&\lim_{x\to a, x>a} \left| \Big(\mathcal{L}_p^\ell f(x-b_\ell)\Big) g(x-b_\ell-\ell)p(x-b_\ell) \right| \\
&\leq \lim_{x\to a, x>a} \Big(|g(x-b_\ell-\ell)|\mathbb{E}[|f(X)|\chi^{-\ell}(x-b_\ell,X)]\mathbb{E}[\chi^\ell(X,x-b_\ell)] \\
&\qquad + |g(x-b_\ell-\ell)|\mathbb{E}[|f(X)|\chi^{\ell}(X,x-b_\ell)]\mathbb{E}[\chi^{-\ell}(x-b_\ell,X)] \Big) \\
&\leq \lim_{x\to a, x>a} |g(x-b_\ell-\ell)|\mathbb{E}[|f(X)|\chi^{-\ell}(x-b_\ell,X)]\mathbb{P}\left(X \leq x-a_\ell-b_\ell\right) \quad = L_1 \\
&\qquad + \lim_{x\to a, x>a} |g(x-b_\ell-\ell)|\mathbb{E}[|f(X)|\chi^{\ell}(X,x-b_\ell)]\mathbb{P}\left(X \geq x\right) \quad = L_2
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{x\to b, x<b} \left| \Big(\mathcal{L}_p^\ell f(x+a_\ell)\Big) g(x+a_\ell-\ell)p(x+a_\ell) \right| \\
&\leq \lim_{x\to b, x<b} \Big(|g(x+a_\ell-\ell)|\mathbb{E}[|f(X)|\chi^{-\ell}(x+a_\ell,X)]\mathbb{E}[\chi^\ell(X,x+a_\ell)] \\
&\qquad + |g(x+a_\ell-\ell)|\mathbb{E}[|f(X)|\chi^{\ell}(X,x+a_\ell)]\mathbb{E}[\chi^{-\ell}(x+a_\ell,X)] \Big) \\
&\leq \lim_{x\to b, x<b} |g(x+a_\ell-\ell)|\mathbb{E}[|f(X)|\chi^{-\ell}(x+a_\ell,X)]\mathbb{P}\left(X \leq x\right) \quad = L_3 \\
&\qquad + \lim_{x\to b, x<b} |g(x+a_\ell-\ell)|\mathbb{E}[|f(X)|\chi^{\ell}(X,x+a_\ell)]\mathbb{P}\left(X \geq x+a_\ell+b_\ell)\right) \quad = L_4.
\end{aligned}$$

Condition 1 guarantees that $L_1 = L_4 = 0$; condition 2 guarantees that $L_2 = L_3 = 0$. If, furthermore, f is bounded then the sufficiency of 1 is immediate; if $f \in L^2(p)$ then it follows from the Cauchy-Schwarz inequality.

□

Remark 3.4.4. *This assumption is closer to what is to be found in the literature, see e.g. Saumard (2019) in the case $\ell = 0$. The main difference between the classical assumptions and ours is that we only impose conditions on one of the functions. We stress that there is a certain degree of redundancy in the items 1 and 2 together with the assumption that $g \in L^1(p)$ and is of bounded variation; the statement could be shortened at the loss of readability.*

In the sequel, to preserve as much generality as possible and not overburden the statements, we will simply require that "the assumptions of Lemma 3.2.19 are satisfied."

5 The inverse Stein operator

We conclude this chapter by exploring easy consequences of the representations from Section 3. These results are also of independent interest to practitioners of Stein's method and we come back to this topic in Chapter 6.

Lemma 3.5.1. *If $f, \mathcal{L}_p^\ell f(X) \in L^1(p)$ then*

$$\mathbb{E}\left[-\mathcal{L}_p^\ell f(X)\right] = \mathbb{E}\left[(X_2 - X_1)^+ (f(X_2) - f(X_1))\right] \tag{3.29}$$

$$= \frac{1}{2}\mathbb{E}\left[(X_2 - X_1)(f(X_2) - f(X_1))\right] \tag{3.30}$$

where $(\cdot)^+$ denotes the positive part of $(\cdot)$. In particular, if the conditions of Lemma 3.2.19 are satisfied with $f(x) = g(x) = \mathrm{Id}$, then $\mathbb{E}[\tau_p^\ell(X)] = \mathrm{Var}(X)$.

Proof. Representation (3.19) gives $\mathbb{E}\big[-\mathcal{L}_p^\ell f(X)\big] = \mathbb{E}[(f(X_2) - f(X_1))\Phi_p^\ell(X_1, X, X_2)]$. Using (3.20) with $f(x) = x$, we have $\mathbb{E}[\Phi_p^\ell(x_1, X, x_2)] = (x_2 - x_1)^+$. Hence, after conditioning with respect to X_1, X_2, the equality (3.29) follows. The second equality (3.30) follows by symmetry. The second claim is immediate under the stated assumptions. □

Remark 3.5.2. *Once again, our assumptions are minimal but not transparent. It is easy to spell out these conditions explicitly for any specific target. For instance if X has bounded support or support $\mathbb{R}$ then finite variance suffices.*

Proposition 3.5.3. *Suppose that all test functions satisfy the conditions in Lemma 3.2.19. Let $\|f\|_{\mathcal{S}(p),\infty} = \sup_{x \in \mathcal{S}(p)} |f(x)|$.*

1. *If f is monotone then $\mathcal{L}_p^\ell f(x)$ does not change sign.*
2. *(Uniform bounds Stein bounds) Consider, for h and η in $L^1(p)$ the function*

$$g_h^{p,\ell,\eta}(x) = \frac{\mathcal{L}_p^\ell h(x + \ell)}{\mathcal{L}_p^\ell \eta(x + \ell)}$$

defined in (3.10) *which solves the η-Stein equation* (3.9) *for h. If η is monotone and $|h(x) - h(y)| \leq k|\eta(x) - \eta(y)|$ for all $x, y \in \mathcal{S}(p)$, then*

$$\|g_h^{p,\ell,\eta}\|_{\mathcal{S}(p),\infty} \leq k.$$

In particular, if $h \in L^1(p)$ is Lipschitz continuous with Lipschitz constant 1 *then the above applies with $\eta(x) = x$, and $\|g_h^{p,0,\mathrm{Id}}\|_{\mathcal{S}(p),\infty} \leq 1$.*

3. (Non uniform bounds Stein bounds)

$$\left|\mathcal{L}_p^\ell f(x)\right| \leq 2\|f\|_{\mathcal{S}(p),\infty}\frac{\mathbb{E}[\chi^\ell(X_1,x)]\mathbb{E}[\chi^{-\ell}(x,X_2)]}{p(x)}$$

for all $x \in \mathcal{S}(p)$.

Proof. Recall representation (3.19) which states that

$$-\mathcal{L}_p^\ell f(x) = \frac{1}{p(x)}\mathbb{E}\left[(f(X_2)-f(X_1))\chi^\ell(X_1,x)\chi^{-\ell}(x,X_2)\right].$$

1. If f is monotone then $f(X_2)-f(X_1)$ is of constant sign conditionally on the event $\chi^\ell(X_1,x)\chi^{-\ell}(x,X_2) = \mathbb{I}[X_1+a_\ell \leq x \leq X_2 - b_\ell] = 1$, because on this event, $X_1 \leq X_2 - \mathbb{I}[\ell \neq 0]$. Hence the first assertion follows.
2. Suppose that the function η is strictly decreasing. By definition of g we have, under the stated conditions,

$$\begin{aligned}
|g(x)| &= \left|\frac{-\mathbb{E}\left[(h(X_2)-h(X_1))\chi^\ell(X_1,x+\ell)\chi^{-\ell}(x+\ell,X_2)\right]}{-\mathbb{E}\left[(\eta(X_2)-\eta(X_1))\chi^\ell(X_1,x+\ell)\chi^{-\ell}(x+\ell,X_2)\right]}\right| \\
&= \frac{\left|-\mathbb{E}\left[(h(X_2)-h(X_1))\chi^\ell(X_1,x+\ell)\chi^{-\ell}(x+\ell,X_2)\right]\right|}{\mathbb{E}\left[(\eta(X_1)-\eta(X_2))\chi^\ell(X_1,x+\ell)\chi^{-\ell}(x+\ell,X_2)\right]} \\
&\leq \frac{\mathbb{E}\left[|h(X_2)-h(X_1)|\chi^\ell(X_1,x+\ell)\chi^{-\ell}(x+\ell,X_2)\right]}{\mathbb{E}\left[|\eta(X_2)-\eta(X_1)|\chi^\ell(X_1,x+\ell)\chi^{-\ell}(x+\ell,X_2)\right]} \\
&\leq k
\end{aligned}$$

   for $x \in \mathcal{S}(p)$.
3. By (3.19),

$$\begin{aligned}
\left|\mathcal{L}_p^\ell f(x)\right|p(x) &= \mathbb{E}\left[|f(X_2)-f(X_1)|\chi^\ell(X_1,x)\chi^{-\ell}(x,X_2)\right] \\
&\leq 2\|f\|_{\mathcal{S}(p),\infty}\mathbb{E}\left[\chi^\ell(X_1,x)\chi^{-\ell}(x,X_2)\right] \\
&\leq 2\|f\|_{\mathcal{S}(p),\infty}\mathbb{E}\left[\chi^\ell(X_1,x)\right]\mathbb{E}\left[\chi^{-\ell}(x,X_2)\right]
\end{aligned}$$

   which leads to the conclusion.

□

Example 3.5.4. *If* $p = \phi$ *is the standard Gaussian with cdf* Φ*, then* $\ell = 0$ *and the third bound in Proposition 3.5.3 reduces to* $2||f||_\infty\Phi(x)(1-\Phi(x))/\phi(x)$*. The ratio* $\Phi(x)\big(1-\Phi(x)\big)/\phi(x)$ *is closely related to Mill's ratio of the standard normal law. The study of such a function is classical and much is known. For instance, we can apply (Baricz, 2008, Theorem 2.3) to get*

$$\frac{1}{\sqrt{x^2+4}+x} \leq \frac{\Phi(x)\big(1-\Phi(x)\big)}{\phi(x)} \leq \frac{4}{\sqrt{x^2+8}+3x}$$

for all $x \geq 0$. *Moreover,* $\Phi(x)\big(1-\Phi(x)\big)/\phi(x) \leq \Phi(0)\big(1-\Phi(0)\big)/\phi(0) = 1/2\sqrt{\pi/2} \approx 0.626$. *In particular Proposition 3.5.3 recovers the well-known bound*

$$\|\mathcal{L}_p^\ell f\|_\infty \leq \sqrt{\pi/2}\|f\|_\infty,$$

see e.g. Nourdin and Peccati (2012, Theorem 3.3.1).

CHAPTER 4

First order covariance identities and inequalities

1 Introduction

Much attention has been given in the literature to the problem of providing sharp tractable estimates on the variance of functions of random variables. Such estimates are directly related to fundamental considerations of pure mathematics (e.g., isoperimetric, logarithmic Sobolev and Poincaré inequalities), as well as essential issues from statistics (e.g., Cramer-Rao bounds, efficiency and asymptotic relative efficiency computations, maximum correlation coefficients, and concentration inequalities).

One of the starting points of this line of research is Chernoff's famous result from Chernoff (1980) which states that, if $N \sim \mathcal{N}(0,1)$, then

$$\mathbb{E}[g'(N)]^2 \leq \mathrm{Var}[g(N)] \leq \mathbb{E}[g'(N)^2] \tag{4.1}$$

for all sufficiently regular functions $g : \mathbb{R} \to \mathbb{R}$. Chernoff obtained the upper bound by exploiting orthogonality properties of the family of Hermite polynomials. The upper bound in (4.1) is, in fact, already available in Nash (1958) and is also a special case of the central inequality in Brascamp and Lieb (1976), see below. Cacoullos (1982) extends Chernoff's bound to a wide class of univariate distributions (including discrete distributions) by proving that if $X \sim p$ has a density function p with respect to the Lebesgue measure then

$$\frac{\mathbb{E}[\tau_p(X)\, g'(X)]^2}{\mathrm{Var}[X]} \leq \mathrm{Var}[g(X)] \leq \mathbb{E}\left[\tau_p(X)\, g'(X)^2\right] \tag{4.2}$$

with $\tau_p(x) = p(x)^{-1} \int_x^\infty (t - \mathbb{E}[X]) p(t) \mathrm{d}t$. It is easy to see that, if p is the standard normal density, then $\tau_p(x) = 1$ so that (4.2) contains (4.1). Cacoullos also obtains

a similar bound as (4.2) for discrete distributions on the positive integers, where the derivative is replaced by the forward difference and the weight becomes $\tau_p(x) = p(x)^{-1} \sum_{t=x+1}^{\infty} tp(t)$.

Variance inequalities such as (4.2) are closely related to the celebrated *Brascamp-Lieb inequality* from Brascamp and Lieb (1976) which, in dimension 1, states that if $X \sim p$ and p is strictly log-concave then

$$\mathrm{Var}[g(X)] \leq \mathbb{E}\left[\frac{(g'(X))^2}{(-\log p)''(X)}\right] \tag{4.3}$$

for all sufficiently regular functions g. In fact, the upper bound from (4.1) is an immediate consequence of (4.3) because, if p is the standard Gaussian density, then $(-\log p)''(x) \equiv 1$. This Brascamp-Lieb inequality is proved in Menz and Otto (2013) to be a consequence of Höffding's classical covariance inequality from Höffding (1940), which states that if (X, Y) is a continuous bivariate random vector with cumulative distribution $H(x, y)$ and marginal cdfs $F(x), G(x)$ then

$$\mathrm{Cov}[f(X), g(Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f'(x)\Big(H(x,y) - F(x)G(y)\Big)g'(y)\,\mathrm{d}x\,\mathrm{d}y \tag{4.4}$$

under weak assumptions on f, g (see e.g. Cuadras, 2002). The freedom of choice in the test functions f, g in (4.4) is exploited by Menz and Otto (2013) to prove that, if X has a C^2 strictly convex absolutely continuous density p then an *asymmetric Brascamp-Lieb inequality* holds:

$$|\mathrm{Cov}[f(X), g(X)]| \leq \sup_x \left\{\frac{|f'(x)|}{(\log p)''(x)}\right\} \mathbb{E}\big[\,|g'(X)|\,\big]. \tag{4.5}$$

Identity (4.4) and inequalities (4.3) and (4.5) are extended to the multivariate setting in Carlen et al. (2013) which also gives connections with logarithmic Sobolev inequalities for spin systems and related inequalities for log-concave densities. This material is revisited and extended in Saumard and Wellner (2019, 2018), Saumard (2019), providing applications in the context of isoperimetric inequalities and weighted Poincaré inequalities. In Cuadras (2002) the identity (4.4) is proved in all generality and used to provide expansions for the covariance in terms of canonical correlations and variables.

Further generalizations of Chernoff's bounds are provided in Chen (1985), Cacoullos and Papathanasiou (1985, 1986), and Karlin (1993) (e.g., Karlin deals with the entire class of log-concave distributions). See also Borovkov and Utev (1984), Cacoullos and Papathanasiou (1989), Korwar (1991), Papathanasiou (1995), Cacoullos and Papathanasiou (1995) for the connection with probabilistic characterizations as

well as Furioli et al. (2017) and Toscani (2019) for several generalizations in particular to stable distributions. Similar inequalities were obtained – often by exploiting properties of suitable families of orthogonal polynomials – for univariate functionals of some specific multivariate distributions e.g., in Cacoullos and Papathanasiou (1992), Cacoullos et al. (1998), Chang and Richards (1999), Landsman et al. (2013), Afendras and Papathanasiou (2014), Landsman et al. (2015). A historical overview as well as a description of the connection between such bounds, the so-called Stein identities from Stein's method (see below) and Sturm-Liouville theory (see Section 3) can be found in Diaconis and Zabell (1991). Finally, we mention that all this material is closely connected to the study of the so-called *spectral gap* of the operator $\mathcal{L}f = f'' + (\log p)' f'$ which, in one dimension, is defined as

$$\lambda = \inf_g \frac{\mathbb{E}[(g'(X))^2]}{\mathrm{Var}[g(X)]}$$

where the infimum is taken over all functions $g \in C_0^\infty(\mathbb{R})$ such that $\mathrm{Var}[g(X)] > 0$. We return to this briefly in Section 3.3, and refer the reader to Bonnefont and Joulin (2014), Bonnefont et al. (2016), Roustant et al. (2017) and Bonnefont and Joulin (2019) for an up-to-date overview of this topic.

To the best of our knowledge, the most general version of (4.1) and (4.2) is due to Klaassen (1985), where the following result is proved

Theorem 4.1.1 (Klaassen bounds)**.** *Let μ be some σ-finite measure. Let $\rho(x, y)$ be a measurable function such that $\rho(x, \cdot)$ does not change sign for μ almost $x \in \mathbb{R}$. Suppose that g is a measurable function such that $G(x) = \int \rho(x, y) g(y)\, \mu(\mathrm{d}y) + c$ is well defined for some $c \in \mathbb{R}$. Let X be a real random variable with density p with respect to μ.*

- *(Klaassen upper variance bound) For all nonnegative measurable functions $h : \mathbb{R} \to \mathbb{R}$ such that $\mu(\{x \in \mathbb{R} \,|\, g(x) \neq 0,\, p(x)h(x) = 0\}) = 0$ we have*

$$\mathrm{Var}[G(X)] \leq \mathbb{E}\left[\frac{g(X)^2}{h(X)} \left(\frac{1}{p(X)} \int \rho(z, X) H(z) p(z) \mu(\mathrm{d}z)\right)\right] \tag{4.6}$$

  with $H : \mathbb{R} \to \mathbb{R}$ supposed well-defined by $H(x) = \int \rho(x, y) h(y)\, \mu(\mathrm{d}y)$.
- *(Cramér-Rao lower variance bound) For all measurable functions $k : \mathbb{R} \to \mathbb{R}$ such that $0 < \mathbb{E}[k^2(X)] < \infty$ and $\mathbb{E}[k(X)] = 0$ we have*

$$\mathrm{Var}[G(X)] \geq \frac{\mathbb{E}\left[g(X)K(X)\right]^2}{\mathrm{Var}[k(X)]} \tag{4.7}$$

  where $K(x) = p(x)^{-1} \int \rho(z, x) k(z) p(z) \mu(\mathrm{d}z)$. Equality in (4.7) holds if and only if G is linear in k, p-almost everywhere.

Klaassen's proof of Theorem 4.1.1 relies on little more than the Cauchy-Schwarz inequality and Fubini's theorem; it has a slightly magical aura as little or no heuristic or context is provided as to the best choices of test functions h, k and kernel ρ or even to the nature of the weights appearing in (4.6) and (4.7). To the best of our knowledge, all available first order variance bounds from the literature can be obtained from either (4.6) or (4.7) by choosing the appropriate test functions h or k and the appropriate kernel ρ. For instance, the weights appearing in the upper bound (4.6) generalize the Stein kernel from Cacoullos' bound (4.2) – both in the discrete and the continuous case. Indeed taking $H(x) = x$ when the distribution p is continuous we see that then $h(x) = 1$ and the weight becomes $p(x)^{-1} \int \rho(z, x) z p(z) \mathrm{d}\mu(z)$ which is none other than $\tau_p(x)$. A similar argument holds as well in the discrete case. In the same way, taking $k(x) = x$ leads to $K(x) = \tau_p(x)$ in (4.7) and thus the lower bound in (4.2) follows as well. The freedom of choice in the function h allows for much flexibility in the quality of the weights; this fact seems somewhat under exploited in the literature. This is perhaps due to the rather obscure nature of Klaassen's weights, a topic which we shall be one of the central learnings of this chapter. Indeed we shall provide a natural theoretical home for Klaassen's result, in the framework of Stein's method.

Several variations on Klaassen's theorem have already been obtained via techniques related to Stein's method. A proper introduction of these techniques is developed in Chapter 3. The gist of the approach can nevertheless be understood very simply in case the underlying distribution is standard normal. Stein's classical identity states that $N \sim \mathcal{N}(0, 1)$ if and only if

$$\mathbb{E}[Ng(N)] = \mathbb{E}[g'(N)] \tag{4.8}$$

for all bounded, continuous g such that $\mathbb{E}[|g'(N)|] < \infty$. By the Cauchy-Schwarz inequality we immediately deduce that, for all appropriate g,

$$\mathbb{E}[g'(N)]^2 = \mathbb{E}[N(g(N) - \mathbb{E}[g(N)])]^2 \leq \mathbb{E}[N^2]\mathbb{E}\left[(g(N) - \mathbb{E}[g(N)])^2\right] \leq \mathrm{Var}[g(N)], \tag{4.9}$$

which gives the lower bound in (4.1). For the upper bound, still by the Cauchy-Schwarz inequality,

$$\mathrm{Var}[g(N)] \leq \mathbb{E}\left[\left(\int_0^N g'(x)\mathrm{d}x\right)^2\right] \leq \mathbb{E}\left[N\int_0^N (g'(x))^2\mathrm{d}x\right] = \mathbb{E}\left[(g'(N))^2\right] \tag{4.10}$$

where the last identity is a direct consequence of Stein's identity (4.8) applied to the function $g(x) = \int_0^x (g'(u))^2 \mathrm{d}u$. This is the upper bound in (4.1). The idea behind this proof is due to Chen (1982). As is now well known (again, we refer the reader to

Chapter 3 for references and details), Stein's identity (4.8) for the normal distribution can be extended to basically any univariate (and even multivariate) distribution via a family of objects called "Stein operators". This leads to a wide variety of Stein-type integration by parts identities and it is natural to wonder whether Chen's approach can be used to obtain generalizations of Klaassen's theorem. First steps in this direction are detailed in Ley and Swan (2013b, 2016); in particular it is seen that general lower variance bounds are easy to obtain from generalized Stein identities in the same way as in (4.9). Nevertheless, the method of proof in (4.10) for the upper bound cannot be generalized to arbitrary targets and, even in cases where the method does apply, the assumptions under which the bounds hold are quite stringent. To the best of our knowledge, the first to obtain upper variance bounds via properties of Stein operators is due to Saumard (2019), by combining generalized Stein identities – expressed in terms of the Stein kernel $\tau_p(x)$ – with Höffding's identity (4.4). The scope of Saumard's weighted Poincaré inequalities is, nevertheless, limited and a general result such as Klaassen's is, to this date, not available in the literature.

There are obviously many applications of such material, not only towards considerations of pure mathematics but also to more applied questions, such as – in no particular order – questions from sensitivity analysis (Roustant et al., 2017), stochastic ordering (Rao, 2006), the study of spin systems (Menz and Otto, 2013), and efficiency considerations (Afendras et al., 2011). We refer to all the above mentioned references for more references and details.

The main contribution of this chapter is a generalization of Klaassen's variance bounds from Theorem 4.1.1 to covariance inequalities of arbitrary functionals of arbitrary univariate targets under minimal assumptions (see Theorems 4.2.1 and 4.2.5). Our results hereby therefore also contains basically the entire literature on the topic, in a unified framework containing in particular both continuous and discrete distributions alike. Moreover, the weights that appear in our bounds bear a clear and natural interpretation in terms of Stein operators which allow for easy computation for a wide variety of targets, as illustrated in the different examples we tackle as well as in Tables 4.1, 4.2 and 4.3 in which we provide explicit variance bounds for univariate target distributions belonging to the classical integrated Pearson and Ord families (see Example 4.2.8 for a definition). Klaassen's bounds, its aforementioned corollaries as well as an (asymetric) Brascamp-Lieb and the weighted Poincaré inequality arise naturally in our setting. Moreover in all these cases we recover freedom of choice in the weights which allows to weaken the underlying assumptions and extend the scope (e.g. to non-absolutely continuous distributions).

2 Covariance identities and inequalities

We start with an easy lower bound inequality, which follows immediately from Lemma 3.2.5.

Proposition 4.2.1 (Cramer-Rao type bound)**.** *Let $g \in L^2(p)$. For any $f \in \mathcal{F}_\ell^{(1)}(p)$ such that $\mathcal{T}_p^\ell f \in L^2(p)$ and the assumptions of Lemma 3.2.5 are satisfied:*

$$\mathrm{Var}[g(X)] \geq \frac{\mathbb{E}\left[f(X)(\Delta^{-\ell} g(X))\right]^2}{\mathbb{E}\left[\left(\mathcal{T}_p^\ell f(X)\right)^2\right]} \tag{4.11}$$

with equality if and only if there exist α, β real numbers such that $g(x) = \alpha \mathcal{T}_p^\ell f(x) + \beta$ for all $x \in \mathcal{S}(p)$.

Proof. The lower bound (4.11) follows from the fact that $\mathcal{T}_p^\ell f \in \mathcal{F}^{(0)}(p)$ for all $f \in \mathcal{F}_\ell^{(1)}(p)$ by Lemma 3.2.4. Therefore, from (3.4), we have

$$\begin{aligned}
\left\{\mathbb{E}\left[f(X)\big(\Delta^{-\ell} g(X)\big)\right]\right\}^2 &= \left\{\mathbb{E}\left[\big(\mathcal{T}_p^\ell f(X)\big) g(X)\right]\right\}^2 \\
&= \left\{\mathbb{E}\left[\big(\mathcal{T}_p^\ell f(X)\big)\big(g(X) - \mathbb{E}[g(X)]\big)\right]\right\}^2 \\
&\leq \mathbb{E}\left[\left(\mathcal{T}_p^\ell f(X)\right)^2\right] \mathrm{Var}[g(X)]
\end{aligned}$$

by the Cauchy-Schwarz inequality. □

Upper bounds require some more work. We start with an easy consequence of our framework.

Corollary 4.2.2 (First order covariance identities)**.** *For all f, g that jointly satisfy the assumptions of Lemma 3.2.19, we have*

$$\mathrm{Cov}[f(X), g(X)] = \mathbb{E}\left[\Delta^{-\ell} f(X) \frac{K_p^\ell(X, X')}{p(X)p(X')} \Delta^{-\ell} g(X')\right]. \tag{4.12}$$

Moreover, if choice $f = \mathrm{Id}$ is allowed, then

$$\mathrm{Cov}[X, g(X)] = \mathbb{E}\left[\tau_p^\ell(X) \Delta^{-\ell} g(X)\right]. \tag{4.13}$$

Remark 4.2.3. *Identity* (4.12) *is provided in Menz and Otto (2013) (see their Equation (11)) in the case $\ell = 0$ for a log-concave density. Some of the history of this identity, including the connection with a classical identity of Höffding (1940), is provided in Saumard and Wellner (2018, Section 2). The earliest version of the same identity (still for $\ell = 0$) we have found in Cuadras (2002), along with applications to measures of correlation as well as further references. A similar identity is provided*

in Menz and Otto (2013), without explicit conditions; a clear statement is given in Saumard and Wellner (2018, Corollary 2.2) where the identity is proved for absolutely continuous $f \in L^r$ and $g \in L^s$ with conjugate exponents. Our approach shows that it suffices to impose regularity on one of the functions for the identity to hold.

Proof. Let $\bar{f}(x) = f(x) - \mathbb{E}[f(X)]$.Note that $\Delta^\ell \bar{f} = \Delta^\ell f$. To obtain (4.12) we start from (3.13) and note that if f, g satisfy the assumptions of Lemma 3.2.19, then

$$\mathrm{Cov}[f(X), g(X)] = \mathbb{E}\left[-\{\mathcal{L}_p^\ell f(X)\}\Delta^{-\ell} g(X)\right].$$

From this equation, (4.13) follows immediately. Applying (3.21) we obtain

$$\mathrm{Cov}[f(X), g(X)] = \mathbb{E}\left[\mathbb{E}\left[\frac{K_p(X', X)}{p(X')p(X)}\Delta^{-\ell} f(X') \mid X\right]\Delta^{-\ell} g(X)\right]$$

which gives the claim after removing the conditioning. □

Example 4.2.4. *Example 3.2.17 and identity* (4.13) *give the following covariance identities.*

- *Binomial distribution: For all functions $g : \mathbb{Z} \to \mathbb{R}$ that are bounded on $[0, n]$,*

  $$\mathrm{Cov}[X, g(X)] = \mathbb{E}\left[(1-\theta)X\Delta^- g(X)\right] = \theta\mathbb{E}\left[(n-X)\Delta^+ g(X)\right].$$

  Combining the two identities we also arrive at

  $$\mathrm{Cov}[X, g(X)] = \mathrm{Var}[X]\mathbb{E}[\nabla_{\mathrm{bin}(n,\theta)} g(X)]$$

  with $\nabla_{\mathrm{bin}(n,\theta)}$ the gradient $\nabla_{\mathrm{bin}(n,\theta)} g(x) = (x/n)\Delta^- g(x) + (1 - x/n)\Delta^+ g(x)$ from Hillion et al. (2014).
- *Beta distribution: For all absolutely continuous g such that $\mathbb{E}\left[|X(1-X)g'(X)|\right] < \infty$,*

  $$\mathrm{Cov}[X, g(X)] = \frac{1}{\alpha+\beta}\mathbb{E}\left[X(1-X)g'(X)\right].$$

It is of interest to work as in Klaassen (1985) to obtain a corresponding upper bound, which would provide some "weighted Poincaré inequality" such as those described in Saumard (2019). The representation formulae (4.12) turns out to simplify the work considerably.

Theorem 4.2.5. *Fix $h \in L^1(p)$ a decreasing function. For all f, g which satisfy the assumptions of Lemma 3.2.19 we have*

$$|\mathrm{Cov}[f(X), g(X)]| \le \sqrt{\mathbb{E}\left[(\Delta^{-\ell} f(X))^2 \frac{-\mathcal{L}_p^\ell h(X)}{\Delta^{-\ell} h(X)}\right]}\sqrt{\mathbb{E}\left[(\Delta^{-\ell} g(X))^2 \frac{-\mathcal{L}_p^\ell h(X)}{\Delta^{-\ell} h(X)}\right]} \tag{4.14}$$

with equality if and only if there exist $\alpha_i, i = 1, \ldots, 4$ real numbers such that $f(x) = \alpha_1 h(x) + \alpha_2$ and $g(x) = \alpha_3 h(x) + \alpha_4$ for all $x \in \mathcal{S}(p)$.

Proof. We simply apply (4.12) and the Cauchy-Schwarz inequality to obtain

$$
\begin{aligned}
|\mathrm{Cov}[f(X), g(X)]| &= \left| \mathbb{E}\left[\Delta^{-\ell} f(X) \frac{K_p^\ell(X, X')}{p(X)p(X')} \Delta^{-\ell} g(X') \right] \right| \\
&= \left| \mathbb{E}\left[\left\{ \frac{\Delta^{-\ell} f(X)}{\sqrt{-\Delta^{-\ell} h(X)}} \sqrt{-\frac{K_p^\ell(X, X')}{p(X)p(X')} \Delta^{-\ell} h(X')} \right\} \right.\right. \\
&\qquad \left.\left. \times \left\{ \frac{\Delta^{-\ell} g(X')}{\sqrt{-\Delta^{-\ell} h(X')}} \sqrt{-\frac{K_p^\ell(X, X')}{p(X)p(X')} \Delta^{-\ell} h(X)} \right\} \right] \right| \\
&\leq \sqrt{\mathbb{E}\left[\frac{(\Delta^{-\ell} f(X))^2}{\Delta^{-\ell} h(X)} \frac{K_p^\ell(X, X')}{p(X)p(X')} \Delta^{-\ell} h(X') \right]} \sqrt{\mathbb{E}\left[\frac{(\Delta^{-\ell} g(X))^2}{\Delta^{-\ell} h(X)} \frac{K_p^\ell(X, X')}{p(X)p(X')} \Delta^{-\ell} h(X') \right]}
\end{aligned}
$$

using (3.21) leads to the inequality.

The only part of the claim that remains to be proved concerns the saturation condition in the inequality. This follows from the Cauchy-Schwarz inequality which is an equality if and only if $\Delta^{-\ell} f(x)/\Delta^{-\ell} h(x) \propto \Delta^{-\ell} g(x')/\Delta^{-\ell} h(x')$ is constant throughout $\mathcal{S}(p)$. This is only possible under the stated condition. □

Remark 4.2.6. *Theorem 4.2.5 can be refined using the exact expression for the remainder in the Cauchy-Schwarz inequality, given by the Lagrange-type identity*

$$
\begin{aligned}
&(\mathbb{E}[f(X_1, X_2) g(X_1, X_2)])^2 \\
&= \mathbb{E}[f^2(X_1, X_2)] \mathbb{E}[g^2(X_1, X_2)] - \frac{1}{2} \mathbb{E}[(f(X_1, X_2) g(X_3, X_4) - f(X_3, X_4) g(X_1, X_2))^2]
\end{aligned}
$$

with X_1, X_2, X_3, X_4 independent copies with density p and $f, g \in L^2(p)$. Fix $h \in L^1(p)$ a decreasing function such that $\|h\|_{\mathcal{S}(p),\infty} < \infty$. For all f, g which satisfy the assumptions of Lemma 3.2.19 we have

$$
\begin{aligned}
&(\mathrm{Cov}[f(X), g(X)])^2 \\
&= \mathbb{E}\left[(\Delta^{-\ell} f(X))^2 \frac{-\mathcal{L}_p^\ell h(X)}{\Delta^{-\ell} h(X)} \right] \mathbb{E}\left[(\Delta^{-\ell} g(X))^2 \frac{-\mathcal{L}_p^\ell h(X)}{\Delta^{-\ell} h(X)} \right] - \frac{1}{2} R(f, g, h)
\end{aligned}
$$

with

$$
\begin{aligned}
R(f, g, h) = \mathbb{E}\left[\left(\Delta^{-\ell} f(X_1) \Delta^{-\ell} g(X_4) - \Delta^{-\ell} f(X_3) \Delta^{-\ell} g(X_2) \frac{\Delta^{-\ell} h(X_1) \Delta^{-\ell} h(X_4)}{\Delta^{-\ell} h(X_2) \Delta^{-\ell} h(X_3)} \right)^2 \right. \\
\left. \frac{\Delta^{-\ell} h(X_2) \Delta^{-\ell} h(X_3)}{\Delta^{-\ell} h(X_1) \Delta^{-\ell} h(X_4)} \frac{K_p^\ell(X_1, X_2) K_p^\ell(X_3, X_4)}{p(X_1) p(X_2) p(X_3) p(X_4)} \right].
\end{aligned}
$$

In particular when $h(x) = x$ the remainder term simplifies to

$$R(f,g,h) = \mathbb{E}\left[\left(\Delta^{-\ell}f(X_1)\Delta^{-\ell}g(X_4) - \Delta^{-\ell}f(X_3)\Delta^{-\ell}g(X_2)\right)^2 \frac{K_p^\ell(X_1,X_2)K_p^\ell(X_3,X_4)}{p(X_1)p(X_2)p(X_3)p(X_4)}\right].$$

Combining Proposition 4.2.1 and Theorem 4.2.5 (applied with $f = g$) we arrive at the following result (applied to a smaller class of functions h) which, as we shall argue below, share a similar flavour to the upper and lower bounds from Theorem 4.1.1.

Corollary 4.2.7 (Klaassen bounds, revisited)**.** *For any decreasing function $h \in L^2(p)$ and all g such that Lemma 3.2.19 applies (with $f = g$), we have*

$$\frac{\mathbb{E}\left[-\mathcal{L}_p^\ell h(X)(\Delta^{-\ell}g(X))\right]^2}{\mathrm{Var}\big(h(X)\big)} \leq \mathrm{Var}[g(X)] \leq \mathbb{E}\left[(\Delta^{-\ell}g(X))^2 \frac{-\mathcal{L}_p^\ell h(X)}{\Delta^{-\ell}h(X)}\right]. \tag{4.15}$$

Equality in the upper bound holds if and only if there exists constants α, β such that $g(x) = \alpha h(x) + \beta$.

Proof. For the lower bound, we apply Proposition 4.2.1 with $f(x) = -\mathcal{L}_p^\ell h(x)$ so that $\mathcal{T}_p^\ell c(x) = h(x) - -\mathbb{E}[h(X)]$. For the upper bound we use Theorem 4.2.5 with $f = g$. □

Example 4.2.8 (Pearson and Ord families)**.** *Tables 4.1, 4.2, and 4.3 present the results for random variables whose distribution belongs to the Pearson and Ord families of distributions. A random variable $X \sim p$ belongs to the* integrated Pearson family *if X is absolutely continuous and there exist $\delta, \beta, \gamma \in \mathbb{R}$ not all equal to 0 such that $\tau_p^\ell(x)\big(:= -\mathcal{L}_p^\ell(\mathrm{Id})\big) = \delta x^2 + \beta x + \gamma$ for all $x \in \mathcal{S}(p)$. Similarly, $X \sim p$ belongs to the cumulative Ord family if X is discrete and there exist $\delta, \beta, \gamma \in \mathbb{R}$ not all equal to 0 such that $\tau_p^\ell(x)\big(:= -\mathcal{L}_p^\ell(\mathrm{Id})\big) = \delta x^2 + \beta x + \gamma$ for all $x \in \mathcal{S}(p)$. In Tables 4.1 and 4.2 the parameters δ, β, γ are defined for $\tau_p^-(x)$. The bounds for these distributions generalize the results e.g. from Afendras et al. (2007). For Integrated Pearson distributions, higher order bounds (that is, bounds in which higher order derivatives of the test functions are considered) are given in Afendras (2013). For the cumulative Ord family we refer to Afendras et al. (2018) for a detailed study of the associated system of orthogonal polynomials.*

Remark 4.2.9 (About the connection with Klaassen's bounds)**.** *The bounds in Corollary 4.2.7 and those from Theorem 4.1.1 are obviously of a similar flavour. Upon closer inspection, however, the connection is not transparent. In order to clarify this point, we follow Klaassen (1985) and restrict our attention to kernels of the*

form

$$\rho_\zeta^+(x,y) = \mathbb{I}[\zeta < y \le x] - \mathbb{I}[x < y \le \zeta] \text{ and } \rho_\zeta^-(x,y) = \mathbb{I}[\zeta \le y < x] - \mathbb{I}[x \le y < \zeta]$$

for some $\zeta \in \mathbb{R}$. *In our notations, these become*

$$\rho_\zeta^\ell(x,y) = \chi^\ell(\zeta,y)\chi^{-\ell}(y,x) - \chi^\ell(x,y)\chi^{-\ell}(y,\zeta)$$

for $\ell \in \{-1,0,1\}$.

We first tackle the relation between the main arguments of the bounds, namely $G(x)$ *and* $g(x)$. *Given a measurable function* g, *we mimic the statement of Theorem 4.1.1 and introduce the generalized primitive* $G(x) = G_\zeta^\ell(x) := \int \rho_\zeta^\ell(x,y)g(y)\mu(\mathrm{d}y) + c$ *with* c *arbitrary, fixed w.l.o.g. to 0. Again in our notations, this becomes*

$$G_\zeta^\ell(x) = \int_{\zeta+a_\ell}^{x-b_\ell} g(y)\mu(\mathrm{d}y)\mathbb{I}[\zeta < x] - \int_{x+a_\ell}^{\zeta-b_\ell} g(y)\mu(\mathrm{d}y)\mathbb{I}[x < \zeta].$$

By construction, $\Delta^{-\ell}G_\zeta^\ell(x) = g(x)$ *for all* ζ *and all* ℓ, *as expected. Nevertheless, in order for* $G_\zeta^\ell(x)$ *to be well-defined, strong (joint) assumptions on* g *and* ζ *are required; for instance, if* $g(x) = 1$ *then* ζ *must be finite and* $G_{\zeta,c}^\ell(x) = x - \zeta$ *while if* $g(x)$ *has* p*-mean 0 then the values* $\zeta = \pm\infty$ *are allowed.*

Next, we examine the connection between the lower bound (4.7) *and the lower bound of* (4.15). *Let* $k \in L^2(p)$ *have* p*-mean 0. Then*

$$\begin{aligned}
\mathbb{E}[k(X)\rho_\zeta^\ell(X,x)] &= \mathbb{E}[k(X)\chi^{-\ell}(x,X)]\chi^\ell(\zeta,x) - \mathbb{E}[k(X)\chi^\ell(X,x)]\chi^{-\ell}(x,\zeta) \\
&= \mathbb{E}[k(X)]\chi^\ell(\zeta,x) - \mathbb{E}[k(X)\chi^\ell(X,x)]\left(\chi^\ell(\zeta,x) + \chi^{-\ell}(x,\zeta)\right) \\
&= -\mathbb{E}[k(X)\chi^\ell(X,x)]
\end{aligned}$$

so that

$$K(x) = \frac{1}{p(x)}\int \rho_\zeta^\ell(z,x)k(z)p(z)\mu(dz) = -\mathcal{L}_p^\ell k(x)$$

and thus (4.7) *follows from the lower bound of* (4.15).

Finally, we consider the upper bounds (4.6) *and* (4.15). *Let* $H(x) = H_\zeta^\ell(x)$ *be a generalized primitive of some nonnegative function* h. *The same manipulations as above lead to*

$$\frac{1}{p(x)}\int \rho_\zeta^\ell(z,x)H(z)p(z)\mu(\mathrm{d}z) = -\mathcal{L}_p^\ell h(x) - \frac{\mathbb{E}[H(X)]}{p(x)}(P(x-a_\ell) - \chi^\ell(\zeta,x)).$$

If, following Klaassen (1985), we choose ζ *in such a way that* $\mathbb{E}[H(X)] = 0$ *(this is equivalent to requiring* $\int_{\zeta+a_\ell}^b h(y)(1-P(y+b_\ell))\mu(\mathrm{d}y) = \int_a^{\zeta-b_\ell} h(y)(P(y-a_\ell))\mu(\mathrm{d}y)$*) then we see that the upper bound in* (4.15) *is equivalent to* (4.6).

Of course there is some gain in generality at allowing for a general kernel ρ as in Theorem 4.1.1, though this comes at the expense of readability: given a positive function h, understanding the form of function H is actually non trivial and our result illuminates Klaassen's discovery by providing the connection with Stein characterizations.

3 About the weights

The freedom of choice in the test functions h appearing in the bounds invite a study of the impact of the choice of h on the validity and quality of the resulting inequalities.

3.1 Score function and a Brascamp-Lieb inequality

The form of the lower bound in Proposition 4.2.1 encourages the choice $f(x) = 1$. This is only permitted if the constant function $1 \in \mathcal{F}_\ell^{(1)}(p)$ and $\mathbb{E}\left[\left(\mathcal{T}_p^\ell 1(X)\right)^2\right] < \infty$; these are two strong assumptions which exclude some natural targets such as e.g. the exponential or beta distributions. If this choice is permitted, then we reap the lower bound

$$\mathrm{Var}[g(X)] \geq \frac{\mathbb{E}[\Delta^{-\ell} g(X)]}{I^\ell(p)}$$

with $I^\ell(p) = \left[\left(\mathcal{T}_p^\ell 1(X)\right)^2\right]$.

The function $\mathcal{T}_p^\ell 1(x) = \Delta^\ell(p(x))/p(x)$ is some form of generalized score function and $I^\ell(p)$ a generalized Fisher information. Indeed, if $\ell = 0$ and $X \sim p$ is absolutely continuous, then $\mathcal{T}_p^\ell 1(x) = (\log p(x))'$ is exactly the (location) score function of p and $I^{(0)}(p)$ is none other than the (location) Fisher information of p. More generally we note that if $1 \in \mathcal{F}_\ell^{(1)}(p)$ then $\mathbb{E}[\mathcal{T}_p^\ell 1(X)] = 0$ and, by Lemma 3.2.5, it satisfies

$$\mathbb{E}[\mathcal{T}_p^\ell 1(X) g(X)] = -\mathbb{E}[\Delta^{-\ell} g(X)]$$

for all appropriate g; this further reinforces the analogy.

The corresponding upper bound from (4.14) is obtained for $h(x) = \mathcal{T}_p^\ell 1(x)$ in (4.15). Suppose that $p(b^- + a_\ell) = p(a^+ - b_\ell) = 0$. By construction, $\mathcal{L}_p^\ell h(x) = \mathbb{I}_{\mathcal{S}(p)}(x)$. If we can further suppose that $\mathcal{T}_p^\ell 1(x)$ is a decreasing function then

$$|\mathrm{Cov}[f(X), g(X)]| \leq \sqrt{\mathbb{E}\left[\frac{(\Delta^{-\ell} f(X))^2}{-\Delta^{-\ell}\mathcal{T}_p^\ell 1(X)}\right]} \sqrt{\mathbb{E}\left[\frac{(\Delta^{-\ell} g(X))^2}{-\Delta^{-\ell}\mathcal{T}_p^\ell 1(X)}\right]}.$$

Taking $g = f$ we deduce the following result whose continuous version (i.e. the case $\ell = 0$) dates back to Brascamp and Lieb (1976).

Corollary 4.3.1 (Brascamp-Lieb inequality)**.** *Under the same conditions as Proposition 3.3.5 we have*

$$\frac{\mathbb{E}\left[(\Delta^{-\ell}g(X))\right]^2}{\mathbb{E}\left[\left(\mathcal{T}_p^\ell 1(X)\right)^2\right]} \leq \mathrm{Var}[g(X)] \leq \mathbb{E}\left[\frac{(\Delta^{-\ell}g(X))^2}{-\Delta^{-\ell}\mathcal{T}_p^\ell 1(X)}\right] \tag{4.16}$$

for all g such that $\mathcal{T}_p^\ell 1, g$ satisfy together the assumptions of Lemma 3.2.19.

Remark 4.3.2. *We refer to* (4.16) *as a "Brascamp-Lieb inequality" because of a result from Brascamp and Lieb (1976) where it is shown that, for a given density p proportional to $\exp(-V)$ with V strictly convex on $\mathbb{R}$ and $V' \in L^2(p)$, we have $\mathrm{Var}[g(X)] \leq \mathbb{E}\left[(g'(X))^2/V''(X)\right]$ where the constant 1 is optimal. Taking $\ell = 0$ in* (4.16)*, one recognizes $-\Delta^{-\ell}\mathcal{T}_p^\ell 1 = V''$ so that the upper bound in* (4.16) *reduces to this Brascamp-Lieb inequality.*

We conclude with a generalized version of the elegant inequality due to Menz and Otto (2013, Lemma 2.11), in the form stated in Carlen et al. (2013, Equation (1.5)).

Corollary 4.3.3 (Asymmetric Brascamp-Lieb inequality)**.** *Under the same conditions as above, if $-\Delta^{-\ell}\mathcal{T}_p^\ell 1 \in L^1(\mu)$ then*

$$|\mathrm{Cov}[f(X), g(X)]| \leq \sup_x \left|\frac{\Delta^{-\ell}f(x)}{\Delta^{-\ell}\mathcal{T}_p^\ell 1(x)}\right| \mathbb{E}\left[\left|\Delta^{-\ell}g(X)\right|\right]$$

for all f, g in $L^2(p)$.

Proof. Under the stated assumptions, we may apply (4.12) to get, after some notational reshuffling,

$$\begin{aligned}
|\mathrm{Cov}[f(X), g(X)]| &\leq \mathbb{E}\left[\frac{\left|\Delta^{-\ell}f(X)\right|}{-\Delta^{-\ell}\mathcal{T}_p^\ell 1(X)}|\Delta^{-\ell}g(X')|\left(-\Delta^{-\ell}\mathcal{T}_p^\ell 1(X)\right)\frac{K_p^\ell(X,X')}{p(X)p(X')}\right] \\
&\leq \sup_x \left|\frac{\Delta^{-\ell}f(x)}{\Delta^{-\ell}\mathcal{T}_p^\ell 1(x)}\right| \mathbb{E}\left[|\Delta^{-\ell}g(X')|\frac{K_p^\ell(X,X')}{p(X)p(X')}\left(-\Delta^{-\ell}\mathcal{T}_p^\ell 1(X)\right)\right] \\
&= \sup_x \left|\frac{\Delta^{-\ell}f(x)}{\Delta^{-\ell}\mathcal{T}_p^\ell 1(x)}\right| \mathbb{E}\left[\left|\Delta^{-\ell}g(X')\right|\right]
\end{aligned}$$

where the last line follows by conditioning on X' and applying Proposition 3.3.5. □

3.2 Stein kernel and Cacoullos' bound

It is natural to consider test function $h = -\mathrm{Id}$ in Theorem 4.2.5. Since $\Delta^{-\ell}h(x) = 1$, we obtain

$$|\mathrm{Cov}[f(X), g(X)]| \leq \sqrt{\mathbb{E}\left[\tau_p^\ell(X)(\Delta^{-\ell}f(X))^2\right]}\sqrt{\mathbb{E}\left[\tau_p^\ell(X)(\Delta^{-\ell}g(X))^2\right]}$$

In particular if $g = f$ then

$$\mathrm{Var}[g(X)] \leq \mathbb{E}\left[\tau_p^\ell(X)(\Delta^{-\ell}g(X))^2\right]$$

in which one recognizes the upper bounds from Cacoullos (1982) and also, when $\ell = 0$, Saumard (2019). The corresponding lower bound in (4.11) is obtained for $f(x) = \tau_p^\ell(x)$ for which $\mathcal{T}_p^\ell f(x) = x\mathbb{I}_{\mathcal{S}(p)}(x)$, and the overall bound becomes

$$\frac{\mathbb{E}\left[\tau_p^\ell(X)(\Delta^{-\ell}g(X))\right]^2}{\mathrm{Var}[X]} \leq \mathrm{Var}[g(X)] \leq \mathbb{E}\left[(\Delta^{-\ell}g(X))^2\tau_p^\ell(X)\right]. \tag{4.17}$$

Example 4.3.4. *In our examples* (4.17) *gives the following covariance identities.*

- *Binomial distribution: Let* $X \sim \mathrm{Bin}(n,\theta)$ *as in Example 3.2.13. From Example 3.2.17 we obtain the upper and lower bounds*

$$\frac{(1-\theta)}{n\theta}\mathbb{E}\left[X\Delta^- g(X)\right]^2 \leq \mathrm{Var}[g(X)] \leq (1-\theta)\mathbb{E}\left[X(\Delta^- g(X))^2\right];$$
$$\frac{\theta}{n(1-\theta)}\mathbb{E}\left[(n-X)\Delta^+ g(X)\right]^2 \leq \mathrm{Var}[g(X)] \leq \theta\mathbb{E}\left[(n-X)(\Delta^+ g(X))^2\right].$$

- *Beta distribution: From Example 4.2.4, for the* $\mathrm{Beta}(\alpha,\beta)$*-distribution with variance* $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$,

$$\frac{(\alpha+\beta+1)}{\alpha\beta}\mathbb{E}\left[X(1-X)g'(X)\right]^2 \leq \mathrm{Var}[g(X)] \leq \frac{1}{\alpha+\beta}\mathbb{E}\left[X(1-X)(g'(X))^2\right].$$

The particular case of other Pearson/Ord distributions is detailed in Tables 4.1, 4.2 and 4.3. The tables include the Binomial distribution and the Beta distribution for easy reference. The Stein operators which are given are those from Example 3.2.18. We remark that Cacoullos' inequality was obtained, in Furioli et al. (2017), through an argument which is some form of dual to that outlined in this section, for continuous distributions. In that article the inequality is coined a "weighted Chernov bound"; we refer to Toscani (2019) as well for extensions towards stable densities.

3.3 Discussion

The theory presented in this chapter is closely connected to several classical topics from functional analysis. First, as already mentioned in the introduction, there is a connection with the *spectral gap and Poincaré inequalities*. Let σ^2 be a positive function on $\mathbb{R}$; the σ^2-weighted Poincaré constant (or spectral gap) of a density p is

$$\lambda_{\sigma^2,p} = \inf_{g\in C_0^\infty(\mathbb{R})} \frac{\mathbb{E}\left[\sigma^2(X)(g'(X))^2\right]}{\mathrm{Var}[g(X)]}.$$

The case $\sigma^2 = 1$ leads to the classical (unweighted) Poincaré inequality or spectral gap; if there exists a function g_{opt} achieving equality, one says that the inequality is saturated at g_{opt}. It is an easy matter to use our notations to extend the above to non absolutely continuous distributions. Exploiting the freedom of choice in the test functions h in Corollary 4.2.7 immediately yields the next result.

Corollary 4.3.5. *Instate all previous assumptions and notations. Then*

$$\lambda_{\sigma^2,p} \geq \left(\sup_h \sup_x \left\{ -\mathcal{L}_p^\ell h(x) / \left(\sigma^2(x) \Delta^{-\ell} h(x) \right\} \right)^{-1}$$

where the supremum is taken over all decreasing functions $h \in L^2(p)$.

It would be of interest to study the connection with the works Bonnefont et al. (2016) and Roustant et al. (2017), and further study the important problem of saturation of the inequalities.

Finally, a connection between variance bounds with Stein's method was already noted e.g. in Diaconis and Zabell (1991). It arises naturally in our context by choosing h in Theorem 4.2.5 such that the corresponding weight $-(\Delta^{-\ell} h(x))^{-1} \mathcal{L}_p^\ell h(x)$ is constant, i.e. any mean zero function h such that there exists $\lambda \in \mathbb{R}$ for which

$$\frac{-\mathcal{L}_p^\ell h(x)}{\Delta^{-\ell} h(x)} = \lambda \text{ for all } x \in \mathcal{S}(p).$$

By construction and Lemma 3.2.7, such functions are solution to the eigenfunction problem

$$h(x) = -\lambda \mathcal{T}_p^\ell (\Delta^{-\ell} h)(x) \quad \text{for all } x \in \mathcal{S}(p)$$

where operator $\mathcal{R}_p^\ell h := \mathcal{T}_p^\ell (\Delta^{-\ell} h)$ is self-adjoint in the sense of that

$$\mathbb{E}[(\mathcal{R}_p^\ell f(X)) g(X)] = \mathbb{E}[f(X)(\mathcal{R}_p^\ell g(X))]$$

for all f, g such that Lemmas 3.2.5 and 3.2.19 apply.

name **parameter**	**p.m.f.** $p(x)$ **support**	**Stein kernel** $\tau^{\ell}(x)$ **Cum. Ord relation**
Poisson (λ)	$e^{-\lambda}\lambda^x/x!$	$\tau^-(x) = \lambda$
$\lambda > 0$	$x = 0, 1, \ldots$	$\tau^+(x) = x$
		$(\delta, \beta, \gamma) = (0, 0, \lambda)$
	Stein operators	$\mathcal{A}^+_{\mathrm{Poi}(\lambda)} g(x) = (x-\lambda)g(x) - x\Delta^- g(x)$
		$\mathcal{A}^-_{\mathrm{Poi}(\lambda)} g(x) = (x-\lambda)g(x) - \lambda\Delta^+ g(x)$
		$\mathcal{A}_{\mathrm{Poi}(\lambda)} g(x) = xg(x) - \lambda g(x+1)$
	Variance bounds	$\lambda\mathbb{E}\left[\Delta^+ g(X)\right]^2 \leq \mathrm{Var}[g(X)] \leq \lambda\mathbb{E}\left[(\Delta^+ g(X))^2\right]$
		$\lambda^{-1}\mathbb{E}\left[X\Delta^- g(X)\right]^2 \leq \mathrm{Var}[g(X)] \leq \mathbb{E}\left[X(\Delta^- g(X))^2\right]$
Binomial (n, θ)	$\binom{n}{x}\theta^x(1-\theta)^{n-x}$	$\tau^-(x) = \theta(n-x)$
$0 < \theta < 1$	$x = 0, 1, \ldots, n$	$\tau^+(x) = (1-\theta)x$
$n = 1, 2, \ldots$		$(\delta, \beta, \gamma) = (0, -\theta, n\theta)$
	Stein operators	$\mathcal{A}^+_{\mathrm{bin}(n,\theta)} g(x) = (x - n\theta)g(x) - (1-\theta)x\Delta^- g(x)$
		$\mathcal{A}^-_{\mathrm{bin}(n,\theta)} g(x) = (x - n\theta)g(x) - \theta(n-x)\Delta^+ g(x)$
		$\mathcal{A}_{\mathrm{bin}(n,\theta)} g(x) = xg(x) + \frac{\theta}{1-\theta}(n-x)g(x+1)$
	Variance bounds	$\frac{\theta}{n(1-\theta)}\mathbb{E}\left[(n-X)\Delta^+ g(X)\right]^2 \leq \mathrm{Var}[g(X)] \leq \theta\mathbb{E}\left[(n-X)(\Delta^+ g(X))^2\right]$
		$\frac{1-\theta}{n\theta}\mathbb{E}\left[X\Delta^- g(X)\right]^2 \leq \mathrm{Var}[g(X)] \leq (1-\theta)\mathbb{E}\left[X(\Delta^- g(X))^2\right]$
Negative Binomial (r, p)	$\binom{r+x-1}{x}p^r(1-p)^x$	$\tau^-(x) = \frac{1-p}{p}(r+x)$
$0 < p < 1$	$x = 0, 1, \ldots$	$\tau^+(x) = \frac{1}{p}x$
$r > 0$		$(\delta, \beta, \gamma) = (0, \frac{1-p}{p}, r\frac{1-p}{p})$
	Stein operators	$\mathcal{A}^+_{\mathrm{NB}(r,p)} g(x) = \left(x - \frac{1-p}{p}r\right) g(x) - \frac{x}{p}\Delta^- g(x)$
		$\mathcal{A}^-_{\mathrm{NB}(r,p)} g(x) = \left(x - \frac{1-p}{p}r\right) g(x) - \frac{1-p}{p}(r+x)\Delta^+ g(x)$
		$\mathcal{A}_{\mathrm{NB}(r,p)} g(x) = xg(x) - (1-p)(r+x)g(x+1)$
	Variance bounds	$\frac{1-p}{r}\mathbb{E}[(X+r)\Delta^+ g(X)]^2 \leq \mathrm{Var}[g(X)] \leq \frac{1-p}{p}\mathbb{E}[(X+r)(\Delta^+ g(X))^2]$
		$\frac{1}{r(1-p)}\mathbb{E}[X\Delta^- g(X)]^2 \leq \mathrm{Var}[g(X)] \leq \frac{1}{p}\mathbb{E}[X(\Delta^- g(X))^2]$

Table 4.1: Specific forms some discrete distributions from the cumulative Ord family. This table is a completed version of Table 1 of Afendras et al. (2007).

name / **parameter**	**p.m.f.** $p(x)$ / **support**	**Stein kernel** $\tau^\ell(x)$ / **Cum. Ord relation**
Hypergeometric (n, K, N) $1 \leq K \leq N$ $n = 1, 2, \ldots, N$	$\binom{K}{x}\binom{N-K}{n-x}(\binom{N}{n})^{-1}$ $0 \leq x \leq \min\{K, n\}$	$\tau^-(x) = \frac{1}{N}(K-x)(n-x)$ $\tau^+(x) = \frac{1}{N}x(x+N-K-n)$ $(\delta, \beta, \gamma) = \left(\frac{1}{N}, \frac{-(n+K)}{N}, \frac{nK}{N}\right)$
	Stein operators	$\mathcal{A}^+_{\mathrm{H}(n,K,N)}g(x) = \left(x - \frac{nK}{N}\right)g(x) - \frac{1}{N}x(N-K-n+x)\Delta^- g(x)$
		$\mathcal{A}^-_{\mathrm{H}(n,K,N)}g(x) = \left(x - \frac{nK}{N}\right)g(x) - \frac{1}{N}(K-x)(n-x)\Delta^+ g(x)$
		$\mathcal{A}_{\mathrm{H}(n,K,N)}g(x) = xg(x) + \frac{1}{N+K+n}x^2 g(x) - \frac{1}{N+K+n}(K-x)(n-x)g(x+1)$
	Variance bounds	$\frac{N-1}{nK(N-K)(N-n)}\mathbb{E}[(K-X)(n-X)\Delta^+ g(X)]^2 \leq \mathrm{Var}[g(X)]$
		$\mathrm{Var}[g(X)] \leq \frac{1}{N}\mathbb{E}[(K-X)(n-X)(\Delta^+ g(X))^2]$
		$\frac{N-1}{nK(N-K)(N-n)}\mathbb{E}[X(N-K-n+X)\Delta^- g(X)]^2 \leq \mathrm{Var}[g(X)] \leq$
		$\mathrm{Var}[g(X)] \leq \frac{1}{N}\mathbb{E}[X(N-K-n+X)(\Delta^- g(X))^2]$
Negative Hyper-geometric (N, K, r) $0 \leq K \leq N$	$\frac{\binom{x+r-1}{x}\binom{N-r-x}{K-x}}{\binom{N}{K}}$ $x = 0, 1, \ldots, K$	$\tau^-(x) = \frac{1}{N-K+1}(K-x)(r+x)$ $\tau^+(x) = \frac{1}{N-K+1}x(N-r+1-x)$ $(\delta, \beta, \gamma) = \left(\frac{-1}{N-K+1}, \frac{K-r}{N-K+1}, \frac{rK}{N-K+1}\right)$
	Stein operators	$\mathcal{A}^+_{\mathrm{NH}(N,K,r)}g(x) = \left(x - \frac{rK}{N-K+1}\right)g(x) - \frac{x(N+1-r-x)}{N-K+1}\Delta^- g(x)$
		$\mathcal{A}^-_{\mathrm{NH}(N,K,r)}g(x) = \left(x - \frac{rK}{N-K+1}\right)g(x) - \frac{(K-x)(r+x)}{N-K+1}\Delta^+ g(x)$
		$\mathcal{A}_{\mathrm{NH}(N,K,r)}g(x) = xg(x) - \frac{1}{N-r+1}x^2 g(x) - \frac{1}{N-r+1}(K-x)(r+x)g(x+1)$
	Variance bounds	$\frac{N-K+2}{r(N+1)K(N-K-r+1)}\mathbb{E}[(K-X)(r+X)\Delta^+ g(X)]^2 \leq \mathrm{Var}[g(X)]$
		$\mathrm{Var}[g(X)] \leq \frac{1}{N-K+1}\mathbb{E}[(K-X)(r+X)(\Delta^+ g(X))^2]$
		$\frac{N-K+2}{r(N+1)K(N-K-r+1)}\mathbb{E}[X(N+1-r-X)\Delta^- g(X)]^2 \leq \mathrm{Var}[g(X)]$
		$\mathrm{Var}[g(X)] \leq \frac{1}{N-K+1}\mathbb{E}[X(N+1-r-X)(\Delta^- g(X))^2]$

Table 4.2: Specific form for some discrete distributions from the cumulative Ord family (second part).

name **parameter**	**p.m.f.** $p(x)$ **support**	$\tau(x)$ **Pearson relation**
Normal(μ, σ^2) $\mu \in \mathbb{R}$, $\sigma^2 > 0$	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$ $x \in \mathbb{R}$	$\tau(x) = \sigma^2$ $(\delta, \beta, \gamma) = (0, 0, \sigma^2)$
	Stein operators	$\mathcal{A}_{\mathrm{N}(\mu,\sigma^2)} g(x) = (x-\mu)g(x) - \sigma^2 g'(x)$
	Variance bounds	$\sigma^2 \mathbb{E}[g'(X)]^2 \leq \mathrm{Var}[f(X)] \leq \sigma^2 \mathbb{E}[g'(X)^2]$
Beta (α, β) $\alpha > 0$, $\beta > 0$	$x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha, \beta)$ $x \in (0, 1)$	$\tau(x) = \frac{x(1-x)}{\alpha+\beta}$ $(\delta, \beta, \gamma) = (\frac{-1}{\alpha+\beta}, \frac{1}{\alpha+\beta}, 0)$
	Stein operators	$\mathcal{A}_{\mathrm{Beta}(\alpha,\beta)} g(x) = \left(x - \frac{\alpha}{\alpha+\beta}\right) g(x) - \frac{x(1-x)}{\alpha+\beta} g'(x)$
	Variance bounds	$\frac{(\alpha+\beta+1)}{\alpha\beta} \mathbb{E}\left[X(1-X)g'(X)\right]^2 \leq \mathrm{Var}[g(X)] \leq \frac{1}{\alpha+\beta} \mathbb{E}\left[X(1-X)(g'(X))^2\right]$
Gamma(μ, σ^2) $\alpha > 0$, $\beta > 0$	$x^{\alpha-1}\beta^{-\alpha} e^{-x/\beta}/\Gamma(\alpha)$ $x \in (0, \infty)$ $(\alpha < 1)$ $x \in [0, \infty)$ $(\alpha \geq 1)$	$\tau(x) = \beta x$ $(\delta, \beta, \gamma) = (0, \beta, 0)$
	Stein operators	$\mathcal{A}_{\mathrm{Gamma}(\alpha,\beta)} g(x) = (x - \alpha\beta)g(x) - \beta x g'(x)$
	Variance bounds	$\frac{1}{\alpha} \mathbb{E}[X g'(X)]^2 \leq \mathrm{Var}[g(X)] \leq \beta \mathbb{E}[X g'(X)^2]$
Student (ν) $\nu > 0$	$\frac{(\nu/(\nu+x^2))^{(1+\nu)/2}}{\nu^{1/2} B(\nu/2, 1/2)}$ $x \in \mathbb{R}$	$\tau(x) = \frac{x^2+\nu}{\nu-1}$ for $\nu > 1$ $(\delta, \beta, \gamma) = \left(\frac{1}{\nu-1}, 0, \frac{\nu}{\nu-1}\right)$
	Stein operators	$\mathcal{A}_{\mathrm{t}(\nu)} g(x) = x g(x) - \frac{x^2+\nu}{\nu-1} g'(x)$
	Variance bounds ($\nu > 2$)	$\frac{(\nu-2)}{\nu(\nu-1)^2} \mathbb{E}[(X^2+\nu)g'(X)]^2 \leq \mathrm{Var}[g(X)] \leq \frac{1}{\nu-1} \mathbb{E}[(X^2+\nu)(g'(X)^2)]$
F distribution (d_1, d_2) $d_1 > 0$, $d_2 > 0$	$\frac{\left(\frac{d_1}{d_2}\right)^{d_1/2} x^{\frac{d_1}{2}-1} \left(1+\frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}}}{B(d_1/2, d_2/2)}$ $x \in (0, \infty)$	$\tau(x) = \frac{2x(d_1 x + d_2)}{d_1(d_2-2)}$ for $d_2 > 2$ $(\delta, \beta, \gamma) = \left(\frac{2d_1}{d_1(d_2-2)}, \frac{2d_2}{d_1(d_2-2)}, 0\right)$
	Stein operators	$\mathcal{A}_{\mathrm{F}(\mathrm{d}_1,\mathrm{d}_2)} g(x) = \left(x - \frac{d_2}{d_2-1}\right) g(x) - \frac{2x(d_2+d_1 x}{d1(d_2-2)} g'(x)$
	Variance bounds ($d_2 > 4$)	$\frac{2(d_2-4)}{d_1 d_2^2 (d_1+d_2-2)} \mathbb{E}[X(d_2 + d_1 X)g'(X)]^2 \leq \mathrm{Var}[g(X)]$ $\mathrm{Var}[g(X)] \leq \frac{2}{d_1(d_2-2)} \mathbb{E}[X(d_2 + d_1 X)g'(X)^2]$

Table 4.3: Specific form for some continuous distributions from the Pearson family. This table is an adapted version of Table 1 of Afendras et al. (2007) using examples from Afendras and Papadatos (2011).

CHAPTER 5

Infinite covariance expansions

1 Introduction

The starting point of this chapter is the famous Gaussian expansion which states that if $N \sim \mathcal{N}(0,1)$, then

$$\mathrm{Var}[g(N)] = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k!} \mathbb{E}\left[g^{(k)}(N)^2\right] \tag{5.1}$$

for all smooth functions $g : \mathbb{R} \to \mathbb{R}$ such that all the expectations exist. Expansion (5.1), whose first order term yields an upper variance bound generalizing Chernoff (1981)'s famous Gaussian bound, has been obtained in a number of different (and often non equivalent) ways. It is proved in Houdré and Kagan (1995) via orthogonality properties of Hermite polynomials, and extensions to multivariate and infinite dimensional settings are given in Houdré and Pérez-Abreu (1995) and Houdré et al. (1998).

Chen (1985) uses martingale and stochastic integrals to obtain a general version of (5.1) (also valid on certain manifolds). The expansion is contextualized in Ledoux (1995) through properties of the Ornstein-Uhlenbeck operator, and it is also shown in that paper that the semi-group arguments carry through to non-Gaussian target distributions under general assumptions. A very general approach to this line of research can be found in Houdré et al. (1998) where similar expansions are obtained by means of an iteration of an interpolation formula for infinitely divisible distributions. The main difference between the univariate standard Gaussian and the general non-Gaussian target is that the explicit weight sequence and simple iterated derivatives appearing in (5.1) need to be replaced by some well-chosen iterated gradients

with weight sequences which can be quite difficult to obtain explicitly (for instance Ledoux' sequence is an iteration of the "carré du champ" operator).

The above references are predated by Papathanasiou (1988) wherein a general version of (5.1) (valid for arbitrary continuous target distributions) is obtained through elementary arguments relying on an iteration of the exact Cauchy-Schwarz equality (via the so-called *Mohr and Noll identity* from Mohr and Noll, 1952) combined with the Lagrange identity for integrals due to Cacoullos and Papathanasiou (1985). Papathanasiou's method of proof is extended in Afendras et al. (2007) to encompass discrete distributions. Both the continuous and discrete expansions are of the same form as (5.1), although the weight sequence $(-1)^k/k!$ is replaced with a target-specific explicit sequence of weights (see equations (5.4) and (5.5) below). To set the scene, we use notation introduced in Chapter 3 which allows to unify the presentation of the results from Papathanasiou (1988) and Afendras et al. (2007).

Notation: For a function $f : \mathbb{R} \to \mathbb{R}$ let $\Delta^\ell f(x) = (f(x+\ell) - f(x))/\ell$ for all $\ell \in \{-1, 0, 1\}$, with the convention that $\Delta^0 f(x) = f'(x)$, with $f'(x)$ the weak derivative defined Lebesgue almost everywhere. The case $\ell = 0$ is referred to as the *continuous case* and $\ell \in \{-1, 1\}$ is referred to as the *discrete case*. For a real-valued function f, in the continuous case $f^{(k)}$ denotes its k^{th} derivative; discrete higher order derivatives $f^{(k)}$ are obtained by iterating the forward derivative $\Delta^+ f(x) = f(x+1) - f(x)$. We use the rising and falling factorial notation

$$f^{[k]}(x) = \prod_{j=0}^{k-1} f(x+j) \text{ and } f_{[k]}(x) = \prod_{j=0}^{k-1} f(x-j), \tag{5.2}$$

with the convention that $f^{[0]}(x) = f_{[0]}(x) = 1$.

Expansion (5.1) can then be seen as a particular instance of the following result (see Papathanasiou, 1988, Theorem 1 and Corollary 1 and Afendras et al., 2007, Theorem 3.1).

Theorem 5.1.1 (Papathanasiou's expansion)**.** *Let X be a random variable with finite $(n+2)^{th}$ moments. Let g be a real-valued function with finite variance with respect to X. Then*

$$\mathrm{Var}[g(X)] = \sum_{k=1}^{n} (-1)^{k-1} \mathbb{E}\left[(g^{(k)}(X))^2 \Gamma_k(X)\right] + (-1)^n R_n \tag{5.3}$$

where R_n is a non-negative remainder term and Γ_k depend on the type of distribution, as follows.

1. *If X is a real random variable with continuous probability density function (pdf) p, then the weights are*

$$\Gamma_k(t) = \frac{(-1)^{k-1}}{k!(k-1)!p(t)}\Big(\mathbb{E}\left[(X-t)^k\right]\int_{-\infty}^{t}(x-t)^{k-1}p(x)dx - \mathbb{E}\left[(X-t)^{k-1}\right]\int_{-\infty}^{t}(x-t)^k p(x)dx\Big), \tag{5.4}$$

   defined for all t such that $p(t) > 0$.

2. *If X is an integer-valued r.v. with probability mass function (pmf) p, then the weights are*

$$\Gamma_k(t) = \frac{(-1)^{k-1}}{k!(k-1)!p(t)}\Big(\mathbb{E}\left[(X-t)_{[k]}\right]\sum_{x<t+1}p(x)(x-(t+1))_{[k-1]} - \mathbb{E}\left[(X-(t+1))_{[k-1]}\right]\sum_{x<t}p(x)(x-t)_{[k]}\Big), \tag{5.5}$$

   defined for all t such that $p(t) > 0$.

It is not hard to show that when $X \sim \mathcal{N}(0,1)$, the weight sequence (5.4) simplifies to $\Gamma_k(t) = 1/k!$ so that (5.3) indeed contains (5.1). More generally, it is shown in Johnson (1993) that if p belongs to the Integrated Pearson (IP) system of distributions (see Definition 4.2.8) then the weights take on a particularly agreeable form, namely $\Gamma_k(t) = \Gamma_1(x)^k/(k!\prod_{j=0}^{k}(1-j\delta))$ and $\delta = \Gamma_1''(x)$ (which is constant if X is Integrated Pearson); many familiar univariate distributions belong to the IP system, such as the normal, beta, gamma, and Student distributions. Similarly as in the continuous case, it is shown by Afendras et al. (2007, Corollary 4.1) that if X belongs to the cumulative Ord family with parameter (δ, β, γ) defined in Definition 4.2.8, then the weights in (5.5) are $\Gamma_k(t) = \Gamma_1^{[k]}(t)/\big(k!\prod_{j=0}^{k}(1-j\delta)\big)$. Like its continuous counterpart, the discrete IP system also contains many familiar univariate distributions such as the binomial, Poisson and geometric distributions.

The list of references presented so far is anything but exhaustive and expansions inspired from (5.1) have attracted a lot of attention over the years, e.g. with extensions to matrix inequalities (Olkin and Shepp, 2005, Wei and Zhang, 2009, Afendras and Papadatos, 2011), to stable distributions (Koldobsky and Montgomery-Smith, 1996), to Bernoulli random vectors (Bobkov et al., 2001); more references shall be provided in the text. Aside from their intrinsic interest, they have many applications and are closely connected to a wide variety of profound mathematical questions. For statistical inference purposes, they can be used in the study of the variance of classes of estimators (see e.g. section 5 Afendras et al., 2007), of copulas (Cuadras and Cuadras, 2008), for problems related to superconcentration (Chatterjee, 2014b, Tanguy, 2017)

or for the study of correlation inequalities (Houdré et al., 1998, Blázquez and Miño, 2014). These expansions can also interpreted as refined log-Sobolev, Poincaré or isoperimetric inequalities (see Saumard, 2019). The weights appearing in the first order ($n = 1$) bounds are crucial quantities in Stein's method (Fathi, 2019, Ledoux et al., 2015) and their higher order extensions are closely connected to eigenvalues and eigenfunctions of certain differential operators (Chen, 1985).

In the present chapter, we combine the method from Papathanasiou (1988), Afendras et al. (2007) with intuition from Klaassen (1985) (and Chapter 4) to unify and extend the results from Theorem 5.1.1 to arbitrary targets under very weak assumptions. The result is given in Theorem 5.3.1 and can be briefly sketched in a simplified form as follows. Fix $(\ell_k)_{k\geq 1}$ a sequence either in $\{-1, 1\}$ or $\{0\}$ and let $h : \mathbb{R} \to \mathbb{R}$ be such that $\Delta^{-\ell_i} h \geq 0$ for all $i \geq 1$. Starting with some functions $f, g : \mathbb{R} \to \mathbb{R}$, we recursively define the sequence $(f_k)_{k\geq 0}$ (resp., $(g_k)_{k\geq 0}$) by $f_0(x) = f(x)$ (resp., $g_0(x) = f(x)$) and $f_i(x) = \Delta^{-\ell} f_{i-1}(x)/\Delta^{-\ell} h(x)$ (resp., $g_i(x) = \Delta^{-\ell} g_{i-1}(x)/\Delta^{-\ell} h(x)$) for all $x \in \mathcal{S}(p)$. Then, for all $n \geq 1$, it holds that if the expectations below are finite then

$$\begin{aligned}&\mathrm{Cov}\left[f(X), g(X)\right]\\ &= \sum_{k=1}^{n} (-1)^{k-1} \mathbb{E}\left[\Delta^{-\ell_k} f_{k-1}(X) \Delta^{-\ell_k} g_{k-1}(X) \frac{\Gamma_k^{\ell}(h)(X)}{\Delta^{-\ell} h(X)}\right] + (-1)^n R_n^{\ell}(h) \qquad (5.6)\end{aligned}$$

where the weight sequences $\Gamma_k^{\ell}(h)$ as well as the non-negative remainder term $R_n^{\ell}(h)$ are given explicitly (see Theorem 5.3.1) and in many cases have a simple form (see Section 4). The expansions from Theorem 5.1.1 are recovered by setting $f = g$, and $h(x) = \mathrm{Id}(x)$ (the identity function) and, in the discrete case, $\ell = -1$. Far from obscuring the message, expansion (5.6), and its more general form provided in Theorem 5.3.1, shed new light on the expansion (5.3) and its available extensions by bringing a new interpretation to the weight sequences in terms of explicit iterated integrals and sums. This is the topic of Section 4. Our results also inscribe the topic within a context which is familiar to practitioners of the famous Stein's method. This last connection nevertheless remains slightly mysterious and will be studied in detail in future contributions.

2 A probabilistic Lagrange inequality

The first ingredient for our results is the following covariance representation For the sake of readability, all proofs are relegated to Appendix.

Lemma 5.2.1. *Let $X \sim p$ with support $\mathcal{S}(p)$. If X_1, X_2 are independent copies of X then*

$$\mathrm{Cov}[f(X), g(X)] = \mathbb{E}\left[\big(f(X_2) - f(X_1)\big)\big(g(X_2) - g(X_1)\big)\mathbb{I}[X_1 < X_2]\right] \tag{5.7}$$

$$= \frac{1}{2}\mathbb{E}\left[\big(f(X_2) - f(X_1)\big)\big(g(X_2) - g(X_1)\big)\right] \tag{5.8}$$

for all $f, g \in L^2(p)$.

A simple representation such as (5.7) is obviously not new, per se; see e.g. the variance expression in Miclo (2008, page 122). In fact, treating the discrete and continuous cases separately, one could also obtain identity (5.7) as a direct application of Lagrange's identity (a.k.a. the Cauchy-Schwarz inequality with remainder) which reads, in the finite discrete case, as

$$\left(\sum_{k=u}^{v} a_k^2\right)\left(\sum_{k=u}^{v} b_k^2\right) - \left(\sum_{k=u}^{v} a_k b_k\right)^2 = \sum_{i=u}^{v-1}\sum_{j=i+1}^{v} (a_i b_j - a_j b_i)^2. \tag{5.9}$$

Using $a_k = g(k)\sqrt{p(k)}$ and $b_k = \sqrt{p(k)}$ for $k = 0, \ldots, n$, identity (5.7) follows in the finite case. Identity (5.9) and its continuous counterpart will play a crucial role in the sequel. As it turns out, they are more suited to our cause under the following form.

Lemma 5.2.2 (A probabilistic Lagrange identity)**.** *Fix some integer $r \in \mathbb{N}_0$ and introduce the (column) vector $\mathbf{v}(x) = (v_1(x), \cdots, v_r(x))' \in \mathbb{R}^r$. Also let $g : \mathbb{R} \to \mathbb{R}$ be any function such that $v_k g \in L^1(p)$ for all $k = 1, \ldots, r$. Then*

$$\mathbb{E}\left[\mathbf{v}(X)g(X)\Phi_p^\ell(u, X, v)\right]\mathbb{E}\left[\mathbf{v}'(X)g(X)\Phi_p^\ell(u, X, v)\right]$$
$$= \mathbb{E}\left[\mathbf{v}(X)\mathbf{v}'(X)\Phi_p^\ell(u, X, v)\right]\mathbb{E}\left[g^2(X)\Phi_p^\ell(u, X, v)\right] - R^\ell(u, v; \mathbf{v}, g), \tag{5.10}$$

where $R^\ell(u, v; \mathbf{v}, g)$ is the $r \times r$ matrix given by

$$R^\ell(u, v; \mathbf{v}, g) = \mathbb{E}\left[(\mathbf{v}_3 g_4 - \mathbf{v}_4 g_3)(\mathbf{v}_3 g_4 - \mathbf{v}_4 g_3)'\Phi_p^\ell(u, X_3, X_4, v)\right] \tag{5.11}$$

with

$$\Phi_p^\ell(u, x_1, x_2, v) = \frac{\chi^\ell(u, x_1)\chi^{\ell^2}(x_1, x_2)\chi^{-\ell}(x_2, v)}{p(x_1)p(x_2)}. \tag{5.12}$$

Here X_3, X_4 denote two independent copies of X and $\mathbf{v}_j = \mathbf{v}(X_j)$ so that $v_{ij} = v_i(X_j)$, and $g_j = g(X_j)$, $i = 3, 4$. If the context is clear, we abbreviate $R^\ell(u, v; \mathbf{v}, g)$ by $R(u, v)$.

3 Papathanasiou-type expansion

Now the necessary ingredients are available to give the main result of this chapter. We use the notation that for a vector $\mathbf{v} = (v_1, \dots, v_r)'$ of functions, the operator Δ^ℓ operates on each component, so that $\Delta^\ell \mathbf{v} = (\Delta^\ell v_1, \dots, \Delta^\ell v_r)'$.

Theorem 5.3.1. *Fix $\ell \in \{-1, 0, 1\}$ and let $\boldsymbol{\ell} = (\ell_n)_{n\geq 1}$ be a sequence such that $\ell_n = 0$ for all n if $\ell = 0$, otherwise $\ell_n \in \{-1, 1\}$ arbitrarily chosen. Let $(h_n)_{n\geq 1}$ be a sequence of real valued functions $h_i : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{P}[\Delta^{-\ell_i} h_i(X) > 0] = 1$ for all $i \geq 1$. Starting with some function $\mathbf{g} : \mathbb{R} \to \mathbb{R}^r$, we recursively define the sequence $(\mathbf{g}_k)_{k\geq 0}$ by $\mathbf{g}_0(x) = \mathbf{g}(x)$ and $\mathbf{g}_i(x) = \Delta^{-\ell_i}\mathbf{g}_{i-1}(x)/\Delta^{-\ell_i} h_i(x)$ for all $x \in \mathcal{S}(p)$. For any sequence $(x_j)_{j\geq 1}$ we let $\Phi_0^{\boldsymbol{\ell}}(x_1, x_2) = 1$ and*

$$\begin{aligned}\Phi_n^{\boldsymbol{\ell}}&(x_1, x_3, \dots, x_{2n-1}, x_{2n+1}, x_{2n+2}, x_{2n}, \dots, x_2)\\ &= \frac{1}{\prod_{i=3}^{2n+2} p(x_i)} \chi^{\ell^2}(x_{2n+1}, x_{2n+2}) \prod_{i=1}^{n} \chi^{\ell_i}(x_{2i-1}, x_{2i+1}) \chi^{-\ell_i}(x_{2i+2}, x_{2i}). \quad (5.13)\end{aligned}$$

Then, for all vectors of functions $\mathbf{f} : \mathbb{R} \to \mathbb{R}^r$ such that the expectations below exist, and all $n \geq 1$, we have

$$\operatorname{Cov}[\mathbf{f}(X)] = \sum_{k=1}^{n} (-1)^{k-1} \mathbb{E}\left[\Delta^{-\ell_k}\mathbf{f}_{k-1}(X) \Delta^{-\ell_k}\mathbf{f}'_{k-1}(X) \frac{\Gamma_k^{\boldsymbol{\ell}}\mathbf{h}(X)}{\Delta^{-\ell_k} h_k(X)}\right] + (-1)^n R_n^{\boldsymbol{\ell}}(\mathbf{h}) \quad (5.14)$$

where the derivatives are taken component-wise, and the weight sequences are

$$\begin{aligned}\Gamma_k^{\boldsymbol{\ell}}\mathbf{h}(x) = \mathbb{E}\Bigg[&(h_k(X_{2k}) - h_k(X_{2k-1})) \prod_{i=1}^{k-1} \Delta^{-\ell_i} h_i(X_{2i+1}, X_{2i+2})\\ &\Phi_p^{\ell_k}(x_{2k-1}, x, x_{2k}) \Phi_{k-1}^{\boldsymbol{\ell}}(X_1, \dots, X_{2k-1}, X_{2k}, \dots, X_2)\Bigg] \quad (5.15)\end{aligned}$$

and

$$\begin{aligned}R_n^{\boldsymbol{\ell}}(\mathbf{h}) = \mathbb{E}\Bigg[&(\mathbf{f}_n(X_{2n+2}) - \mathbf{f}_n(X_{2n+1}))(\mathbf{f}_n(X_{2n+2}) - \mathbf{f}_n(X_{2n+1}))'\\ &\prod_{i=1}^{n} \Delta^{-\ell_i} h_i(X_{2i+1}, X_{2i+2}) \Phi_n^{\boldsymbol{\ell}}(X_1, \dots X_{2n+1}, X_{2n+2}, \dots, X_2)\Bigg] \quad (5.16)\end{aligned}$$

where $\Delta^\ell h_k(x, y) = \Delta^\ell h_k(x) \Delta^\ell h_k(y)$ and an empty product is set to 1.

Remark 5.3.2. *If $R_n^{\boldsymbol{\ell}}(\mathbf{h}) \to 0$ as $n \to \infty$ then, under the conditions of Theorem 5.3.1,*

$$\mathrm{Cov}\,[\mathbf{f}(X)] = \sum_{k=1}^{\infty}(-1)^{k-1}\mathbb{E}\left[\Delta^{-\ell_k}\mathbf{f}_{k-1}(X)\Delta^{-\ell_k}\mathbf{f}'_{k-1}(X)\frac{\Gamma_k^{\boldsymbol{\ell}}\mathbf{h}(X)}{\Delta^{-\ell_k}h_k(X)}\right]. \tag{5.17}$$

In particular when $\mathbf{f}$ is a dth-degree polynomial, then $R_n^{\boldsymbol{\ell}}(\mathbf{h})$ vanishes for $n \geq d$ and (5.14) *is an exact expansion of the variance in* (5.14) *with respect to the $\Gamma_k^{\boldsymbol{\ell}}\mathbf{h}(x)$ functions ($k = 1, \ldots, d$).*

Remark 5.3.3. *A stronger sufficient condition on the functions h_i is that they be strictly increasing throughout $\mathcal{S}(p)$, in which case the condition $\Delta^{-\ell_i}h_i > 0$ is guaranteed. Under this assumption, the matrix $R_n^{\boldsymbol{\ell}}(\mathbf{h})$ defined in* (5.16) *is non-negative definite so that, in particular, taking $h_i = h$ for all $i \geq 1$ and fixing $r = 2$ we recover the expansion* (5.6) *as stated in the Introduction.*

Remark 5.3.4. *When $\ell \neq 0$ then the condition that $\mathbb{P}[\Delta^{-\ell_i}h_i(X) > 0] = 1$ is itself also too restrictive because, as will have been made clear in the proof (see the Appendix), the recurrence only implies that $\Delta^{-\ell_i}h_i(x)$ needs to be positive on some interval $[a + \mathbf{a}_i; b - \mathbf{b}_i] \subset [a, b]$ where $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are positive integers (they will be properly defined in* (5.24)*). In particular when $\ell \neq 0$ the sequence necessarily stops if $\mathcal{S}(p)$ is bounded, since after a certain number of iterations the indicator functions defining $\Phi_{n,j}^{\boldsymbol{\ell}}$ will be 0 everywhere.*

Suppose that the assumption of Remark 5.3.3 applies, so that the remainder is non negative definite. Then, taking $n = 1$ in (5.14) gives an upper bound, and taking $n = 2$ gives a lower bound, on the covariance, and the following holds (stated again in the case $r = 2$, for the sake of clarity).

Corollary 5.3.5. *Let all the conditions in Theorem 5.3.1 prevail for $n = 2$. Then*

$$\begin{aligned}
&\mathbb{E}\left[\Delta^{-\ell_1}f(X)\Delta^{-\ell_1}g(X)\frac{\Gamma_1^{\ell_1}h_1(X)}{\Delta^{-\ell_1}h_1(X)}\right] \\
&\quad - \mathbb{E}\left[\Delta^{-\ell_2}\left(\frac{\Delta^{-\ell_1}f(X)}{\Delta^{-\ell_1}h(X)}\right)\Delta^{-\ell_2}\left(\frac{\Delta^{-\ell_1}g(X)}{\Delta^{-\ell_1}h(X)}\right)\frac{\Gamma_2^{\ell_1,\ell_2}(h_1,h_2)(X)}{\Delta^{-\ell_2}h_2(X)}\right] \\
&\leq \mathrm{Cov}[f(X), g(X)] \leq \mathbb{E}\left[\Delta^{-\ell_1}f(X)\Delta^{-\ell_1}g(X)\frac{\Gamma_1^{\ell_1}h_1(X)}{\Delta^{-\ell_1}h_1(X)}\right].
\end{aligned}$$

Remark 5.3.6. *When $f = g$, the upper bound for $n = 1$ is a weighted Poincaré inequality of the same essence as the upper bound provided in Klaassen (1985) (as revisited in Chapter 4), whereas the lower bound obtained with $n = 2$ is of a different flavour.*

Of course such identities and expansions are only useful if the weights are of a manageable form. This is exactly the topic of the next section.

4 About the weights in Theorem 5.3.1

The crucial quantities in Theorem 5.3.1 are the sequences of weights $\Gamma_k^{\boldsymbol{\ell}}\mathbf{h}$ defined in (5.15). For $k = 1$, the expression are straightforward to obtain (see equations (5.21) for the continuous case $\ell_1 = 0$ and (5.25) for the discrete case $\ell_1 \in \{-1, 1\}$). For larger k the situation is not so straightforward. Relevance of the higher order terms in the covariance expansions (5.14) then hinges on the tractability of these weights, which itself depends on the choice of functions $h_1, h_2, \ldots$. In this section we restrict attention to the (natural) choice $h_k(x) = h(x)$ for all k. Then, writing $\Gamma_k^{\boldsymbol{\ell}}h(x)$ instead of $\Gamma_k^{\boldsymbol{\ell}}(h, h, \ldots)(x)$ we can express the sequence of weights as $\Gamma_k^{\boldsymbol{\ell}}h(x) =: \mathbb{E}\left[\gamma_k^{\boldsymbol{\ell}}h(X_1, x, X_2)\right]$ where, for all $k \geq 1$, we set

$$\gamma_k^{\boldsymbol{\ell}}h(x_1, x, x_2) = \mathbb{E}\Big[(h(X_{2k}) - h(X_{2k-1}))\prod_{i=1}^{k-1}\Delta^{-\ell_i}h(X_{2i+1}, X_{2i+2}) \\ \Phi_p^{\ell_k}(X_{2k-1}, x, X_{2k})\Phi_{k-1}^{\boldsymbol{\ell}}(x_1, X_3 \ldots, X_{2k-1}, X_{2k}, \ldots, x_2)\Big]. \quad (5.18)$$

We now study (5.18) and the resulting expressions for the weights under different sets of assumptions.

4.1 General considerations

When no specific assumptions are made on p or h, we find it easier to separate the continuous case (i.e. $\ell = 0$) from the discrete one (i.e. $\ell \in \{-1, 1\}$).

The continuous case

The continuous case is quite easy as (5.13) simplifies when all the test functions h_i are equal and the expressions follow directly from the structure of the weight sequence, which turn out to be straightforward iterated integrals. We note that such iterated integrals have a structure which may be of independent interest; all details are provided in the Appendix.

Lemma 5.4.1. *Fix $\boldsymbol{\ell} = (0, 0, \ldots)$ and let h be non-decreasing. Then for all $k \geq 1$,*

$$\gamma_k^0 h(x_1, x, x_2) = (h(x) - h(x_1))^{k-1}(h(x_2) - h(x))^{k-1}(h(x_2) - h(x_1))\frac{\mathbb{I}[x_1 \leq x \leq x_2]}{p(x)k!(k-1)!} \quad (5.19)$$

and

$$\Gamma_k^0 h(x) = \frac{1}{k!(k-1)!}\frac{1}{p(x)}\mathbb{E}\Big[\big(h(x)-h(X_1)^{k-1}(h(X_2)-h(x)\big)^{k-1}\big(h(X_2)-h(X_1)\big) \mathbb{I}[X_1 \le x \le X_2]\Big]. \tag{5.20}$$

Specific instantiations for different explicit distributions are given in Section 4.3. We nevertheless note that, letting $\nu(h)$ denote the mean $\mathbb{E}[h(X)]$ we get

$$\begin{aligned}\Gamma_1^0 h(x) &= \frac{1}{p(x)}\mathbb{E}\left[(h(X_2)-h(X_1))\mathbb{I}[X_1 \le x \le X_2]\right] \\ &= \frac{1}{p(x)}\mathbb{E}[(\nu(h)-h(X))\mathbb{I}[x \le X]]\end{aligned} \tag{5.21}$$

which one may recognize as the inverse of the canonical Stein operator (see (5.26)); in particular taking $h(x) = \mathrm{Id}(x) = x$ the identity function, (5.21) yields the Stein kernel. For more information on the connection with Stein operators, see Section 4.1.

The discrete case

In the discrete case, simplifications of $\Gamma_k^{\boldsymbol{\ell}} h(x)$ are more difficult as (5.13) depends strongly on the chosen sequence $\boldsymbol{\ell}$. Let $\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots) \in \{-1,+1\}^\infty$. Recall the notations in (3.2) and set $a_{\ell_i} = a_i$, $b_{\ell_i} = b_i$ for $i \ge 1$. Applying the definitions leads to

$$\gamma_1^{\ell_1} h(x_1, x, x_2) = (h(x_2)-h(x_1))\frac{\mathbb{I}[x_1 + a_1 \le x \le x_2 - b_1]}{p(x)} \tag{5.22}$$

$$\gamma_2^{\ell_1,\ell_2} h(x_1, x, x_2) = \left(\sum_{x_3 = x_1 + a_1}^{x - a_2} \sum_{x_4 = x + b_2}^{x_2 - b_1} (h(x_4) - h(x_3))\Delta^{-\ell_1} h(x_3, x_4)\right) \times \frac{\mathbb{I}[x_1 + a_1 + a_2 \le x \le x_2 - b_1 - b_2]}{p(x)}. \tag{5.23}$$

In order to generalize to arbitrary $k \ge 3$, we introduce

$$\mathbf{a}_k = \sum_{i=1}^k a_i \text{ and } \mathbf{b}_k = \sum_{i=1}^k b_i. \tag{5.24}$$

Note that $\mathbf{a}_k (= \mathbf{a}_k(\boldsymbol{\ell}))$ counts the number of "+" in the first k components of $\boldsymbol{\ell}$ and $\mathbf{b}_k (= \mathbf{b}_k(\boldsymbol{\ell}))$ counts the corresponding number of "−", so that $\mathbf{a}_k + \mathbf{b}_k = k$. Then for $k \ge 2$ we have (sums over empty sets are set to 1):

$$
\begin{aligned}
&\gamma_k^{\boldsymbol{\ell}} h(x_1, x, x_2) \\
&= \Bigg(\sum_{x_3 = x_1 + \mathbf{a}_{k-1}}^{x - a_k} \sum_{x_4 = x + b_k}^{x_2 - \mathbf{b}_{k-1}} (h(x_4) - h(x_3)) \Delta^{-\ell_{k-1}} h(x_3, x_4) \times \\
&\quad \sum_{x_5 = x_1 + \mathbf{a}_{k-2}}^{x_3 - a_{k-1}} \sum_{x_6 = x_4 + b_{k-1}}^{x_2 - \mathbf{b}_{k-2}} \Delta^{-\ell_{k-2}} h(x_5, x_6) \cdots \\
&\quad \sum_{x_{2k-1} = x_1 + a_1}^{x_{2k-3} - a_2} \sum_{x_{2k+1} = x_{2k-2} + b_2}^{x_2 - b_1} \Delta^{-\ell_1} h(x_{2k-1}, x_{2k}) \Bigg) \frac{\mathbb{I}[x_1 + \mathbf{a}_k \leq x \leq x_2 - \mathbf{b}_k]}{p(x)}
\end{aligned}
$$

for all $x \in \mathcal{S}(p)$ and all x_1, x_2. This is a proof of the next result.

Proposition 5.4.2. *Instate all previous notations. For all $k \geq 1$,*

$$
\begin{aligned}
\gamma_k^{\boldsymbol{\ell}} h(x_1, x, x_2) = &\Bigg(\sum_{x_3 = x_1 + \mathbf{a}_{k-1}}^{x - a_k} \sum_{x_4 = x + b_k}^{x_2 - \mathbf{b}_{k-1}} (h(x_4) - h(x_3)) \psi_{k-1}^{\boldsymbol{\ell}} h(x_1, x_3, x_4, x_2) \Bigg) \times \\
&\frac{\mathbb{I}[x_1 + \mathbf{a}_k \leq x \leq x_2 - \mathbf{b}_k]}{p(x)}
\end{aligned}
$$

where $\psi_0^{\boldsymbol{\ell}} h(x_1, x_3, x_4, x_2) = 1$ and, for $k \geq 2$,

$$
\psi_{k-1}^{\boldsymbol{\ell}} h(x_1, x_3, x_4, x_2) = \psi_{k-1,1}^{\boldsymbol{\ell}} h(x_1, x_3) \psi_{k-1,2}^{\boldsymbol{\ell}} h(x_4, x_2)
$$

with

$$
\begin{aligned}
&\psi_{k-1,1}^{\boldsymbol{\ell}} h(x_1, x_3) \\
&= \Delta^{-\ell_{k-1}} h(x_3) \sum_{x_5 = x_1 + \mathbf{a}_{k-2}}^{x_3 - a_{k-1}} \Bigg(\Delta^{-\ell_{k-2}} h(x_5) \sum_{x_7 = x_1 + \mathbf{a}_{k-4}}^{x_5 - a_{k-2}} \Bigg(\cdots \sum_{x_{2k-1} = x_1 + a_1}^{x_{2k-3} - a_2} \Delta^{-\ell_1} h(x_{2k-1}) \Bigg) \Bigg) \\
&\psi_{k-1,2}^{\boldsymbol{\ell}} h(x_4, x_2) \\
&= \Delta^{-\ell_{k-1}} h(x_4) \sum_{x_6 = x_4 + b_{k-1}}^{x_2 - \mathbf{b}_{k-2}} \Bigg(\Delta^{-\ell_{k-2}} h(x_6) \sum_{x_8 = x_6 + b_{k-2}}^{x_2 - \mathbf{b}_{k-3}} \Bigg(\cdots \sum_{x_{2k} = x_{2k-2} + b_2}^{x_2 - b_1} \Delta^{-\ell_1} h(x_{2k}) \Bigg) \Bigg)
\end{aligned}
$$

for all $x_1 + \mathbf{a}_{k-1} \leq x_3 \leq x_4 \leq x_2 - \mathbf{b}_{k-1}$.

Taking expectations in (5.22) and (5.23) we obtain

$$
\Gamma_1^{\ell_1} h(x) = \frac{1}{p(x)} \mathbb{E}\left[(h(X_2) - h(X_1)) \mathbb{I}[X_1 + a_1 \leq x \leq X_2 - b_1] \right] \tag{5.25}
$$

$$\Gamma_2^{\ell_1,\ell_2} h(x) = \frac{1}{p(x)} \mathbb{E}\Bigg[\sum_{x_3 = X_1 + a_1}^{x - a_2} \sum_{x_4 = x + b_2}^{X_2 - b_1} (h(x_4) - h(x_3)) \Delta^{-\ell_1} h(x_3, x_4) \\ \mathbb{I}[X_1 + \mathbf{a}_2 \leq x \leq X_2 - \mathbf{b}_2] \Bigg].$$

The expressions for higher orders are easy to infer, but this seems to be the best we can do because the expressions in Proposition 5.4.2 are obscure and, unfortunately, we have not been able to devise a formula as transparent as (5.19) for general h in the discrete case. Nevertheless, simple manageable expressions are obtainable for certain specific choices of h, particularly the case $h(x) = \mathrm{Id}(x)$ as we shall see in Section 4.2.

Connection with Stein operators

In Chapter 3, we introduced the *canonical inverse Stein operator*

$$\mathcal{L}_p^\ell h(x) = \mathbb{E}\Big[(h(X_1) - h(X_2)) \Phi_p^\ell(X_1, x, X_2) \Big] \tag{5.26}$$

for $h \in L^1(p)$ and X_1, X_2 independent copies of $X \sim p$. This operator has the property of yielding solutions to so-called Stein equations, both in discrete and continuous setting; it has many important properties within the context of Stein's method. In particular it provides generalized covariance identities and, when $h(x) = \mathrm{Id}(x)$ is the identity function, it provides

$$\tau_p^\ell(x) = -\mathcal{L}_p^\ell \mathrm{Id}(x) \tag{5.27}$$

the all-important Stein kernel of p. This function, first introduced in Stein (1986), has long been known to provide a crucial handle on the properties of p and is now studied as an object of intrinsic interest (see e.g. Courtade et al., 2019, Fathi, 2019).

From (5.21) and (5.25), we immediately recognize that $\Gamma_1^{\ell_1} h(x) = -\mathcal{L}_p^{\ell_1} h(x)$, in other words the first order weight in our expansion is given by a Stein operator. There is also a connection between $\Gamma_k^{\boldsymbol{\ell}} h$ and "higher order" Stein kernels. To see this, restrict to the continuous case $\boldsymbol{\ell} = 0$ and introduce $H_x^k(y) = (h(y) - h(x))^k / k!$. Then (5.19) becomes

$$\Gamma_k^{\mathbf{0}} h(x) = (-1)^k \left(\mathbb{E}\big[H_x^{k-1}(X)\big] \mathcal{L}_p^0 H_x^k(x) - \mathbb{E}\big[H_x^k(X)\big] \mathcal{L}_p^0 H_x^{k-1}(x) \right) \tag{5.28}$$

(see the Appendix for a proof). In the case $h(x) = x$ the expression (5.28) simplifies to Papathanasiou's weights from (5.4). This allows to make the connection between considerations related to Stein's method and the weights appearing in the expansions, as has already been observed (see e.g. Afendras et al., 2007). We do not pursue this line of research here, except to point out that our result provides a framework to

the important works (Papathanasiou, 1988, Korwar, 1991, Johnson, 1993, Afendras et al., 2007, 2018), which focus on particular families of distributions, see Section 4.3. Further study of this connection, in line e.g. with Fathi (2018), is outside the scope of this thesis and deferred to a future work.

4.2 Handpicking the test functions

We now focus on particular choices of h. To begin with, we consider the most intuitive choice (and the only one studied in the literature): $h(x) = \mathrm{Id}(x)$. In this case we abbreviate $\Gamma_k^{\ell}\mathbf{h}(x) = \Gamma_k^{\ell}(x)$. If $\boldsymbol{\ell} = \mathbf{0}$ we have

$$\Gamma_k^{\mathbf{0}}(x) = \frac{1}{k!(k-1)!p(x)}\mathbb{E}\left[(X_2 - x)^{k-1}(x - X_1)^{k-1}(X_2 - X_1)\mathbb{I}[X_1 \le x \le X_2]\right].$$

The discrete case is less transparent, but direct computations for the first two weights in the discrete case lead to

$$\Gamma_1^{\ell_1}(x) = \mathbb{E}\left[(X_2 - X_1)\frac{\mathbb{I}[X_1 + a_1 \le x \le X_2 - b_1]}{p(x)}\right]$$

$$\Gamma_2^{\ell_1,\ell_2}(x) = \mathbb{E}\left[(X_2 - x - \mathbf{b}_2 + 1)(x - X_1 - \mathbf{a}_2 + 1)(X_2 - X_1)\frac{\mathbb{I}[X_1 + \mathbf{a}_2 \le x \le X_2 - \mathbf{b}_2]}{2p(x)}\right].$$

More generally we have the following.

Lemma 5.4.3. *If $\boldsymbol{\ell} \in \{-1, 1\}^{\infty}$ then for all $k \ge 1$*

$$\begin{aligned}\Gamma_k^{\boldsymbol{\ell}}(x) =& \mathbb{E}\Bigg[(X_2 - x - \mathbf{b}_k + 1)^{[k-1]}(x - X_1 - \mathbf{a}_k + 1)^{[k-1]}(X_2 - X_1) \\ & \frac{\mathbb{I}[X_1 + \mathbf{a}_k \le x \le X_2 - \mathbf{b}_k]}{p(x)k!(k-1)!}\Bigg].\end{aligned} \tag{5.29}$$

We can unify the continuous and the discrete settings, to reap

$$\Gamma_k^{\boldsymbol{\ell}}(x) = \mathbb{E}\left[(X_2 - x)_{\{k-1;\boldsymbol{\ell}\}}(x - X_1)^{\{k-1;\boldsymbol{\ell}\}}(X_2 - X_1)\frac{\mathbb{I}[X_1 + \mathbf{a}_k \le x \le X_2 - \mathbf{b}_k]}{p(x)k!(k-1)!}\right]$$

where $f_{\{k,\boldsymbol{\ell}\}}(x) = \prod_{j=1}^{k} f(x + \mathbf{a}_k - |\ell|j)$ and $f^{\{k,\boldsymbol{\ell}\}}(x) = \prod_{j=1}^{k} f(x - \mathbf{a}_k + |\ell|j)$ or equivalently

$$f_{\{k,\boldsymbol{\ell}\}}(x) = \begin{cases} f(x)^k & \text{if } \boldsymbol{\ell} = \mathbf{0}, \\ \prod_{j=1}^{k} f(x + \mathbf{a}_k - j) = f_{[k]}(x + \mathbf{a}_k - 1) & \text{else;} \end{cases}$$

$$f^{\{k,\boldsymbol{\ell}\}}(x) = \begin{cases} f(x)^k & \text{if } \boldsymbol{\ell} = \mathbf{0}, \\ \prod_{j=1}^{k} f(x - \mathbf{a}_k + j) = f^{[k]}(x - \mathbf{a}_k + 1) & \text{else.} \end{cases}$$

and the empty product equals 1.

Remark 5.4.4. *As already noted in Section 4.1, the expression of the weights in the continuous case is already known and can be traced back to works as early as Papathanasiou (1988); the expression for the discrete case (namely equation* (5.29)*) is new, although a version with* $\boldsymbol{\ell} = (-1, -1, -1, \ldots)$ *is available from Afendras et al. (2007).*

Another natural choice in the continuous case $\boldsymbol{\ell} = 0$, of increasing function h to plug into the weights is $h(x) = P(x)$ with P the cdf of p. Then the following holds.

Lemma 5.4.5. *If* $\boldsymbol{\ell} = 0$ *and* $X \sim p$ *has cdf* P *then*

$$\Gamma_k^0 P(x) = \frac{1}{k!(k+1)!p(x)} P(x)^k (1 - P(x))^k.$$

A final natural choice occurs whenever p is log-concave. Indeed in this case the function $h_1 = -(\log p)'$ is increasing. In particular, $\Gamma_1^0 h_1(x) = -\mathcal{L}_p^0 h_1(x) = 1$, which allows us to rewrite the first order expansion as

$$\mathrm{Cov}\left[f(X), g(X)\right] = \mathbb{E}\left[\frac{f'(X)g'(X)}{-(\log p)''(X)}\right] - R_1^0(\mathbf{h}).$$

This expression generalizes the Brascamp-Lieb inequality from Chapter 4. For simple expressions of $R_1^0(\mathbf{h})$ one may like to choose $h_2 = h_3 = \cdots = \mathrm{Id}$. This example thus benefits from the flexibility in choosing a sequence of functions $\mathbf{h}$.

4.3 Illustrations

The weights for Integrated Pearson family

Based on the definition 4.2.8, the following results hold (to facilitate comparison of the results we use the same notations as in Afendras and Papadatos, 2014).

Proposition 5.4.6. *If* $X \sim p$ *is integrated Pearson distributed with Stein kernel* $\tau_p(x) = \tau_p^{\mathbf{0}}(x) = -\mathcal{L}_p^0(\mathrm{Id}) = \delta x^2 + \beta x + \gamma$ *then*

$$\Gamma_k^0(x) = \frac{\tau_p(x)^k}{k! \prod_{j=0}^{k-1}(1 - j\delta)}. \tag{5.30}$$

The coefficient (δ, β, γ) of the Stein kernel are explicitly given in Table 4.3. These coefficients allow us to directly obtain the infinite expansion of covariance for the integrated Pearson family. We give the expansions for two distributions in the following examples.

Example 5.4.7 (Normal expansion)**.** *The standard normal distribution ϕ is an element of the integrated Pearson family with $\delta = 0, \beta = 0$, and $\gamma = 1$. Direct computations show that if $X \sim \mathcal{N}(0,1)$ then $\tau_\phi(x) = 1$ so that $\Gamma_k^0(x) = \frac{1}{k!}$ for all k and*

$$\mathrm{Cov}[f(X), g(X)] = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k!} \mathbb{E}\left[f^{(k)}(X) g^{(k)}(X)\right],$$

which extends the variance expansion (5.1) *to a covariance expansion.*

Example 5.4.8 (Beta expansion)**.** *The Beta(a,b) distribution is an element of the integrated Pearson family with $\delta = -\frac{1}{a+b}, \beta = \frac{1}{a+b}$, and $\gamma = 0$; then $\tau_{\mathrm{Beta(a,b)}}(x) = \frac{x(1-x)}{a+b}$. Direct computations show that if $X \sim \mathrm{Beta}(a,b)$ then $\Gamma_k^0(x) = (x(1-x))^k/(k!(a+b)^{[k]})$ for $k \geq 1$, so that*

$$\mathrm{Cov}[f(X), g(X)] = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k!(a+b)^{[k]}} \mathbb{E}\left[f^{(k)}(X) g^{(k)}(X) X^k (1-X)^k\right].$$

The weights for Cumulative Ord family

Based on the definition 4.2.8, the following results hold (to facilitate comparison of the results we use the exact same notations as in Afendras et al., 2007).

Proposition 5.4.9. *If $X \sim p$ is cumulative Ord distributed with $\tau_p^-(x) = \delta x^2 + \beta x + \gamma$ (and hence $\tau_p^+(x) = \frac{p(x-1)}{p(x)} \tau_p^-(x-1) = x(\delta x + \beta + 1)$), then*

$$\Gamma_k^{\boldsymbol{\ell}}(x) = \frac{1}{k! \prod_{j=0}^{k-1}(1 - j\delta)} \left(\tau_p^+(x)\right)_{[\mathbf{a}_k]} \left(\tau_p^-(x)\right)^{[\mathbf{b}_k]}. \tag{5.31}$$

Remark 5.4.10. *By taking only k forward difference, i.e., $\boldsymbol{\ell} = (-1, \ldots, -1)$, we deduce the result of Afendras et al. (2007, Theorem 4.1). In particular, their Table 1 illustrates the expression of $\Gamma_k^{\boldsymbol{\ell}}(x)$ for some discrete distributions from the cumulative Ord family. Tables 4.1 and 4.2 give explicit expressions of Stein kernels for many standard distributions.*

In the discrete case, there is much more flexibility in the construction of the bounds as any permutation of $+1$ and -1 is allowed for every k, leading to:

$$\mathrm{Var}[g(X)] = \mathbb{E}\left[\Gamma_1^+(X)(\Delta^- g(X))^2\right] - R_1^+ = \mathbb{E}\left[\Gamma_1^-(X)(\Delta^+ g(X))^2\right] - R_1^-$$

and for an order 2 expansion, for any of the four choices of $(\ell_1, \ell_2) \in \{-1, +1\}^2$,

$$\mathrm{Var}[g(X)] = \mathbb{E}\left[\Gamma_1^{\ell_1}(X)(\Delta^{-\ell_1} g(X))^2\right] - \mathbb{E}\left[\Gamma_2^{\ell_1,\ell_2}(X)(\Delta^{-\ell_1,-\ell_2} g(X))^2\right] + R_2^{\ell_1,\ell_2}$$

where we use the concise notation $\Delta^{\ell_1,\ell_2} g(X)$ for $\Delta^{\ell_2}\left(\Delta^{\ell_1} g(X)\right)$.

Example 5.4.11 (Binomial expansion)**.** *The Binomial(n,θ) distribution is an element of the cumulated Ord family with $\delta = 0, \beta = -\theta$, and $\gamma = n\theta$; its Stein kernels are $\tau^-(x) = \theta(n-x)$ and $\tau^+(x) = (1-\theta)x$. Hence $\Gamma_1^+(x) = (1-\theta)x$, $\Gamma_1^-(x) = \theta(n-x)$ so that the order 1 expansions are*

$$\begin{aligned}\mathrm{Var}[g(X)] &= (1-\theta)\mathbb{E}\left[X(\Delta^- g(X))^2\right] - R_1^+ &(5.32)\\ &= \theta\mathbb{E}\left[(n-X)(\Delta^+ g(X))^2\right] - R_1^-; &(5.33)\end{aligned}$$

choosing a linear combination of (5.32) *and* (5.33) *with weights θ and $1-\theta$, respectively, yields*

$$\mathrm{Var}[g(X)] = n\theta(1-\theta)\mathbb{E}\left[\frac{X}{n}(\Delta^- g(X))^2 + \frac{n-X}{n}(\Delta^+ g(X))^2\right] - \theta R_1^+ - (1-\theta)R_1^-. \tag{5.34}$$

We note that Hillion et al. (2014, Theorem 1.3) introduce the "natural binomial derivative" $\nabla_n g(x) = \frac{x}{n}\Delta^- g(x) + \frac{n-x}{n}\Delta^+ g(x)$ and prove – by arguments which are specific to the binomial distribution – the Poincaré inequality

$$\mathrm{Var}[g(X)] \le n\theta(1-\theta)\mathbb{E}\left[\left(\nabla_n g(X)\right)^2\right].$$

The connection with (5.34) *is easy to see (see e.g. Hillion et al., 2014, Remark 3.3)*

$$\left(\nabla_n g(x)\right)^2 = \frac{x}{n}(\Delta^- g(x))^2 + \frac{n-x}{n}(\Delta^+ g(x))^2 - \frac{x(n-x)}{n^2}(\Delta^{+-} g(x))^2.$$

Moving to the second order, direct computations show that

$$\begin{aligned}\Gamma_2^{+,+}(x) &= \frac{1}{2}(1-\theta)^2 x(x-1)\mathbb{I}[1 \le x \le n],\\ \Gamma_2^{+,-}(x) &= \Gamma_2^{-,+}(x) = \frac{1}{2}\theta(1-\theta)x(n-x)\mathbb{I}[0 \le x \le n],\\ \Gamma_2^{-,-}(x) &= \frac{1}{2}\theta^2(n-x)(n-x-1)\mathbb{I}[0 \le x \le n-1]\end{aligned}$$

leading to the order 2 expansions

$$\begin{aligned}&\mathrm{Var}[g(X)]\\ &= (1-\theta)\mathbb{E}\left[X(\Delta^- g(X))^2\right] - \frac{1}{2}(1-\theta)^2\mathbb{E}\left[X(X-1)(\Delta^{--} g(X))^2\right] + R_2^{++}\\ &= (1-\theta)\mathbb{E}\left[X(\Delta^- g(X))^2\right] - \frac{1}{2}\theta(1-\theta)\mathbb{E}\left[X(n-X)(\Delta^{-+} g(X))^2\right] + R_2^{+-}\\ &= \theta\mathbb{E}\left[(n-X)(\Delta^+ g(X))^2\right] - \frac{1}{2}\theta(1-\theta)\mathbb{E}\left[X(n-X)(\Delta^{+-} g(X))^2\right] + R_2^{-+}\\ &= \theta\mathbb{E}\left[(n-X)(\Delta^+ g(X))^2\right] - \frac{1}{2}\theta^2\mathbb{E}\left[(n-X-1)(n-X)(\Delta^{++} g(X))^2\right] + R_2^{--}.\end{aligned}$$

Using the notation ∇_n from above, we deduce from a combination of the second and third identities the lower variance bound

$$\mathrm{Var}[g(X)] \geq n\theta(1-\theta)\left\{\mathbb{E}\left[\left(\nabla_n g(X)\right)^2\right] - \frac{n-2}{2}\mathbb{E}\left[\frac{X(n-X)}{n^2}(\Delta^{+-}g(X))^2\right]\right\}.$$

Combining these inequalities yields that for $0 < \theta < 1$,

$$\mathbb{E}\left[\left(\nabla_n g(X)\right)^2\right] - \frac{n-2}{2}\mathbb{E}\left[\frac{X(n-X)}{n^2}(\Delta^{+-}g(X))^2\right] \leq \frac{\mathrm{Var}[g(X)]}{n\theta(1-\theta)} \leq \mathbb{E}\left[\left(\nabla_n g(X)\right)^2\right].$$

Examples which are not integrated Pearson or cumulative Ord distributions

Example 5.4.12 (Laplace expansion)**.** *Direct computations show that if $p(x) = e^{-|x|}/2$ on $\mathbb{R}$, i.e. $X \sim \mathrm{Laplace}(0,1)$, then $\Gamma_1^0(x) = 1+|x|$ and $\Gamma_2^0(x) = \frac{1}{2}x^2+|x|+1$ so that the first two bounds become*

$$\begin{aligned}\mathrm{Var}[g(X)] &= \mathbb{E}\left[(1+|X|)g'(X)^2\right] - R_1 \\ &= \mathbb{E}\left[(1+|X|)g'(X)^2\right] - \mathbb{E}\left[(1+|X|+X^2/2)g''(X)^2\right] + R_2.\end{aligned}$$

Despite this distribution not being a member of the Pearson family, the general expression for Γ_k is quite simple:

$$\Gamma_k^0(x) = \sum_{j=0}^{k}\frac{|x|^j}{j!}.$$

The structure of this sequence seems to indicate that this distribution is of a different nature than integrated Pearson distributions; this is also illustrated in the properties of the corresponding Stein operator (which is best described as a second order differential operator), see Eichelsbacher and Thäle (2015), Pike and Ren (2014).

Example 5.4.13 (Rayleigh expansion)**.** *Direct computations show that if $X \sim \mathrm{Rayleigh}(0,1)$ (i.e. $p(x) = xe^{-x^2/2}$ on $\mathbb{R}^+$) then $\tau_p^0(x)$ does not take on an agreeable form. Nevertheless the choice $h(x) = x^2$ leads to*

$$\frac{\Gamma_k^0 h(x)}{h'(x)} = \frac{2^{k-2}}{k!}x^{2(k-1)}.$$

Example 5.4.14 (Cauchy expansion)**.** *The standard Cauchy distribution lacks moments; nevertheless taking $h(x) = \arctan(x)$ leads to*

$$\frac{\Gamma_k^0 h(x)}{h'(x)} = \frac{1}{4^k(k+1)!(k)!}(1+x^2)^2\left(\pi^2 - 4\arctan(x)^2\right)^k.$$

Example 5.4.15 (Levy expansion)**.** *The pdf of the standard Levy distribution is given by* $(2\pi)^{-\frac{1}{2}} e^{\frac{1}{2x}} x^{-\frac{3}{2}}$*. Similarly as in the previous example, taking* $h(x) = P(x)$*,*

$$\frac{\Gamma_k^0 h(x)}{h'(x)} = \binom{k+1}{2} \frac{1}{k!(k+1)!} \pi e^{1/x} x^3 \big((1-P(x))P(x)\big)^k.$$

A Appendix: proofs

Proof of Lemma 5.2.1. The equivalence between (5.8) and (5.7) follows from the fact that $\mathbb{I}[X_1 < X_2] + \mathbb{I}[X_1 = X_2] + \mathbb{I}[X_1 > X_2] = 1$ and

$$\begin{aligned}&\mathbb{E}\left[\big(f(X_2) - f(X_1)\big)\big(g(X_2) - g(X_1)\big)\mathbb{I}[X_1 < X_2]\right] \\ &\quad = \mathbb{E}\left[\big(f(X_2) - f(X_1)\big)\big(g(X_2) - g(X_1)\big)\mathbb{I}[X_2 < X_1]\right].\end{aligned}$$

Without loss of generality in (5.8) it can be assumed that $\mathbb{E}[f(X)] = \mathbb{E}[g(X)] = 0$. Evaluating the expectation (5.8) through expanding the product yields the assertion. □

Proof of Lemma 5.2.2. First, from (3.16) in Lemma 3.3.1 it follows directly that

$$\Phi_p^\ell(u, x_1, x_2, v)\mathbb{I}[x_1 \neq x_2] = \mathbb{I}[x_1 \neq x_2]\chi^{\ell^2}(x_1, x_2)\Phi_p^\ell(u, x_1, v)\Phi_p^\ell(u, x_2, v). \tag{5.35}$$

With the abbreviations as introduced in the statement of the lemma, the (i, j) entry of the $r \times r$ matrix $R(u, v)$ is

$$\begin{aligned}&(R(u,v))_{i,j} \\ &:= \mathbb{E}\left[(v_{i3}g_4 - v_{i4}g_3)(v_{j3}g_4 - v_{j4}g_3)\Phi_p^\ell(u, X_3, X_4, v)\right] \\ &= \mathbb{E}\left[\mathbb{I}[X_3 \neq X_4](v_{i3}g_4 - v_{i4}g_3)(v_{j3}g_4 - v_{j4}g_3)\chi^{\ell^2}(X_3, X_4)\Phi_p^\ell(u, X_3, v)\Phi_p^\ell(u, X_4, v)\right],\end{aligned}$$

where we used (5.35) in the last step. Next, again using Lemma 3.3.1, $\mathbb{I}[x_1 \neq x_2](\chi^{\ell^2}(x_1, x_2) + \chi^{\ell^2}(x_2, x_1)) = \mathbb{I}[x_1 \neq x_2]$ and by symmetry,

$$\begin{aligned}&\mathbb{E}\left[\mathbb{I}[X_3 \neq X_4](v_{i3}g_4 - v_{i4}g_3)(v_{j3}g_4 - v_{j4}g_3)\chi^{\ell^2}(X_3, X_4)\Phi_p^\ell(u, X_3, v)\Phi_p^\ell(u, X_4, v)\right] \\ &= \mathbb{E}\left[\mathbb{I}[X_4 \neq X_3](v_{i3}g_4 - v_{i4}g_3)(v_{j3}g_4 - v_{j4}g_3)\chi^{\ell^2}(X_4, X_3)\Phi_p^\ell(u, X_3, v)\Phi_p^\ell(u, X_4, v)\right].\end{aligned}$$

Thus

$$
\begin{aligned}
&2(R(u,v))_{i,j}\\
&= \mathbb{E}\left[\mathbb{I}[X_3 \neq X_4](v_{i3}g_4 - v_{i4}g_3)(v_{j3}g_4 - v_{j4}f_3)\chi^{\ell^2}(X_3,X_4)\Phi_p^\ell(u,X_3,v)\Phi_p^\ell(u,X_4,v)\right]\\
&+ \mathbb{E}\left[\mathbb{I}[X_4 \neq X_3](v_{i3}g_4 - v_{i4}g_3)(v_{j3}g_4 - v_{j4}g_3)\chi^{\ell^2}(X_4,X_3)\Phi_p^\ell(u,X_3,v)\Phi_p^\ell(u,X_4,v)\right]\\
&= \mathbb{E}\left[\mathbb{I}[X_3 \neq X_4](v_{i3}g_4 - v_{i4}g_3)(v_{j3}g_4 - v_{j4}g_3)\Phi_p^\ell(u,X_3,v)\Phi_p^\ell(u,X_4,v)\right]\\
&= \mathbb{E}\left[(v_{i3}g_4 - v_{i4}g_3)(v_{j3}g_4 - v_{j4}g_3)\Phi_p^\ell(u,X_3,v)\Phi_p^\ell(u,X_4,v)\right].
\end{aligned}
$$

Now we exploit the independence of X_3 and X_4 to obtain

$$
\begin{aligned}
2(R(u,v))_{i,j} &= 2\mathbb{E}\left[v_{i3}v_{j3}\Phi_p^\ell(u,X_3,v)\right]\mathbb{E}\left[g_4^2\Phi_p^\ell(u,X_4,v)\right]\\
&\quad -2\mathbb{E}\left[v_{i3}g_3\Phi_p^\ell(u,X_3,v)\right]\mathbb{E}\left[v_{j4}g_4\Phi_p^\ell(u,X_4,v)\right].
\end{aligned}
$$

The assertion follows by dividing by 2 and re-arranging the equation. □

Proof of Theorem 5.3.1. First by direct verification we note that the following recursion for $\Phi_n^{\boldsymbol{\ell}}$ holds. Starting from $\Phi_1^{\boldsymbol{\ell}}(x_1,x_3,x_4,x_2) = \Phi_p^{\ell_1}(x_1,x_3,x_4,x_2)$ we have for $n \geq 2$

$$
\begin{aligned}
&\Phi_n^{\boldsymbol{\ell}}(x_1,x_3,\ldots,x_{2n-1},x_{2n+1},x_{2n+2},x_{2n},\ldots,x_2)\\
&\quad = \Phi_p^{\ell_n}(x_{2n-1},x_{2n+1},x_{2n+2},x_{2n})\Phi_{n-1}^{\boldsymbol{\ell}}(x_1,x_3,\ldots,x_{2n-1},x_{2n},\ldots,x_2) \qquad (5.36)
\end{aligned}
$$

for any sequence $(x_j)_{j\geq 1}$. We abbreviate

$$
\begin{aligned}
&\Phi_{n,1}^{\boldsymbol{\ell}}(x_1,\ldots,x_{2n-1},x,x_{2n},\ldots,x_2)\\
&\quad = \Phi_p^{\ell_n}(x_{2n-1},x,x_{2n})\Phi_{n-1}^{\boldsymbol{\ell}}(x_1,x_3,\ldots,x_{2n-1},x_{2n},\ldots,x_2). \qquad (5.37)
\end{aligned}
$$

The proof uses induction in n. First consider $n = 1$. Let X_1, X_2, X_3, X_4 be independent copies of X. Starting from (5.7),

$$
\begin{aligned}
\operatorname{Cov}[\mathbf{f}(X)] &= \mathbb{E}[(\mathbf{f}(X_2) - \mathbf{f}(X_1))(\mathbf{f}(X_2) - \mathbf{f}(X_1))'\mathbb{I}[X_1 < X_2]]\\
&= \mathbb{E}\Big[\mathbb{E}\left[\Phi_p^{\ell_1}(X_1,X_3,X_2)\Delta^{-\ell_1}\mathbf{f}(X_3)\,|\,X_1,X_2\right] \times\\
&\qquad \mathbb{E}\left[\Phi_p^{\ell_1}(X_1,X_4,X_2)\Delta^{-\ell_1}\mathbf{f}(X_4)\,|\,X_1,X_2\right]'\mathbb{I}[X_1 < X_2]\Big]
\end{aligned}
$$

where we used (3.20) in the last step. Now for any h_1 such that $\mathbb{P}[\Delta^{-\ell_1}h_1(X) > 0] = 1$, dividing and multiplying by $\sqrt{\Delta^{-\ell_1}h_1(X)}$ and applying Lemma 5.2.2 (Lagrange identity) with

$$
\mathbf{v}(x) = \frac{\Delta^{-\ell_1}\mathbf{f}(x)}{\sqrt{\Delta^{-\ell_1}h_1(x)}} \quad \text{and } g(x) = \sqrt{\Delta^{-\ell_1}h_1(x)} \qquad (5.38)
$$

gives note re-arrangement

$$\begin{aligned}
&\mathrm{Cov}\left[\mathbf{f}(X)\right] + \mathbb{E}\left[R^{\ell_1}(X_1, X_2; \mathbf{v}, g)\mathbb{I}[X_1 < X_2]\right] \\
&= \mathbb{E}\Bigg[\mathbb{E}\left[\mathbf{v}(X)\mathbf{v}'(X)\Phi_p^{\ell_1}(X_1, X, X_2) \,|\, X_1, X_2\right] \times \\
&\qquad \mathbb{E}\left[g^2(X)\Phi_p^{\ell_1}(X_1, X, X_2) \,|\, X_1, X_2\right] \mathbb{I}[X_1 < X_2]\Bigg] \\
&= \mathbb{E}\Bigg[\mathbb{E}\left[\frac{\Delta^{-\ell_1}\mathbf{f}(X)\Delta^{-\ell_1}\mathbf{f}'(X)}{\Delta^{-\ell_1}h_1(X)}\Phi_p^{\ell_1}(X_1, X, X_2)|X_1, X_2\right] \times \\
&\qquad \mathbb{E}\left[\Delta^{-\ell_1}h_1(X)\Phi_p^{\ell_1}(X_1, X, X_2)|X_1, X_2\right] \mathbb{I}[X_1 < X_2]\Bigg] \\
&= \mathbb{E}\Bigg[\mathbb{E}\left[\frac{\Delta^{-\ell_1}\mathbf{f}(X)\Delta^{-\ell_1}\mathbf{f}'(X)}{\Delta^{-\ell_1}h_1(X)}\Phi_p^{\ell_1}(X_1, X, X_2)|X_1, X_2\right] \times \\
&\qquad (h_1(X_2) - h(X_1))\mathbb{I}[X_1 < X_2]\Bigg] \qquad (5.39)
\end{aligned}$$

with the last equality following from (3.20). Note that, in the discrete case, the strict inequality in the indicator $\mathbb{I}[X_1 < X_2]$ is implicit in

$$\Phi_p^{\ell_1}(X_1, X, X_2) = \frac{\chi^{\ell_1}(X_1, X)\chi^{-\ell_1}(X, X_2)}{p(X)}$$

(and hence a fortiori also in $\Phi_p^{\ell_1}(X_1, X_3, X_4, X_2)$; in the continuous case there is no difference between $\mathbb{I}[X_1 < X_2]$ and $\mathbb{I}[X_1 \leq X_2]$. Hence unconditioning yields

$$\begin{aligned}
&\mathbb{E}\left[\frac{\Delta^{-\ell_1}\mathbf{f}(X)\Delta^{-\ell_1}\mathbf{f}'(X)}{\Delta^{-\ell_1}h_1(X)}\Phi_p^{\ell_1}(X_1, X, X_2)(h_1(X_2) - h_1(X_1))\mathbb{I}[X_1 < X_2]\right] \\
&= \mathbb{E}\left[\frac{\Delta^{-\ell_1}\mathbf{f}(X)\Delta^{-\ell_1}\mathbf{f}'(X)}{\Delta^{-\ell_1}h_1(X)}\Phi_p^{\ell_1}(X_1, X, X_2)(h_1(X_2) - h_1(X_1))\right] \\
&= \mathbb{E}\left[\Delta^{-\ell_1}\mathbf{f}(X)\Delta^{-\ell_1}\mathbf{f}'(X)\frac{\Gamma_1^{\ell_1}h_1(X)}{\Delta^{-\ell_1}h_1(X)}\right],
\end{aligned}$$

giving the first term in the covariance expansion (5.14). With the notation (5.38), the remainder term in (5.39) is

$$\begin{aligned}
&\mathbb{E}\left[R^{\ell_1}(X_1, X_2; \mathbf{v}, g)\mathbb{I}[X_1 < X_2]\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[(\mathbf{v}_3 g_4 - \mathbf{v}_4 g_3)(\mathbf{v}_3 g_4 - \mathbf{v}_4 g_3)'\Phi_p^{\ell_1}(X_1, X_3, X_4, X_2)|X_1, X_2\right]\mathbb{I}[X_1 < X_2]\right].
\end{aligned}$$

Now,

$$\mathbf{v}_3 g_4 = \frac{\Delta^{-\ell_1}\mathbf{f}(X_3)}{\sqrt{\Delta^{-\ell_1}h_1(X_3)}}\sqrt{\Delta^{-\ell_1}h_1(X_4)} = \frac{\Delta^{-\ell_1}\mathbf{f}(X_3)}{\Delta^{-\ell_1}h_1(X_3)}\sqrt{\Delta^{-\ell_1}h_1(X_3)\Delta^{-\ell_1}h_1(X_4)}$$

and $\sqrt{\Delta^{-\ell_1}h_1(X_3)\Delta^{-\ell_1}h_1(X_4)}$ is a common factor, so that

$$\begin{aligned}
&\mathbb{E}\left[R^{\ell_1}(X_1,X_2;\mathbf{v},g)\mathbb{I}[X_1<X_2]\right]\\
&= \mathbb{E}\left[\left(\frac{\Delta^{-\ell_1}\mathbf{f}(X_3)}{\Delta^{-\ell_1}h_1(X_3)}-\frac{\Delta^{-\ell_1}\mathbf{f}(X_4)}{\Delta^{-\ell_1}h_1(X_4)}\right)\left(\frac{\Delta^{-\ell_1}\mathbf{f}(X_3)}{\Delta^{-\ell_1}h_1(X_3)}-\frac{\Delta^{-\ell_1}\mathbf{f}(X_4)}{\Delta^{-\ell_1}h_1(X_4)}\right)'\right.\\
&\qquad\left.\times\left(\sqrt{\Delta^{-\ell_1}h_1(X_3)\Delta^{-\ell_1}h_1(X_4)}\right)^2\Phi_p^{\ell_1}(X_1,X_3,X_4,X_2)\mathbb{I}[X_1<X_2]\right]\\
&= \mathbb{E}\big[(\mathbf{f}_1(X_3)-\mathbf{f}_1(X_4))(\mathbf{f}_1(X_3)-\mathbf{f}_1(X_4))'\Delta^{-\ell_1}h_1(X_3)\Delta^{-\ell_1}h_1(X_4)\\
&\qquad\Phi_p^{\ell_1}(X_1,X_3,X_4,X_2)\big]\\
&= R_1^{\ell_1}(\mathbf{h})
\end{aligned}$$

as required; here $\mathbf{h}=h_1$. Thus the assertion holds for $n=1$.

To obtain the complete claim, we proceed by induction and suppose that the claim holds at some n. It remains to show that

$$R_n^{\boldsymbol{\ell}}(\mathbf{h})=\mathbb{E}\left[\Delta^{-\ell_{n+1}}\mathbf{f}_n(X)\Delta^{-\ell_{n+1}}\mathbf{f}_n'(X)\frac{\Gamma_{n+1}^{\boldsymbol{\ell}}\mathbf{h}(X)}{\Delta^{-\ell_{n+1}}h_{n+1}(X)}\right]-R_{n+1}^{\boldsymbol{\ell}}(\mathbf{h}). \tag{5.40}$$

To this purpose, starting from (5.16), we simply apply the same process as above: for $x_{2n+1}<x_{2n+2}$, we use

$$\mathbf{f}_n(x_{2n+2})-\mathbf{f}_n(x_{2n+1})=\mathbb{E}\left[\Delta^{-\ell_{n+1}}\mathbf{f}_n(X)\Phi_p^{\ell_{n+1}}(x_{2n+1},X,x_{2n+2})\right]$$

as well as the Lagrange identity (5.10) and simple conditioning to obtain that

$$\begin{aligned}
R_n^{\boldsymbol{\ell}}(\mathbf{h}) &= \mathbb{E}\Bigg[\left(\mathbf{f}_n(X_{2n+2})-\mathbf{f}_n(X_{2n+1})\right)\left(\mathbf{f}_n(X_{2n+2})-\mathbf{f}_n(X_{2n+1})\right)'\\
&\qquad\Phi_n^{\boldsymbol{\ell}}(X_1,\dots X_{2n+1},X_{2n+2},\dots,X_2)\prod_{i=1}^n\Delta^{-\ell_i}h_i(X_{2i+1},X_{2i+2})\Bigg]\\
&= \mathbb{E}\Bigg[\mathbb{E}\left[\Delta^{-\ell_{n+1}}\mathbf{f}_n(X_{2n+3})\Phi_p^{\ell_{n+1}}(X_{2n+1},X_{2n+3},X_{2n+2})|X_{2n+1},X_{2n+2}\right]\\
&\qquad\mathbb{E}\left[\Delta^{-\ell_{n+1}}\mathbf{f}_n'(X_{2n+4})\Phi_p^{\ell_{n+1}}(X_{2n+1},X_{2n+4},X_{2n+2})|X_{2n+1},X_{2n+2}\right]\\
&\qquad\Phi_n^{\boldsymbol{\ell}}(X_1,\dots X_{2n+1},X_{2n+2},\dots,X_2)\prod_{i=1}^n\Delta^{-\ell_i}h_i(X_{2i+1},X_{2i+2})\Bigg].
\end{aligned}$$

Now for any h_{n+1} such that $\mathbb{P}[\Delta^{-\ell_{n+1}}h_{n+1}(X)>0]=1$, dividing and multiplying by $\sqrt{\Delta^{-\ell_{n+1}}h_{n+1}(X)}$ and applying Lemma 5.2.2 with

$$\mathbf{v}_{n+1}(x)=\frac{\Delta^{-\ell_{n+1}}\mathbf{f}_n(x)}{\sqrt{\Delta^{-\ell_{n+1}}h_{n+1}(x)}}\quad\text{and } g_{n+1}(x)=\sqrt{\Delta^{-\ell_{n+1}}h_{n+1}(x)} \tag{5.41}$$

we obtain with (5.15)

$$
\begin{aligned}
R_n^{\boldsymbol{\ell}}(\mathbf{h}) - \mathbb{E}\Big[& \mathbb{E}\left[R^{\ell_{n+1}}(X_{2n+1}, X_{2n+2}; \mathbf{v_{n+1}}, g_{n+1}) | X_{2n+1}, X_{2n+2}\right] \\
& \mathbb{I}[X_{2n+1} < X_{2n+2}] \Phi_n^{\boldsymbol{\ell}}(X_1, \dots X_{2n+1}, X_{2n+2}, \dots, X_2) \prod_{i=1}^{n} \Delta^{-\ell_i} h_i(X_{2+1}, X_{2i+2}) \Big] \\
= \mathbb{E}\Big[& \mathbb{E}\left[\mathbf{v}_{n+1}(X)\mathbf{v}'_{n+1}(X)\Phi_p^{\ell_{n+1}}(X_{2n+1}, X, X_{2n+2}) | X_{2n+1}, X_{2n+2}\right] \\
& \times \mathbb{E}\left[g_{n+1}^2(X)\Phi_p^{\ell_{n+1}}(X_{2n+1}, X, X_{2n+2}) | X_{2n+1}, X_{2n+2}\right] \mathbb{I}[X_{2n+1} < X_{2n+2}] \\
& \Phi_n^{\boldsymbol{\ell}}(X_1, \dots X_{2n+1}, X_{2n+2}, \dots, X_2) \prod_{i=1}^{n} \Delta^{-\ell_i} h_i(X_{2+1}, X_{2i+2}) \Big] \\
= \mathbb{E}\Big[& \mathbb{E}\left[\mathbf{v}_{n+1}(X)\mathbf{v}'_{n+1}(X)\Phi_p^{\ell_{n+1}}(X_{2n+1}, X, X_{2n+2})\right] \left(h_{n+1}(X_{2n+2}) - h_{n+1}(X_{2n+1})\right) \\
& \Phi_n^{\boldsymbol{\ell}}(X_1, \dots X_{2n+1}, X_{2n+2}, \dots, X_2) \prod_{i=1}^{n} \Delta^{-\ell_i} h_i(X_{2i+1}, X_{2i+2}) \Big] \\
= \mathbb{E}\Bigg[& \Delta^{-\ell_{n+1}} \mathbf{f}_n(X) \Delta^{-\ell_{n+1}} \mathbf{f}'_n(X) \frac{\Gamma_{n+1}^{\boldsymbol{\ell}} \mathbf{h}(X)}{\Delta^{-\ell_{n+1}} h_{n+1}(X)} \Bigg]
\end{aligned} \tag{5.42}
$$

where we used (5.41) in the last step. Thus we have recovered the first summand in (5.40). For the remainder term in (5.42), leaving out the negative sign, the notation (5.41) gives

$$
\begin{aligned}
\mathbb{E}\Big[& \mathbb{E}\left[R^{\ell_{n+1}}(X_{2n+1}, X_{2n+2}; \mathbf{v_{n+1}}, g_{n+1}) | X_{2n+1}, X_{2n+2}\right] \mathbb{I}[X_{2n+1} < X_{2n+2}] \\
& \Phi_n^{\boldsymbol{\ell}}(X_1, \dots X_{2n+1}, X_{2n+2}, \dots, X_2) \prod_{i=1}^{n} \Delta^{-\ell_i} h_i(X_{2i+1}, X_{2i+2}) \Big] \\
= \quad & \mathbb{E}\Big[(\mathbf{v}_{n+1,2n+3} g_{n+1,2n+4} - \mathbf{v}_{n+1,2n+4} g_{n+1,2n+3}) \\
& (\mathbf{v}_{n+1,2n+3} g_{n+1,2n+4} - \mathbf{v}_{n+1,2n+4} g_{n+1,2n+3})' \\
& \Phi_p^{\ell_{n+1}}(X_{2n+1}, X_{2n+3}, X_{2n+4}, X_{2n+2}) \Phi_n^{\boldsymbol{\ell}}(X_1, \dots X_{2n+1}, X_{2n+2}, \dots, X_2) \\
& \prod_{i=1}^{n} \Delta^{-\ell_i} h_i(X_{2i+1}, X_{2i+2}) \Big]
\end{aligned}
$$

Again extracting the common factor $\sqrt{\Delta^{-\ell_{n+1}} h_{n+1}(X_{2n+3}) \Delta^{-\ell_{n+1}} h_{n+1}(X_{2n+4})}$ and re-arranging yields the assertion. □

Proof of Lemma 5.4.1. Let $x_1 \le x \le x_2$ and h an increasing function. Direct application of the definitions with (5.13) lead to

$$
\begin{aligned}
&p(x)\gamma_k^0 h(x_1, x, x_2) \\
&= \int_{x_1}^{x}\int_{x}^{x_2}\int_{x_3}^{x}\int_{x}^{x_4}\cdots\int_{x_{2k-3}}^{x}\int_{x}^{x_{2k-2}} (h(x_{2k}) - h(x_{2k-1})h'(x_{2k-1})h'(x_{2k})\mathrm{d}x_{2k}\mathrm{d}x_{2k-1} \\
&\qquad \cdots h'(x_5)h'(x_6)\mathrm{d}x_6\mathrm{d}x_5 h'(x_3)h'(x_4)\mathrm{d}x_4\mathrm{d}x_3.
\end{aligned}
$$

Applying the change of variables $u_k = h(x_k), k = 1, \ldots, 2k$ and setting $u = h(x)$ we see that the sequence $\gamma_k^0 h$ depends only on the iterated integrals

$$
\iota_k(u_1, u, u_2) := \int_{u_1}^{u}\int_{u}^{u_2}\cdots\int_{u_{2k-3}}^{u}\int_{u}^{u_{2k-2}} (u_{2k} - u_{2k-1})\mathrm{d}u_{2k}\mathrm{d}u_{2k-1}\cdots\mathrm{d}u_4\mathrm{d}u_3
$$

which we can write recursively as

$$
\begin{aligned}
\iota_1(u_1, u, u_2) &= u_2 - u_1 \\
\iota_k(u_1, u, u_2) &= \int_{u_1}^{u}\int_{u}^{u_2} \iota_{k-1}(u_3, u, u_4)\mathrm{d}u_4\mathrm{d}u_3, \qquad k \geq 2.
\end{aligned}
$$

It remains to show that

$$
\iota_k(u_1, u, u_2) = (u_2 - u)^{k-1}(u - u_1)^{k-1}(u_2 - u_1)\frac{\mathbb{I}[u_1 \leq u \leq u_2]}{k!(k-1)!} \tag{5.43}
$$

for all $k \geq 1$. We proceed by induction on k. Clearly $\iota_1(u_1, u, u_2) = (u_2 - u_1)\mathbb{I}[u_1 \leq u \leq u_2]$, as required. Next suppose that (5.43) holds. Then

$$
\begin{aligned}
\iota_{k+1}(u_1, u, u_2) &= \frac{1}{k!(k-1)!}\int_{u_1}^{u}\int_{u}^{u_2}(u_4 - u)^{k-1}(u - u_3)^{k-1}(u_4 - u_3)\mathrm{d}u_4\mathrm{d}u_3 \\
&= \frac{1}{k!(k-1)!}\int_{u_1}^{u}\int_{u}^{u_2}(u_4 - u)^{k}(u - u_3)^{k-1}\mathrm{d}u_4\mathrm{d}u_3 \\
&\quad + \frac{1}{k!(k-1)!}\int_{u_1}^{u}\int_{u}^{u_2}(u_4 - u)^{k}(u - u_3)^{k-1}\mathrm{d}u_4\mathrm{d}u_3 \\
&= \frac{(u_2 - u)^{k+1}(u - u_1)^k + (u_2 - u)^k(u - u_1)^{k+1}}{(k+1)!k!}
\end{aligned}
$$

which leads to the claim. □

Proof of Identity (5.28). Identity (5.28) follows from Lemma 5.4.1 by using $h(X_2) - h(X_1) = h(X_2) - h(x) + h(x) - h(X_1)$ and $\mathbb{I}[X_1 \leq x \leq X_2]\mathbb{I}[X_1 \neq X_2] = \mathbb{I}[X_1 \leq x]\mathbb{I}[X_2 \geq x]\mathbb{I}[X_1 \neq X_2]$ to get

$$
\begin{aligned}
\Gamma_k^{\mathbf{0}} h(x) &= (-1)^{k-1} \frac{1}{p(x)} \mathbb{E}\big[H_x^{k-1}(X)\mathbb{I}[X \le x]\big]\mathbb{E}\big[H_x^{k}(X)\mathbb{I}[X \ge x]\big] \\
&\quad + (-1)^{k} \frac{1}{p(x)} \mathbb{E}\big[H_x^{k}(X)\mathbb{I}[X \le x]\big]\mathbb{E}\big[H_x^{k-1}(X)\mathbb{I}[X \ge x]\big] \\
&= (-1)^{k-1} \mathbb{E}\big[H_x^{k-1}(X)\big] \frac{1}{p(x)} \mathbb{E}\big[H_x^{k}(X)\mathbb{I}[X \ge x]\big] \\
&\quad + (-1)^{k} \mathbb{E}\big[H_x^{k}(X)\big] \frac{1}{p(x)} \mathbb{E}\big[H_x^{k-1}(X)\mathbb{I}[X \ge x]\big]
\end{aligned} \tag{5.44}
$$

where the last equality follows from

$$
\mathbb{E}\big[H_x^k(X)\big] = \mathbb{E}\big[H_x^k(X)\mathbb{I}[X \le x]\big] + \mathbb{E}\big[H_x^k(X)\mathbb{I}[X \ge x]\big].
$$

Upon noting that

$$
\begin{aligned}
&- \mathcal{L}_p^0 H_x^k(x) \\
&= \frac{1}{p(x)} \left\{ \mathbb{E}\big[H_x^k(X_2)\mathbb{I}[X_1 < x < X_2]\big] - \mathbb{E}\big[H_x^k(X_1)\mathbb{I}[X_1 < x < X_2]\big] \right\} \\
&= \frac{1}{p(x)} \left\{ \mathbb{E}\big[H_x^k(X_2)\mathbb{I}[x < X_2]\big]\mathbb{P}[x > X_1] - \mathbb{E}\big[H_x^k(X_1)\mathbb{I}[X_1 < x]\big]\mathbb{P}[x < X_2] \right\} \\
&= \frac{1}{p(x)} \big\{ \mathbb{E}\big[H_x^k(X_2)\mathbb{I}[x < X_2]\big] - \mathbb{E}\big[H_x^k(X_2)\mathbb{I}[x < X_2]\big]\mathbb{P}[x < X_1] \\
&\qquad - \mathbb{E}\big[H_x^k(X_1)\mathbb{I}[X_1 < x]\big]\mathbb{P}[x < X_2] \big\}
\end{aligned}
$$

with $P(x) = \mathbb{P}[X \le x]$ we obtain

$$
\frac{1}{p(x)} \mathbb{E}\big[H_x^k(X)\mathbb{I}[X \ge x]\big] = -\mathcal{L}_p^0 H_x^k(x) + \frac{1 - P(x)}{p(x)} \mathbb{E}\big[H_x^k(X)\big],
$$

the required result is obtained after straightforward simplifications by writing

$$
\begin{aligned}
\Gamma_k^{\mathbf{0}} h(x) &= (-1)^{k-1} \mathbb{E}\big[H_x^{k-1}(X)\big] \frac{1}{p(x)} \mathbb{E}\big[H_x^{k}(X)\mathbb{I}[X \ge x]\big] \\
&\quad + (-1)^{k} \mathbb{E}\big[H_x^{k}(X)\big] \frac{1}{p(x)} \mathbb{E}\big[H_x^{k-1}(X)\mathbb{I}[X \ge x]\big] \\
&= (-1)^{k-1} \left(-\mathbb{E}\big[H_x^{k-1}(X)\big] \mathcal{L}_p^0 H_x^k(x) + \mathbb{E}\big[H_x^k(X)\big] \mathcal{L}_p^0 H_x^{k-1}(x) \right) \\
&\quad + (-1)^{k-1} \frac{1 - P(x)}{p(x)} \left(\mathbb{E}\big[H_x^{k-1}(X)\big]\mathbb{E}\big[H_x^k(X)\big] - \mathbb{E}\big[H_x^k(X)\big]\mathbb{E}\big[H_x^{k-1}(X)\big] \right)
\end{aligned}
$$

and noticing that the last term cancels. □

Proof of Lemma 5.4.3. We shall prove that

$$\gamma_k^{\boldsymbol{\ell}}(x_1,x,x_2) := \gamma_k^{\boldsymbol{\ell}}\mathrm{Id}(x_1,x,x_2) \tag{5.45}$$

$$= (x_2-x)_{\{k-1;\boldsymbol{\ell}\}}(x-x_1)^{\{k-1;\boldsymbol{\ell}\}}(x_2-x_1)\frac{\mathbb{I}[x_1+\mathbf{a}_k\le x\le x_2-\mathbf{b}_k]}{p(x)k!(k-1)!}. \tag{5.46}$$

The claim is obvious from (5.19) in the continuous case. For the discrete case, the assertion is proved by induction in k; the cases $k=1$ and $k=2$ need to be asserted to start the induction. The case $k=1$ is immediate. For $k=2$, we show that, for $\ell_i\in\{-1,1\}$,

$$\gamma_2^{\ell_1,\ell_2}(X_1,x,X_2) = \frac{1}{2}(x-X_1-a_{\boldsymbol{\ell}}(2)+1)(X_2-x-b_{\boldsymbol{\ell}}(2)+1)(X_2-X_1)\frac{\mathbb{I}[X_1+a_{\boldsymbol{\ell}}(2)\le x\le X_2-b_{\boldsymbol{\ell}}(2)]}{p(x)}.$$

To this end, from Proposition 5.4.2 where we sum over (x_3,x_4) instead of (y,z), we obtain

$$\gamma_2^{\ell_1,\ell_2}(x_1,x,x_2) = \sum_{x_3=x_1+a_1}^{x-a_2}\sum_{x_4=x+b_2}^{x_2-b_1}(x_4-x_3)\frac{\mathbb{I}[x_1+\mathbf{a}_2\le x\le x_2-\mathbf{b}_2]}{p(x)}$$
$$= \frac{1}{2}(x-x_1-\mathbf{a}_2+1)(x_2-x-\mathbf{b}_2+1)(x_2-x_1)\frac{\mathbb{I}[x_1+\mathbf{a}_2\le x\le x_2-\mathbf{b}_2]}{p(x)}$$

as required.

To conclude the argument, we prove the identity (5.46) by induction: we suppose the claims hold for k and investigate its validity for $k+1$. The definition of $\Gamma_k^{\boldsymbol{\ell}}$ in (5.15) gives

$$\gamma_{k+1}^{\boldsymbol{\ell}}(x_1,x,x_2) = \mathbb{E}\left[\frac{\chi^{\ell_1}(x_1,X_3)}{p(X_3)}\frac{\chi^{-\ell_1}(X_4,x_2)}{p(X_4)}\gamma_k^{\ell_2,\ldots,\ell_{k+1}}(X_3,x,X_4)\right]. \tag{5.47}$$

Now we can plug-in the induction assumption (5.46) into (5.47):

$$\gamma_{k+1}^{\boldsymbol{\ell}}(x_1,x,x_2)$$
$$= \mathbb{E}\Bigg[(X_4-x-\mathbf{b}'_k+1)^{[k-1]}(x-X_3-\mathbf{a}'_k+1)^{[k-1]}(X_4-X_3)$$
$$\frac{\mathbb{I}[X_3+\mathbf{a}'_k\le x\le X_4-\mathbf{b}'_k]}{p(x)k!(k-1)!}\frac{\chi^{\ell_1}(x_1,X_3)}{p(X_3)}\frac{\chi^{-\ell_1}(X_4,x_2)}{p(X_4)}\Bigg]$$

$$
\begin{aligned}
&\gamma^{\boldsymbol{\ell}}_{k+1}(x_1, x, x_2) \\
&= \sum_{x_3=x_1+a_1}^{x-\mathbf{a}'_k} \sum_{x_4=x+\mathbf{b}'_k}^{x_2-b_1} (x_4 - x - \mathbf{b}'_k + 1)^{[k-1]} (x - x_3 - \mathbf{a}'_k + 1)^{[k-1]} (x_4 - x_3) \\
&\qquad \frac{\mathbb{I}[x_1 + \mathbf{a}_{k+1} \le x \le x_2 - \mathbf{b}_{k+1}]}{p(x)} \\
&= (x_2 - x - \mathbf{b}_{k+1} + 1)^{[k]} (x - x_1 - \mathbf{a}_{k+1} + 1)^{[k]} (x_2 - x_1) \\
&\qquad \frac{\mathbb{I}[x_1 + \mathbf{a}_{k+1} \le x \le x_2 - \mathbf{b}_{k+1}]}{p(x)}
\end{aligned}
$$

where $\mathbf{a}'_k = \sum_{i=2}^{k+1} a_i$ and $\mathbf{b}'_k = \sum_{i=2}^{k+1} b_i$.

□

Proof of Lemma 5.4.5. By Lemma 5.4.1 and (5.44), we have

$$
\begin{aligned}
&\Gamma^0_k P(x) \\
&= \frac{1}{p(x)k!(k-1)!} \mathbb{E}\left[(P(x) - P(X_1))^{k-1} \mathbb{I}[X_1 \le x]\right] \mathbb{E}\left[(P(X_2) - P(x))^k \mathbb{I}[X_2 \ge x]\right] \\
&+ \frac{1}{p(x)k!(k-1)!} \mathbb{E}\left[(P(x) - P(X_1))^k \mathbb{I}[X_1 \le x]\right] \mathbb{E}\left[(P(X_2) - P(x))^{k-1} \mathbb{I}[X_2 \ge x]\right].
\end{aligned}
$$

Moreover, using integration by substitution,

$$
\begin{aligned}
\mathbb{E}\left[(P(x) - P(X_1))^k \mathbb{I}[X_1 \le x]\right] &= \int_a^x (P(x) - P(x_1))^k p(x_1) dx_1 \\
&= -\int_{P(x)}^0 u^k du = \frac{P(x)^{k+1}}{k+1} \\
\mathbb{E}\left[(P(X_2) - P(x))^k \mathbb{I}[X_2 \ge x]\right] &= \int_x^b (P(x_2) - P(x))^k p(x_2) dx_2 \\
&= \int_0^{1-P(x)} u^k du = \frac{(1-P(x))^{k+1}}{k+1},
\end{aligned}
$$

and the conclusion follows.

□

Proof of Proposition 5.4.6. The argument for the integrated Pearson system is inspired from Johnson (1993, Theorem 2). By Lemma 5.4.1, note that

$$\begin{aligned}\gamma_k^0(x_1, x, x_2) &= (x - x_1)^{k-1}(x_2 - x)^{k-1}(x_2 - x_1)\frac{\mathbb{I}[x_1 \leq x \leq x_2]}{p(x)k!(k-1)!} \\ &= (x - x_1)^{k-1}(x_2 - x)^{k-1}(x_2 - \mu + \mu - x_1)\frac{\mathbb{I}[x_1 \leq x]\mathbb{I}[x \leq x_2]}{p(x)k!(k-1)!}\end{aligned}$$

Therefore, $\Gamma_k^0(x)$ can be decomposed using simple expectations:

$$\begin{aligned}\Gamma_k^0(x) =& \frac{1}{p(x)k!(k-1)!}\Bigg(\mathbb{E}\left[(x - X_1)^{k-1}\mathbb{I}[X_1 \leq x]\right]\mathbb{E}\left[(X_2 - \mu)(X_2 - x)^{k-1}\mathbb{I}[x \leq X_2]\right] \\ &+ \mathbb{E}\left[(\mu - X_1)(x - X_1)^{k-1}\mathbb{I}[X_1 \leq x]\right]\mathbb{E}\left[(X_2 - x)^{k-1}\mathbb{I}[x \leq X_2]\right]\Bigg) \qquad (5.48)\end{aligned}$$

In the continuous setting, the Stein kernel τ_p is such that is satisfies for $X \sim p$ with mean μ and differentiable f such that the expectations exist,

$$\mathbb{E}[(X - \mu)f(X)] = \mathbb{E}[\tau_p(X)f'(X)].$$

Integrating by parts we thus obtain

$$\mathbb{E}\left[(X_2 - \mu)(X_2 - x)^{k-1}\mathbb{I}[X_2 \geq x]\right] = \mathbb{E}\left[\tau_p(X_2)(k-1)(X_2 - x)^{k-2}\mathbb{I}[X_2 \geq x]\right]$$

and

$$\mathbb{E}\left[(\mu - X_1)(x - X_1)^{k-1}\mathbb{I}[X_1 \leq x]\right] = \mathbb{E}\left[\tau_p(X_1)(k-1)(x - X_1)^{k-2}\mathbb{I}[X_1 \leq x]\right].$$

When we plug it into (5.48), we get

$$\begin{aligned}\Gamma_k^0(x) = \frac{k-1}{p(x)k!(k-1)!}\mathbb{E}\Bigg[&(x - X_1)^{k-2}(X_2 - x)^{k-2}\mathbb{I}[X_1 \leq x \leq X_2] \\ &(\tau_p(X_2)(x - X_1) + \tau_P(X_1)(X_2 - x))\Bigg].\end{aligned}$$

Using the particular form of τ_p for the integrated Pearson family, Taylor expansion of $\tau_p(X)$ around x gives

$$(x - x_1)\tau_p(x_2) + (x_2 - x)\tau_p(x_1) = \tau_p(x)(x_2 - x_1) + \frac{\tau_p''(x)}{2}(x - x_1)(x_2 - x)(x_2 - x_1)$$

Therefore,

$$\begin{aligned}\Gamma_k^0(x) =& \frac{k-1}{k!(k-1)!}\frac{1}{p(x)}\mathbb{E}\Big[(x-X_1)^{k-2}(X_2-x)^{k-2}\mathbb{I}[X_1\le x\le X_2]\\ &\left(\tau_p(x)(X_2-X_1)+\frac{\tau_p''(x)}{2}(x-X_1)(X_2-x)(X_2-X_1)\right)\Big]\\ =& \frac{\tau_p(x)}{k}\Gamma_{k-1}^0(x)+\frac{\tau_p''(x)(k-1)}{2}\Gamma_k^0(x)\\ =& \frac{1}{k\left(1-\frac{k-1}{2}\tau_p''(x)\right)}\tau_p(x)\Gamma_{k-1}^0(x)\end{aligned}$$

The assertion follows from iterating this expression and using $\Gamma_1^0(x) = \tau_p(x)$ and $\tau_p''(x) = 2\delta$. □

Proof of Proposition 5.4.9. By induction, we only have to prove the relation with respect to ℓ_{k+1}, i.e.,

$$\Gamma_{k+1}^{\boldsymbol{\ell},1}(x) = \frac{\tau_p^+(x-\mathbf{a}_k)}{(k+1)(1-k\delta)}\Gamma_k^{\boldsymbol{\ell}}(x) \text{ and } \Gamma_{k+1}^{\boldsymbol{\ell},-1}(x) = \frac{\tau_p^-(x+\mathbf{b}_k)}{(k+1)(1-k\delta)}\Gamma_k^{\boldsymbol{\ell}}(x).$$

The following argument is inspired from Afendras et al. (2007). Using (5.46) and a similar proof as in the Pearson case (Proposition 5.4.6), we may rewrite $\Gamma_{k+1}^{\boldsymbol{\ell},1}(x)$ using simple expectations:

$$\begin{aligned}\Gamma_{k+1}^{\boldsymbol{\ell},1}(x) = \frac{1}{p(x)}\frac{1}{k!(k+1)!}\Bigg(&\mathbb{E}\left[(x-X_1-\mathbf{a}_k)^{[k]}\mathbb{I}[X_1+\mathbf{a}_k+1\le x]\right]\times\\ &\mathbb{E}\left[(X_2-\mu)(X_2-x-\mathbf{b}_k+1)^{[k]}\mathbb{I}[x\le X_2-\mathbf{b}_k]\right]\\ +\,&\mathbb{E}\left[(\mu-X_1)(x-X_1-\mathbf{a}_k)^{[k]}\mathbb{I}[X_1+\mathbf{a}_k+1\le x]\right]\times\\ &\mathbb{E}\left[(X_2-x-\mathbf{b}_k+1)^{[k]}\mathbb{I}[x\le X_2-\mathbf{b}_k]\right]\Bigg).\end{aligned} \tag{5.49}$$

With the notation (5.2) is it straightforward to verify that for all x we have

$$\Delta^\ell\left(f^{[k]}(x)\right) = f^{[k-1]}(x+a_\ell)\sum_{j=0}^{k-1}\Delta^\ell f(x+j). \tag{5.50}$$

In particular, for all x, a, we have

$$\begin{aligned}\Delta^-\left((x-a+1)^{[k]}\mathbb{I}[x\ge a]\right) &= k(x-a+1)^{[k-1]}\mathbb{I}[x\ge a]\\ \Delta^+\left((a+1-x)^{[k]}\mathbb{I}[x\le a]\right) &= -k(a+1-x)^{[k-1]}\mathbb{I}[x\le a]\end{aligned}$$

$$\Delta^{-}\left((a-x)^{[k]}\mathbb{I}[x<a]\right) = -k(a-x+1)^{[k-1]}\mathbb{I}[x\leq a]$$

The Stein kernel τ_p^ℓ for discrete distributions satisfies for $X \sim p$ with mean μ and functions f such that the expectations exist,

$$\mathbb{E}[(X-\mu)f(X)] = \mathbb{E}[\tau_p^\ell(X)\Delta^{-\ell}f(X-\ell)],$$

see for example Ley et al. (2017b). Hence, with (5.50), we may use the discrete integration by parts formula to rewrite

$$\begin{aligned}
&\mathbb{E}\Big[(X_2-\mu)(X_2-x-\mathbf{b}_k+1)^{[k]}\mathbb{I}[x\leq X_2-\mathbf{b}_k]\Big] \\
&\quad = k\mathbb{E}\Big[\tau_p^+(X_2)(X_2-x-\mathbf{b}_k+1)^{[k-1]}\mathbb{I}[x\leq X_2-\mathbf{b}_k]\Big]
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\Big[(\mu-X_1)(x-X_1-\mathbf{a}_k)^{[k]}\mathbb{I}[X_1\leq x-\mathbf{a}_k-1]\Big] \\
&\quad = \mathbb{E}\Big[(\mu-X_1)(x-X_1-\mathbf{a}_k)^{[k]}\mathbb{I}[X_1\leq x-\mathbf{a}_k]\Big] \\
&\quad = k\mathbb{E}\Big[\tau_p^+(X_1)(x-X_1-\mathbf{a}_k+1)^{[k-1]}\mathbb{I}[X_1\leq x-\mathbf{a}_k]\Big].
\end{aligned}$$

After plugging these equations into (5.49) and some further algebraic developments (which we omit), we obtain

$$\begin{aligned}
&\Gamma_{k+1}^{\boldsymbol{\ell},1}(x) \\
=&\frac{1}{p(x)}\frac{1}{k!(k+1)!}\Bigg(k\tau_p^+(x-\mathbf{a}_k) \\
&\mathbb{E}\Big[(x-X_1-\mathbf{a}_k+1)^{[k-1]}(X_2-x-\mathbf{b}_k+1)^{[k-1]}(X_2-X_1)\mathbb{I}[X_1+\mathbf{a}_k\leq x\leq X_2-\mathbf{b}_k]\Big] \\
&+\delta k\mathbb{E}\Big[(X_2-X_1)(x-X_1-\mathbf{a}_k)(X_2-x+k-\mathbf{b}_k)\mathbb{I}[X_1+\mathbf{a}_k+1\leq x\leq X_2-\mathbf{b}_k]\Big]\Bigg) \\
=&\frac{\tau_p^+(x-\mathbf{a}_k)}{k+1}\Gamma_k^{\boldsymbol{\ell}}(x)+\delta k\Gamma_{k+1}^{\boldsymbol{\ell},1}(x)
\end{aligned}$$

which gives the assertion. The same result can easily be obtained for $\Gamma_{k+1}^{\boldsymbol{\ell},-1}(x)$. □

CHAPTER 6

Stein factors and distances between distributions

1 Introduction

Consider two random variables $X_n, X_\infty \in \mathbb{R}$ such that $\mathcal{L}(X_n) \approx \mathcal{L}(X_\infty)$. There are many ways of quantifying this proximity:

- Kolmogorov distance: $\mathrm{Kol}(X_n, X_\infty) = \sup_{z\in\mathbb{R}} |\mathbb{P}(X_n \le z) - \mathbb{P}(X_\infty \le z)|$
- Total variation distance: $\mathrm{TV}(X_n, X_\infty) = \sup_{B\subset\mathbb{R}} |\mathbb{P}(X_n \in B) - \mathbb{P}(X_\infty \in B)|$
- Wasserstein distance: $\mathrm{Wass}(X_n, X_\infty) = \int_{-\infty}^{\infty} |\mathbb{P}(X_n \le z) - \mathbb{P}(X_\infty \le z)|\,\mathrm{d}z$

and many more (Hellinger, Lévy, Prokhorov, f-divergences, relative entropy, ...). It is generally non-trivial to determine bounds $L_1 \le \mathcal{D}(X_n, X_\infty) \le L_2$ with L_1, L_2 meaningful and computable quantities.

Example 6.1.1 (Berry-Esseen bound $\sim$ 1942)**.** *Let $X_n = n^{-1/2}\sum_{i=1}^n X_i$ with X_i iid mean 0 variance 1 and $X_\infty \sim \mathcal{N}(0,1)$. Then* $\mathrm{Kol}(X_n, X_\infty) \le Cn^{-1/2}\mathbb{E}\big[|X_1|^3\big]$ *for $C \in (0.40973, 0.4748)$.*

Example 6.1.2 (Le Cam's inequality $\sim$ 1960)**.** *Let $X_n = \sum_{i=1}^n X_i$ with $X_i \overset{\text{ind}}{\sim}$ $\mathrm{Bern}(\theta_i)$ and $X_\infty \sim \mathrm{Poi}(\lambda)$ with $\lambda = \sum_{i=1}^n \theta_i$. Here and throughout we write $a \wedge b = \min(a,b)$ and $a \vee b = \max(a,b)$. Then*

$$(1 \wedge \lambda^{-1}) \sum_{i=1}^n \theta_i^2/32 \le \mathrm{TV}(X_n, X_\infty) \le (1 - e^{-\lambda})\lambda^{-1} \sum_{i=1}^n \theta_i^2$$

(the constants are due to Barbour and Hall, 1984).

Examples 6.1.1 and 6.1.2 illustrate situations wherein the "target" law ($\mathcal{L}(X_\infty)$, say) is easy and explicit while the "approximating" law ($\mathcal{L}(X_n)$) is unknown and unfathomable. There is also interest for situations wherein both the target and the approximating distributions are known explicitly.

Example 6.1.3 (Duembgen et al., 2019)**.** *For the sake of illustration we cite the work Duembgen et al. (2019) who provide, by means of direct analysis of the maximal ratio* $\rho(X_n, X_\infty) = \sup_A \mathbb{P}(X_n \in A)/\mathbb{P}(X_\infty \in A)$ *many very competitive bounds for the total variation, including the following.*

- $\mathrm{TV}(\mathrm{Hyp}(N, L, n), \mathrm{Bin}(n, L/N)) \leq (n-1)/N$
- $\mathrm{TV}(\mathrm{Bin}(n, \lambda/n), \mathrm{Poi}(\lambda)) \leq 1 - \left(1 - \frac{\lceil \lambda \rceil}{n}\right)^{1/2}$
- $\mathrm{TV}(\mathrm{Beta}(a, b), \mathrm{Gamma}(a, a+b)) \leq 1 - \left(1 - \frac{a+1}{a+b}\right)^{1/2}$.

There are many ways to prove estimates such as those provided in Examples 6.1.1, 6.1.2, and 6.1.3, such as Fourier methods, couplings or, whenever possible, direct analysis of the densities involved. A now well established general technique for dealing with such problems is *Stein's method*, about which the present work is concerned. In Chapter 3, we prove (see IBP Lemmas 3.2.5 and 3.2.19) that under reasonable conditions on X_∞'s distribution one can associate to it two linear operators $\mathcal{T}_\infty^\ell$ and $\mathcal{L}_\infty^\ell$ such that the "Stein identities"

$$\mathrm{Cov}[f(X_\infty), g(X_\infty)] = \mathbb{E}\left[-\mathcal{L}_\infty^\ell f(X_\infty) \Delta^{-\ell} g(X_\infty)\right] \tag{6.1}$$

$$\mathbb{E}\left[\big(\mathcal{T}_\infty^\ell f(X_\infty)\big) g(X_\infty)\right] = -\mathbb{E}\left[f(X_\infty) \Delta^{-\ell} g(X_\infty)\right] \tag{6.2}$$

are valid for all sufficiently regular functions f, g.

Example 6.1.4. *Take* X_∞ *standard Gaussian with density* $\varphi(x) = (2\pi)^{-1} e^{-x^2/2}$. *Then* $\ell = 0$, $\mathcal{T}_\infty^0 f(x) = f'(x) - xf(x)$, $\mathcal{L}_\infty f(x) = e^{x^2/2} \int_{-\infty}^x (f(u) - \mathbb{E}[f(X_\infty)])\mathrm{d}u$, *so that* (6.1) *and* (6.2) *read as*

$$\mathrm{Cov}[f(X_\infty), g(X_\infty)] = \mathbb{E}\left[\left(e^{X_\infty^2/2} \int_{-\infty}^{X_\infty} (\mathbb{E}[f(X_\infty)] - f(u)) e^{-u^2/2} \mathrm{d}u\right) g'(X_\infty)\right]$$

$$\mathbb{E}\left[\big(f'(X_\infty) - X_\infty f(X_\infty)\big) g(X_\infty)\right] = -\mathbb{E}\left[f(X_\infty) g'(X_\infty)\right]$$

which hold for all $f \in L^1(\varphi)$ *and absolutely continuous functions* g. *Both identities are a straightforward consequence of Fubini's theorem.*

If, in (6.1) or (6.2), we take expectations with respect to some X_n rather than X_∞, absence of equality in either identities for some functions f, g indicates absence of equality between the laws of X_n and X_∞. Stein's method consists in transforming this observation into estimates on relevant probability distances between the laws of X_n and X_∞. More precisely, the method advocates to fix f in (6.1) or (6.2) some "well chosen" function (e.g. $f(x) = 1$, but this is not always ideal) and use the numbers

$$\mathcal{S}_A(X_n, X_\infty, \mathcal{G}) := \sup_{g \in \mathcal{G}} \left|\mathrm{Cov}[f(X_n), g(X_n)] + \mathbb{E}\left[\big(\mathcal{L}_\infty^\ell f(X_n)\big) \Delta^{-\ell} g(X_n)\right]\right| \tag{6.3}$$

$$\mathcal{S}_B(X_n, X_\infty, \mathcal{G}) := \sup_{g\in\mathcal{G}} \left|\mathbb{E}\left[\left(\mathcal{T}_\infty^\ell f(X_n)\right)g(X_n) + f(X_n)\Delta^{-\ell}g(X_n)\right]\right| \tag{6.4}$$

(with $\mathcal{G}$ "some class of functions" to be determined) to quantify the difference between the laws of X_n and X_∞.

Example 6.1.5. *If X_∞ is standard normal, fixing $f(x) = x$ in (6.1) (or $f(x) = 1$ in (6.2)) leads to the discrepancy measure $\sup_{g\in\mathcal{G}} |\mathbb{E}[g'(X_n) - X_n g(X_n)]|$ which, in light of Stein's characterization of the normal distribution, is 0 if and only if X_n is itself Gaussian – at least when $\mathcal{G}$ is a sufficiently large class of test functions. Other choices of f are possible, see Goldstein and Reinert (2005).*

Before diving into the study of the numbers $\mathcal{S}_\bullet(X_n, X_\infty, \mathcal{G})$, it is first necessary to argue as to why such numbers indeed metrize convergence in distribution in terms of relevant metrics. To this end, it suffices to notice that discrepancies $\mathcal{S}_\bullet(X_n, X_\infty, \mathcal{G})$ contain (at least formally) any distance that can be represented as an *Integral Probability Metric* (IPM):

$$\mathcal{D}_\mathcal{H}(X_n, X_\infty) = \sup_{h\in\mathcal{H}} |\mathbb{E}h(X_n) - \mathbb{E}h(X_\infty)|. \tag{6.5}$$

To see why this holds true, fix $f = \eta$ in (6.1) or $f = c$ in (6.2) (the difference in notation is cosmetic but will help at a later stage) and consider the *Stein equations*

$$(\eta(x) - \mathbb{E}\eta(X_\infty))g_h(x) + \left(\mathcal{L}_\infty^\ell \eta(x)\right)\Delta^{-\ell}g_h(x) = h(x) - \mathbb{E}h(X_\infty) \tag{6.6}$$

$$\mathcal{T}_\infty^\ell c(x) g_h^*(x) + c(x)\Delta^{-\ell}g_h^*(x) = h(x) - \mathbb{E}h(X_\infty) \tag{6.7}$$

for all $x \in \mathcal{S}(p_\infty)$. Lemma 3.2.12 guarantees that if $\mathcal{H}$ is reasonable, then for any well-chosen η or c, to every $h \in \mathcal{H}$ we can associate (uniquely) a function g_h or g_h^* such that either (6.6) or (6.7) holds at all x in the support of the law of X_∞. Let $\mathcal{G}_\mathcal{H} = \{g_h \,|\, h \in \mathcal{H}\}$ and $\mathcal{G}_\mathcal{H}^* = \{g_h^\star \,|\, h \in \mathcal{H}\}$ be the collection of all these solutions. Then simple computations show that

$$\mathcal{D}_\mathcal{H}(X_n, X_\infty) = \mathcal{S}_A(X_n, X_\infty, \mathcal{G}_\mathcal{H}) = \mathcal{S}_B(X_n, X_\infty, \mathcal{G}_\mathcal{H}^*).$$

In other words, under non-stated regularity conditions which basically require that all quantities be defined, the IPMs (6.5) can be interpreted as specific instances of Stein's discrepancies $\mathcal{S}_\bullet$.

Example 6.1.6. *Still in the case where X_∞ is standard Gaussian, fix $f = c = 1$ in (6.2) and consider the Stein equation*

$$g'(x) - xg(x) = h(x) - \mathbb{E}h(X_\infty) \tag{6.8}$$

over $x \in \mathbb{R}$. For each $h \in L^1(X_\infty)$ there exists a unique bounded solution given by $g_h(x) = -\mathcal{L}_\infty^0(x) = e^{x^2/2}\int_{-\infty}^x (h(u) - \mathbb{E}h(X_\infty))e^{u^2/2}\mathrm{d}u$, so that

$$\mathcal{D}_{\mathcal{H}}(X_n, X_\infty) = \sup_{h\in\mathcal{H}} |\mathbb{E}[g_h'(X_n) - X_n g_h(X_n)]|$$

and all IPMs with Gaussian target are indeed Stein discrepancies.

Many classical metrics can be represented as IPMs, most notably for us the Kolmogorov, total variation and Wasserstein distances with respective classes

$$\mathcal{H}_{\mathrm{Kol}} = \{h(x) = \mathbb{I}[x \in (-\infty, z]] \text{ such that } z \in \mathbb{R}\}$$
$$\mathcal{H}_{\mathrm{TV}} = \{h(x) = \mathbb{I}[x \in B] \text{ such that } B \in \mathcal{B}(\mathbb{R})\}$$
$$\mathcal{H}_{\mathrm{Wass}} = \mathrm{Lip}(1) = \{h(x) \text{ such that } |h(x) - h(y)| \leq |x - y| \text{ for all } x, y \in \mathbb{R}\}$$

To summarize what has just been written, the heuristic behind our version of Stein's method for a metric of the form (6.5) is to tackle the problem of bounding an IPM by contemplating the identities

$$\begin{aligned}\mathcal{D}_{\mathcal{H}}(X_n, X_\infty) &= \sup_{h\in\mathcal{H}} \left|\mathbb{E}\left[(\eta(x) - \mathbb{E}\eta(X_\infty))g_h(x) + \left(\mathcal{L}_\infty^\ell \eta(x)\right)\Delta^{-\ell} g_h(x)\right]\right| \\ &= \sup_{h\in\mathcal{H}} \left|\mathbb{E}\left[\mathcal{T}_\infty^\ell c(X_n) g_h(X_n) + c(X_n)\Delta^{-\ell} g_h(X_n)\right]\right|\end{aligned}$$

where $g_h(x)$ is solution to either (6.6) (first case) or (6.7) (second case). It remains of course to be able to choose η or c in such a way that the resulting expressions are tractable *and* the corresponding solutions g_h are well behaved.

It is now extremely well documented that, for many classic targets (particularly the normal and Poisson), this approach is powerful because there are many handles for dealing with these numbers, be it via exchangeable pairs, zero- and size bias, Malliavin-Stein, etc. We refer the reader to Barbour et al. (1992), Chen et al. (2011) and Nourdin and Peccati (2012) (among many other possible references) for an in depth overview of a broad variety of applications around the Gaussian and Poisson cases. In this chapter, we adopt the abstract formalism developed in the previous chapters to provide a new point of view on the properties of the solutions to equations (6.6) and (6.7). We have two main types of results.

The first, developed in Section 2.2, is of a classical nature within the theory on Stein's method, and summarized in Proposition 6.2.23: we provide explicit uniform and non-uniform bounds on the solutions to Stein equations and on their derivatives. In all the examples we have considered, our bounds are easily computed and competitive with existing bounds. For instance, applying our bounds to the Gaussian case leads (see Example 6.2.26) to the fact that the solutions to equation (6.8) satisfy

$$|g(x)| \leq \min\left(\kappa_1 \frac{\Phi(x)(1-\Phi(x))}{\varphi(x)}, \kappa_2\right) \leq \min\left(\kappa_1 \frac{1}{2}\sqrt{\frac{\pi}{2}}, \kappa_2\right)$$

$$|g'(x)| \leq \kappa_1 \left(1 + |x| \frac{\Phi(x)(1-\Phi(x))}{\varphi(x)}\right) \leq 2\kappa_1$$
$$|g'(x)| \leq 2\kappa_2 \min\left(|x|, \frac{\int_{-\infty}^{x} \Phi(u)\mathrm{d}u \int_{x}^{\infty}(1-\Phi(u))\mathrm{d}u}{\varphi(x)}\right) \leq 2\kappa_2 \min\left(\sqrt{\frac{2}{\pi}}, |x|\right)$$

where $\kappa_1 \leq 2\|h\|_\infty$ and $\kappa_2 \leq \|h'\|_\infty$. We also compute the bounds for the Poisson (Example 6.2.28) and the exponential (Example 6.2.27).

Our second main result is developed in Section 3, where we propose probabilistic representations of differences between expectations which allow to dispense with the need to bound solutions to Stein equations. As applications we provide new representations for (and bounds on) the Kolmogorov, total variation and Wasserstein distances whenever the target and the approximating random variables are continuous w.r.t. the same dominating measure. For instance in the case of a Gaussian target we obtain (see Example 6.3.7) that if $X_n \sim p_n$ has support $\mathbb{R}$ and score function $\rho_n(x)$ then

$$\begin{aligned}
\mathrm{Kol}(X_n, X_\infty) &= \sup_z \left|\mathbb{E}\left[(X_n + \rho_n(X_n))\frac{\Phi(X_n \wedge z)\bar{\Phi}(X_n \vee z)}{\varphi(X_n)}\right]\right| \\
&\leq \mathbb{E}\left[|X_n + \rho_n(X_n)|\frac{\Phi(X_n)\bar{\Phi}(X_n)}{\varphi(X_n)}\right] \\
&\leq \frac{1}{2}\sqrt{\frac{\pi}{2}}\mathbb{E}\left[|X_n + \rho_n(X_n)|\right],
\end{aligned}$$

and also provide bounds on total variation and Wasserstein distances. We also compare with other available bounds. Our results appear to be competitive with or improve on the current literature on the topic.

The structure of the chapter is as follows. In Section 2, we complete the formalism of Stein's method introduced in Chapter 3. We discuss the properties of solutions to Stein equations in Section 2.1, and provide explicit uniform and non uniform bounds in Section 2.2. In Section 3 we provide new representations for and bounds on the IPMs between densities sharing a common dominating measure, and we apply these in several examples. Most proofs are either omitted or delayed to the Appendix.

2 Stein operators, equations and solutions

Functions of the form $x \mapsto \mathcal{T}_p^\ell f(x)$ or $x \mapsto \mathcal{L}_p^\ell h(x)$, for given special choices of f, h, will play a crucial role in the sequel. Of particular importance is the choice of the constant function $f(x) = 1$, on the one hand, which gives the score function of p, $\rho_p^\ell(x) = \mathcal{T}_p^\ell 1(x) = \Delta^\ell p(x)/p(x)$ (see Definition 3.2.2), and the linear function $h(x) = x$ on the other hand, which defined the Stein kernel, $\tau_p^\ell(x) = -\mathcal{L}_p^\ell \mathrm{Id}(x)$ (see

Definition 3.2.15). The Stein kernels of classic distributions are already provided in Tables 4.1-4.3.

Example 6.2.1 (Gaussian target)**.** *Consider a standard Gaussian target with density* $\varphi(x) \propto e^{-x^2/2}$*. Then* $\ell = 0$*. Simple computations show that* $\rho_\varphi(x) = -x$ *and* $\tau_\varphi(x) = 1$*.*

Example 6.2.2 (Exponential target)**.** *Consider a rate* λ *exponential target with density* $p_{\exp}(x) = \lambda e^{-\lambda x}\mathbb{I}[x \geq 0]$*. Then* $\ell = 0$*. Simple computations show that* $\rho_{\exp}(x) = -\lambda\mathbb{I}[x \geq 0]$ *and* $\tau_{\exp}(x) = x/\lambda$*.*

Example 6.2.3 (Poisson target)**.** *The discrete Poisson target density is* $p_{\text{pois}}(x) = e^{-\lambda}\lambda^x/x!\mathbb{I}[x \geq 0]$*. Then,* $\ell = -1$ *or 1. Simple computations show that* $\rho^+_{\text{pois}}(x) = \lambda/(x+1) - 1$ *and* $\rho^-_{\text{pois}}(x) = 1 - x/\lambda$*,* $\tau^+_{\text{pois}}(x) = x$ *and* $\tau^-_{\text{pois}}(x) = \lambda$*, in all cases for* $x \in \mathbb{N}$*, and 0 elsewhere.*

Another way of writing the Definition 3.2.8 of standardisations of the operator is to consider a function c instead of $\mathcal{L}^\ell_p\eta$, which softens the conditions on Stein class. This is sometimes a better choice because one might want more general coefficient. Therefore, we consider hereafter the adapted definition.

Definition 6.2.4 (Standardizations of the operator)**.** *Let* $\text{dom}(\mathcal{T}^\ell_p)$ *be the collection of functions such that* $c(\cdot)p(\cdot)$ *belongs to* $\text{dom}(\Delta^\ell)$*. A* standardization *of the canonical operator* $\mathcal{T}^\ell_p$ *is any linear operator of the form* $\mathcal{A}g = \mathcal{T}^\ell_p\left(c(\cdot)g(\cdot - \ell)\right)$ *for some* $c \in \text{dom}(\mathcal{T}^\ell_p)$*. That is,*

$$\mathcal{A}g(x) = \mathcal{T}^\ell_p c(x)g(x) + c(x)\Delta^{-\ell}g(x). \tag{6.9}$$

Given some function c*, the corresponding* standardized Stein class *is the collection* $\mathcal{F}(\mathcal{A})$ *of test functions* g *such that* $c(\cdot)g(\cdot - \ell) \in \mathcal{F}^{(1)}_\ell(p)$ *and* $c(\cdot)\Delta^{-\ell}g(\cdot) \in L^1(p)$*.*

By the definitions, it is evident that $\mathbb{E}[\mathcal{A}g(X)] = 0$ for all $g \in \mathcal{F}(\mathcal{A})$. Moreover, we have

$$\mathbb{E}[\mathcal{A}g(X)] = \mathbb{E}[c(X)\Delta^{-\ell}g(X)] + \mathbb{E}[\mathcal{T}^\ell_p c(X)g(X)] = 0 \tag{6.10}$$

for all such g. Equation (6.10) is a *Stein identity*; such identities have many applications as already pointed out in the previous chapters.

Remark 6.2.5. *The most common examples of functions* c *are* $c(x) = 1$ *and* $c(x) = \tau^\ell_p(x)$*; many other choices are of course possible.*

Example 6.2.6 (Gaussian target)**.** *Consider a Gaussian target as in Example 6.2.1. Taking* $c(x) = 1$ *in* (6.9) *(or* $\eta(x) = -x$ *in* (3.8)*) leads to the classic operator* $\mathcal{A}g(x) = g'(x) - xg(x)$ *acting on* $\mathcal{F}(\mathcal{A})$ *the collection of test functions such that*

$$\int_{-\infty}^{\infty} |(g(x)\varphi(x))'|\mathrm{d}x < \infty \text{ and } \lim_{x\to\infty} g(x)\varphi(x) = \lim_{x\to-\infty} g(x)\varphi(x).$$

This is satisfied by all differentiable functions such that $g' \in L^1(\varphi)$, which is the classic class of test functions in this case, see e.g. Nourdin and Peccati (2012, Lemma 3.1.2). Other choices of functions c are possible, leading to other operators for the standard Gaussian.

Example 6.2.7 (Exponential target)**.** *Consider an exponential target as in Example 6.2.2.*

Taking $c(x) = 1$ in (6.9) leads to the operator $\mathcal{A}_1 g(x) = (g'(x) - \lambda g(x))\mathbb{I}[x \geq 0]$, acting on $\mathcal{F}(\mathcal{A}_1)$ the collection of test functions such that

$$\int_0^\infty |(\lambda g(x) e^{-\lambda x})'| \mathrm{d}x < \infty \text{ and } \lim_{x\to\infty} \lambda g(x) e^{-\lambda x} = g(0).$$

In particular, all functions g such that $g(0) = 0$ and $g' \in L^1(p_{\exp})$ are in this class.

Taking $\eta(x) = -x$ in (3.8) (or $c(x) = x/\lambda$ in (6.9)) leads to the operator $\mathcal{A}_2 g(x) = (x/\lambda g'(x) - (x - 1/\lambda)g(x))\mathbb{I}[x \geq 0]$ acting on $\mathcal{F}(\mathcal{A}_2)$ the collection of test functions such that

$$\int_0^\infty |(\lambda x g(x) e^{-\lambda x})'| \mathrm{d}x < \infty \text{ and } \lim_{x\to\infty} x g(x) e^{-\lambda x} = 0.$$

In particular, all functions g such that $xg'(x)$ are in $L^1(p_{\exp})$.

Example 6.2.8 (Poisson target)**.** *Consider a Poisson target as in Example 6.2.3.*

Taking $c(x) = 1$ in (6.9) leads to the operators $\mathcal{A}_1^+ g(x) = \big((\lambda/(x+1) - 1)g(x) + \Delta^- g(x)\big)\mathbb{I}[x \geq 0]$ and $\mathcal{A}_1^- g(x) = \big((1 - x/\lambda)g(x) + \Delta^+ g(x)\big)\mathbb{I}[x \geq 0]$ acting respectively on $\mathcal{F}(\mathcal{A}_1^+)$ the collection of test functions such that

$$\sum_{x=0}^\infty |\Delta^+(g(x)p_{\mathrm{pois}}(x))| < \infty \text{ and } \lim_{x\to\infty} g(x)p_{\mathrm{pois}}(x) = g(0)e^{-\lambda}$$

(in particular all functions g such that $g(0) = 0$ and $\Delta^+ g \in L^1(p_{\mathrm{pois}})$ are in this class) and $\mathcal{F}(\mathcal{A}_1^-)$ the collection of test functions such that

$$\sum_{x=0}^\infty |\Delta^-(g(x)p_{\mathrm{pois}}(x))| < \infty \text{ and } \lim_{x\to\infty} g(x)p_{\mathrm{pois}}(x) = 0$$

(in particular all functions g such that $\Delta^- g \in L^1(p_{\mathrm{pois}})$ are in this class).

Taking $\eta(x) = -x$ in (3.8) leads to the operators $\mathcal{A}_2^+ g(x) = \big((\lambda - x)g(x) + x\Delta^- g(x)\big)\mathbb{I}[x \geq 0]$ and $\mathcal{A}_2^- g(x) = \big((\lambda - x)g(x) + \lambda\Delta^+ g(x)\big)\mathbb{I}[x \geq 0]$ acting respectively on $\mathcal{F}(\mathcal{A}_2^+)$ the collection of test functions such that

$$\sum_{x=0}^\infty |\Delta^+(xg(x)p_{\mathrm{pois}}(x))| < \infty \text{ and } \lim_{x\to\infty} xg(x)p_{\mathrm{pois}}(x) = 0$$

and $\mathcal{F}(\mathcal{A}_2^-)$ the collection of test functions such that $\sum_{x=0}^\infty |\Delta^-(\lambda g(x)p_{\mathrm{pois}}(x))| < \infty$ and $\lim_{x\to\infty} \lambda g(x)p_{\mathrm{pois}}(x) = 0$.

Remark 6.2.9. *If $c \in \mathcal{F}_\ell^{(1)}(p)$, then $\mathcal{F}(\mathcal{A})$* always *contains the constant functions $g(x) = \alpha \in \mathbb{R}$. For instance in the exponential case, $\mathcal{F}(\mathcal{A}_2)$ contains constant functions, whereas $\mathcal{F}(\mathcal{A}_1)$ does not.*

The final ingredient of the theory is to consider a family of Stein equations. According to the adapted standardisation of Definition 6.2.4, we may write the Stein equation (3.9) with respect to the c function.

Definition 6.2.10 (Stein equation)**.** *Let $c \in \mathrm{dom}(\mathcal{T}_p^\ell)$ be such that $c(x) \neq 0$ for all $x \in \mathrm{int}(\mathcal{S}(p))$ the interior of the support (in the discrete case we call $\{a+1, \dots, b-1\}$ the interior). The c-Stein equation for p is*

$$\mathcal{T}_p^\ell c(x) g(x) + c(x) \Delta^{-\ell} g(x) = h(x) - \mathbb{E}\big[h(X)\big] =: \bar{h}(x) \tag{6.11}$$

considered at all $x \in \mathcal{S}(p)$.

Lemma 3.2.12 provide conditions under which, for any $h \in L^1(p)$, there exists a solution $g \in \mathcal{F}(\mathcal{A})$ to (6.6) whose derivative is well defined almost everywhere. The following Lemma is its adaptation to the c-Stein equation (6.11).

Lemma 6.2.11 (Stein solution)**.** *The solution to* (6.11) *is $g_h^{p,\ell,c} =: g$ defined by*

$$g(x) = \frac{\mathcal{L}_p^\ell h(x+\ell)}{c(x+\ell)}. \tag{6.12}$$

with the convention that $g(x) = 0$ for all $x + \ell$ outside of $\mathcal{S}(p)$. This function admits a derivative defined almost everywhere as

$$\Delta^{-\ell} g(x) = \frac{\bar{h}(x) - \mathcal{T}_p^\ell c(x) g(x)}{c(x)} \tag{6.13}$$

$$= \frac{\bar{h}(x) c(x+\ell) - \mathcal{T}_p^\ell c(x) \mathcal{L}_p^\ell h(x+\ell)}{c(x) c(x+\ell)} \tag{6.14}$$

at all $x \in \mathrm{int}(\mathcal{S}(p))$. Moreover, in the discrete case, if $\mathcal{S}(p) = \mathbb{N} \cap [a, b]$, then $\Delta^{-\ell} g(a) = g(a + b_\ell)$ and $\Delta^{-\ell} g(b) = -g(b - a_\ell)$.

Example 6.2.12 (Gaussian target)**.** *Consider a Gaussian target as in Example 6.2.6. The operator leads to the Stein equation $g'(x) - x g(x) = h(x) - \mathbb{E}h(X)$ whose solution in $\mathcal{F}(\mathcal{A})$ is given by*

$$g(x) = e^{x^2/2} \int_{-\infty}^{x} (h(u) - \mathbb{E}h(X)) e^{-u^2/2} \mathrm{d}u. \tag{6.15}$$

Illustrations are provided for $h(x) = \mathbb{I}[x \leq \xi]$ indicator of half lines in Lemma 6.2.15 and Figure 6.1.

Example 6.2.13 (Exponential target)**.** *Consider an exponential target as in Example 6.2.7. The first operator $\mathcal{A}_1$ leads to the Stein equation $g_1'(x) - \lambda g_1(x) = h(x) - \mathbb{E}[h(X)]$ on $[0,\infty)$ whose solution in $\mathcal{F}(\mathcal{A}_1)$ is given by*

$$g_1(x) = \left(e^{\lambda x} \int_0^x (h(u) - \mathbb{E}h(X)) e^{-\lambda u} \mathrm{d}u \right) \mathbb{I}[x \geq 0]. \tag{6.16}$$

Illustrations are provided for $h(x) = \mathbb{I}[x \leq \xi]$ indicator of half lines in Lemma 6.2.15 and Figure 6.2. The second operator $\mathcal{A}_2$ leads to the Stein equation $x/\lambda g_2'(x) - (x - 1/\lambda) g_2(x) = h(x) - \mathbb{E}[h(X)]$ (still restricted to $[0,\infty)$) whose solution in $\mathcal{F}(\mathcal{A}_2)$ is given by

$$g_2(x) = \left(\frac{\lambda}{x} e^{\lambda x} \int_0^x (h(u) - \mathbb{E}h(X)) e^{-\lambda u} \mathrm{d}u \right) \mathbb{I}[x \geq 0]. \tag{6.17}$$

Illustrations are provided for $h(x) = \mathbb{I}[x \leq \xi]$ indicator of half lines in Lemma 6.2.15 and Figure 6.3.

Example 6.2.14 (Poisson target)**.** *Consider a Poisson target as in Example 6.2.8. The first operators $\mathcal{A}_1^+$ and $\mathcal{A}_1^-$ leads to the Stein equations $(\lambda/(x+1) - 1) g_1^+(x) + \Delta^- g_1^+(x) = h(x) - \mathbb{E}[h(X)]$ and $(1 - x/\lambda) g_1^-(x) + \Delta^+ g_1^-(x) = h(x) - \mathbb{E}[h(X)]$ on positive integers whose solutions in $\mathcal{F}(\mathcal{A}_1^+)$ and $\mathcal{F}(\mathcal{A}_1^-)$ are given by*

$$g_1^+(x) = \left(\frac{1}{p_{\mathrm{pois}}(x+1)} \sum_{j=0}^{x} (h(j) - \mathbb{E}h(X)) p_{\mathrm{pois}}(j)) \right) \mathbb{I}[x \geq 0],$$

$$g_1^-(x) = \left(\frac{1}{p_{\mathrm{pois}}(x-1)} \sum_{j=0}^{x-1} (h(j) - \mathbb{E}h(X)) p_{\mathrm{pois}}(j)) \right) \mathbb{I}[x > 0].$$

Illustrations are provided for the point mass $h(x) = \mathbb{I}[x = \xi]$ in Lemma 6.2.16 and Figure 6.4.

The other operators $\mathcal{A}_2^+$ and $\mathcal{A}_2^-$ leads to the Stein equations $(\lambda - x) g_2^+(x) + x\Delta^- g_2^+(x) = h(x) - \mathbb{E}[h(X)]$ and $(\lambda - x) g_2^-(x) + \lambda \Delta^+ g_2^-(x) = h(x) - \mathbb{E}[h(X)]$ on positive integers whose solutions in $\mathcal{F}(\mathcal{A}_2^+)$ and $\mathcal{F}(\mathcal{A}_2^-)$ are given by

$$g_2^+(x) = \left(\frac{1}{(x+1) p_{\mathrm{pois}}(x+1)} \sum_{j=0}^{x} (h(j) - \mathbb{E}h(X)) p_{\mathrm{pois}}(j)) \right) \mathbb{I}[x \geq 0], \tag{6.18}$$

$$g_2^-(x) = \left(\frac{1}{\lambda p_{\mathrm{pois}}(x-1)} \sum_{j=0}^{x-1} (h(j) - \mathbb{E}h(X)) p_{\mathrm{pois}}(j)) \right) \mathbb{I}[x > 0]. \tag{6.19}$$

Illustrations are provided for the point mass $h(x) = \mathbb{I}[x = \xi]$ in Lemma 6.2.16.

In this chapter we shall concentrate on four classes of test functions $\mathcal{H}$: (i) Lipschitz, (ii) bounded, (iii) indicator for a half-line (i.e. $h(x) = \mathbb{I}[x \leq z]$ for some z), and (iv) Dirac delta at some point ($h(x) = \mathbb{I}[x = \xi]$ for some $\xi \in \mathcal{S}(p)$). As mentioned in the Introduction, these choices correspond in the Steinian approach to some of the more classic integral probability metrics, namely the Wasserstein distance (case (i)), the total variation distance (cases (ii) and (iv)), and the Kolmogorov distance, case (iii). There is, however, in principle no need to restrict only to this choice of classes of test functions.

2.1 The solutions to Stein equations

We study the solutions g_h and their derivatives $\Delta^{-\ell} g_h$ from Lemma 6.2.11. If P is the cdf of a density p, its survival function is $\bar{P} = 1 - P$.

Lemma 6.2.15 (Lower half-line indicators, $\ell = 0$)**.** *Let $\ell = 0$ (i.e. p is absolutely continuous w.r.t. the Lebesgue measure). If $h(x) = \mathbb{I}[x \leq \xi]$, the Stein equation* (6.11) *for p is*

$$\mathcal{T}_p^0 c(x) g(x) + c(x) g'(x) = \mathbb{I}[x \leq \xi] - P(\xi).$$

The solutions (6.12) *are*

$$g(x) = \frac{1}{c(x)} \frac{P(\xi \wedge x)\bar{P}(\xi \vee x)}{p(x)} \tag{6.20}$$

still with the convention that the functions are set to 0 outside the support of p. The derivatives (6.13) *of these solutions are*

$$g'(x) = \frac{\mathbb{I}[x \leq \xi] - P(\xi)}{c(x)} - \frac{\mathcal{T}_p^0 c(x)}{c^2(x)} \frac{P(\xi \wedge x)\bar{P}(\xi \vee x)}{p(x)}. \tag{6.21}$$

Lemma 6.2.16 (Point mass, $\ell = \pm 1$)**.** *Let $\ell = \pm 1$ (i.e. p is absolutely continuous w.r.t. the counting measure). Let $h(x) = \mathbb{I}[x = \xi]$. The Stein equation* (6.11) *for p is*

$$\mathcal{T}_p^\ell c(x) g(x) + c(x) \Delta^{-\ell} g(x) = \mathbb{I}[x = \xi] - p(\xi) \tag{6.22}$$

and the solutions (6.12) *are given by*

$$g_\xi^\ell(x) = \frac{p(\xi)}{c(x+\ell)p(x+\ell)} \left(\mathbb{I}[x \geq \xi + b_\ell] - P(x - b_\ell) \right) \tag{6.23}$$

If, moreover, $c = \tau_p^\ell$ then the derivatives (6.13) *satisfy*

$$\Delta^{-\ell} g_\xi^\ell(x) = \frac{\mathbb{I}[x = \xi] - p(\xi)}{\tau_p^+(x)} + \frac{p(\xi)(\mathbb{I}[x \geq \xi] - P(x))}{p(x)} \left(\frac{1}{\tau_p^-(x)} - \frac{1}{\tau_p^+(x)} \right) \tag{6.24}$$

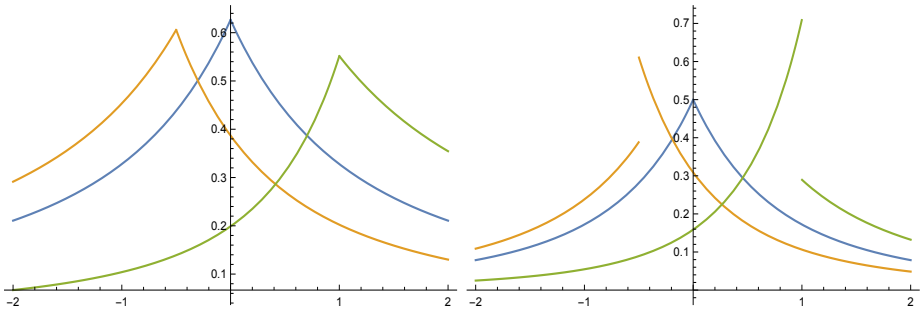


Figure 6.1: Solution (6.20) (left plot) and absolute value of its derivative (6.21) (right plot) for Gaussian target with $c(x) = 1$ and, in both plots, $\xi = -0.5$ (orange curves), $\xi = 0$ (blue curves) and $\xi = 1$ (green curves).

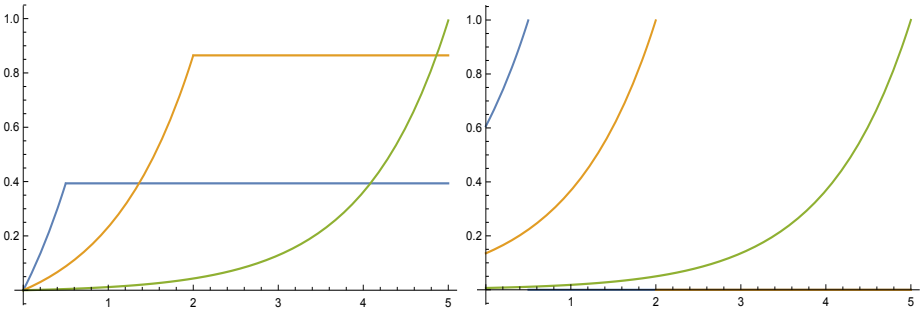


Figure 6.2: Solution (6.20) (left plot) and absolute value of its derivative (6.21) (right plot) for exponential target with $c(x) = 1$ and, in both plots, $\xi = 0.5$ (blue curves), $\xi = 2$ (orange curves) and $\xi = 5$ (green curves).

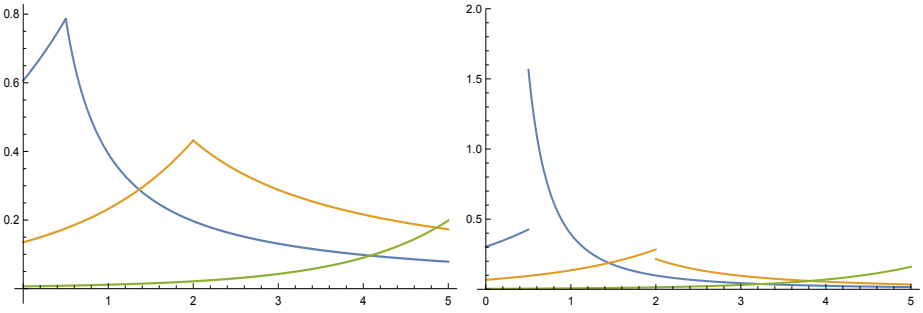


Figure 6.3: Solution (6.20) (left plot) and and absolute value of its derivative (6.21) (right plot) for exponential target with $c(x) = x$ and, in both plots, $\xi = .05$ (blue curves), $\xi = 2$ (orange curves) and $\xi = 5$ (green curves).

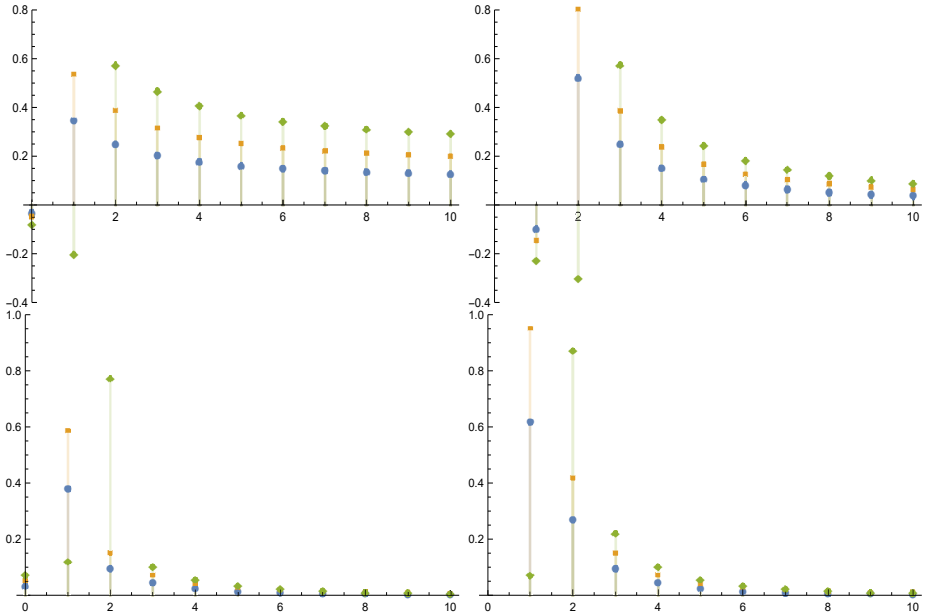


Figure 6.4: Solutions (6.23) (upper panels) and and absolute value of their derivatives (6.24) (lower panels) for Poisson target of parameter 3 with $c(x) = 1$, $\ell = 1$ (left plot) and $\ell = -1$ (right plot) and, in all plots, $\xi = 0.5$ (blue curves), $\xi = 1$ (orange curves) and $\xi = 2$ (green curves).

Remark 6.2.17. *The result of point mass can easily be extended to any Borel set A. Following the proof of Barbour et al. (1992, Lemma 1.1.1), for any $A \subset \mathcal{S}(p)$, the Stein equation* (6.11) *for p can be written*

$$\mathcal{T}_p^\ell c(x)g(x) + c(x)\Delta^{-\ell}g(x) = \mathbb{I}_A(x) - \mathbb{E}[\mathbb{I}_A(X)]$$

and the solutions (6.12) *are now given by*

$$g_A^\ell(x) = \frac{1}{c(x+\ell)p(x+\ell)} \left(\sum_{\xi \in A} p(\xi)\mathbb{I}[x \geq \xi + b_\ell] - P(x - b_\ell)\mathbb{E}[\mathbb{I}_A(X)] \right) = \sum_{\xi \in A} g_\xi^\ell(x)$$

if g_ξ^ℓ is the solution of Stein equation (6.22) *for the point mass function $h_\xi(x) = \mathbb{I}[x = \xi]$.*

In order to work for unspecified functions h, consider several probabilistic representations of the inverse operator presented in Chapter 3, namely Equations (3.17), (3.19) and (3.21). Throughout the section, all results are stated with the implicit assumption that all functions exist and that the various expectations are defined. The following Lemma is only a rewriting of the last representation.

Lemma 6.2.18. *We define the symmetric positive kernel*

$$\tilde{K}_p^\ell(x,y) = \frac{P(x \wedge y - a_\ell)\bar{P}(x \vee y - a_\ell)}{p(x)p(y)}.$$

Then, for all functions $h \in L^1(p)$, we have

$$-\mathcal{L}_p^\ell h(x) = \mathbb{E}\left[\tilde{K}_p^\ell(x,X)\Delta^{-\ell}h(X)\right]. \tag{6.25}$$

The proof of the next useful lemma is detailed in the Appendix.

Lemma 6.2.19. *Define*

$$R_p^\ell(x,y) = \chi^{-\ell}(y,x)\frac{P(y-a_\ell)}{p(y)} - \chi^\ell(x,y)\frac{\bar{P}(y-a_\ell)}{p(y)} = \frac{1}{p(y)}\big(P(y-a_\ell) - \chi^\ell(x,y)\big).$$

Then

$$\bar{h}(x) = \mathbb{E}[R_p^\ell(x,X)\Delta^{-\ell}h(X)]. \tag{6.26}$$

Remark 6.2.20. *It is easy to show that $\mathbb{E}\left[\tilde{K}_p^\ell(x,X)\right] = \tau_p^\ell(x)$ (the Stein kernel of p), and $\mathbb{E}[R_p^\ell(x,X)] = x - \mathbb{E}[X]$.*

With these notations in hand, the following result holds (proof in the Appendix).

Lemma 6.2.21 (Representation formulae)**.** *The solutions* (6.12) *can be written:*

$$g(x) = -\frac{\mathbb{E}\big[(h(X_2)-h(X_1))\chi^\ell(X_1,x+\ell,X_2)\big]}{p(x+\ell)c(x+\ell)} \tag{6.27}$$

$$= -\frac{\mathbb{E}\left[\tilde{K}_p^\ell(x+\ell,X)\Delta^{-\ell}h(X)\right]}{c(x+\ell)}. \tag{6.28}$$

The derivatives (6.13) *can be written:*

$$\Delta^{-\ell}g(x) = \frac{\bar{h}(x)}{c(x)} + \frac{\mathcal{T}_p^\ell c(x)}{c(x)}\frac{\mathbb{E}[(h(X_2)-h(X_1))\chi^\ell(X_1,x+\ell,X_2)]}{c(x+\ell)p(x+\ell)} \tag{6.29}$$

$$= \frac{\mathbb{E}\Big[\big(R_p^\ell(x,X)c(x+\ell) + \mathcal{T}_p^\ell c(x)\tilde{K}_p^\ell(x+\ell,X)\big)\Delta^{-\ell}h(X)\Big]}{c(x)c(x+\ell)}. \tag{6.30}$$

If, moreover, $c \in \mathcal{F}_\ell^{(1)}(p)$ then, setting $\bar{\eta}(x) = \mathcal{T}_p^\ell c(x)$, the derivatives (6.13) *can further be simplified as:*

$$\Delta^{-\ell}g(x) = \frac{\mathbb{E}\left[\big(\bar{\eta}(x)\big(h(X_2)-h(X_1)\big) - \bar{h}(x)\big(\eta(X_2)-\eta(X_1)\big)\big)\,\chi^\ell(X_1,x+\ell,X_2)\right]}{p(x+\ell)\mathcal{L}_p^\ell\eta(x)\mathcal{L}_p^\ell\eta(x+\ell)} \tag{6.31}$$

(6.32)

and

$$\Delta^{-\ell} g(x) = \frac{1}{p(x+\ell)\mathcal{L}_p^\ell \eta(x)\mathcal{L}_p^\ell \eta(x+\ell)} \times \tag{6.33}$$
$$\left(\mathbb{E}\left[\Delta^{-\ell} h(X) \frac{\bar{P}(X-a_\ell)}{p(X)} \chi^\ell(x,X)\right] \mathbb{E}\left[\Delta^{-\ell}\eta(X)\frac{P(X-a_\ell)}{p(X)}\chi^{-\ell}(X,x)\right] \right.$$
$$\left. - \mathbb{E}\left[\Delta^{-\ell} h(X) \frac{P(X-a_\ell)}{p(X)} \chi^{-\ell}(X,x)\right] \mathbb{E}\left[\Delta^{-\ell}\eta(X)\frac{\bar{P}(X-a_\ell)}{p(X)}\chi^{\ell}(x,X)\right] \right).$$

2.2 Stein factors

We start with the discrete case, by following arguments in Ehm (1991), Barbour et al. (1992), Erhardsson (2005) to obtain the following result (proof in the Appendix).

Lemma 6.2.22 (Discrete case, point mass)**.** *Let $\ell = \pm 1$. Consider g_ξ^ℓ the solution to the Stein equation*

$$\tau_p^\ell(x)\Delta^{-\ell} g(x) - (x - \mathbb{E}[X])g(x) = \mathbb{I}[x=\xi] - p(\xi) \tag{6.34}$$

If the ratio $\frac{P(x-1)}{\tau_p^+(x)p(x)}$ is non decreasing for $x \le \xi$ and the ratio $\frac{1-P(x-1)}{\tau_p^+(x)p(x)}$ is non increasing for $x > \xi$ then

$$\|g_\xi^\ell\|_\infty \le \max\left\{ \frac{P(\xi-1)}{\tau_p^+(\xi)}, \frac{1-P(\xi)}{\tau_p^-(\xi)} \right\}, \tag{6.35}$$

and

$$\|\Delta g_\xi^\ell\|_\infty = \frac{P(\xi-1)}{\tau_p^+(\xi)} + \frac{1-P(\xi)}{\tau_p^-(\xi)} \le \begin{cases} \frac{1-p(\xi)}{\tau_p^+(\xi)} & \text{if } \xi \le \mathbb{E}[X] \\ \frac{1-p(\xi)}{\tau_p^-(\xi)} & \text{if } \xi \ge \mathbb{E}[X] \end{cases} \tag{6.36}$$
$$\le \frac{1-p(\xi)}{\min\{\tau_p^+(\xi), \tau_p^-(\xi)\}}$$

More generally, for any Borel set A,

$$\|g_A^\ell\|_\infty \le \left(\sum_{j\in A} p(j)\right) \sup_{\xi\in A}\left\{ \frac{1}{\tau_p^+(\xi)p(\xi)}, \frac{1}{\tau_p^-(\xi)p(\xi)} \right\} \tag{6.37}$$

and

$$\|\Delta g_A^\ell\|_\infty \le \sup_{\xi\in A}\left(\frac{P(\xi-1)}{\tau_p^+(\xi)} + \frac{1-P(\xi)}{\tau_p^-(\xi)} \right) =: \sup_{\xi\in A} B_p(\xi) \tag{6.38}$$

For general h, representations (6.27) to (6.33) lead to the following bounds.

Proposition 6.2.23. *Let*

$$\kappa_1(h) = \sup_{y\in\mathcal{S}(p)} h(y) - \inf_{y\in\mathcal{S}(p)} h(y) \text{ and } \kappa_2(h) = \sup_{y\in\mathcal{S}(p)} |\Delta^{-\ell}h(y)|.$$

Let g be the function defined by (6.12). Suppose that $c > 0$ on the interior of the support of p. Then

1. *If h is bounded then*

$$|g(x)| \le \kappa_1(h)\frac{P(x-b_\ell)\bar{P}(x-b_\ell)}{p(x+\ell)}\frac{1}{c(x+\ell)} \tag{6.39}$$

and

$$\left|\Delta^{-\ell}g(x)\right| \le \kappa_1(h)\frac{1}{c(x)}\left(1 + \frac{|\mathcal{T}_p^\ell c(x)|}{c(x+\ell)}\frac{P(x-b_\ell)\bar{P}(x-b_\ell)}{p(x+\ell)}\right). \tag{6.40}$$

2. *If $\Delta^{-\ell}h$ exists and is bounded then*

$$|g(x)| \le \kappa_2(h)\frac{\tau_p^\ell(x+\ell)}{c(x+\ell)} \tag{6.41}$$

and

$$\left|\Delta^{-\ell}g(x)\right| \le \kappa_2(h)\left(\frac{|x-\mathbb{E}[X]|}{c(x)} + \frac{|\mathcal{T}_p^\ell c(x)|}{c(x)}\frac{\tau_p^\ell(x+\ell)}{c(x+\ell)}\right). \tag{6.42}$$

If, moreover, $c \in \mathcal{F}_\ell^{(1)}(p)$ is of the form $c = -\mathcal{L}_p^\ell\eta$, then the following also hold true.

3. *If h satisfies $|h(x)-h(y)| \le k|\eta(x)-\eta(y)|$ then*

$$\|g\|_\infty \le k. \tag{6.43}$$

4. *If h is bounded then*

$$|\Delta^{-\ell}g(x)| \le \kappa_1(h)\frac{1}{-\mathcal{L}_p^\ell\eta(x)}\left(1 + \frac{|\bar{\eta}(x)|}{-\mathcal{L}_p^\ell\eta(x+\ell)}\frac{P(x-b_\ell)\bar{P}(x+a_\ell)}{p(x+\ell)}\right). \tag{6.44}$$

5. *If $\Delta^{-\ell}h$ exists and is bounded then*

$$\begin{aligned}|\Delta^{-\ell}g(x)| \le{}& \kappa_2(h)\frac{1}{p(x+\ell)\big(-\mathcal{L}_p^\ell\eta(x)\big)\big(-\mathcal{L}_p^\ell\eta(x+\ell)\big)}\\ &\times\left(\mathbb{E}\left[\frac{\bar{P}(X+b_\ell)}{p(X)}\chi^\ell(x,X)\right]\mathbb{E}\left[\Delta^{-\ell}\eta(X)\frac{P(X-a_\ell)}{p(X)}\chi^{-\ell}(X,x)\right]\right.\\ &\left.+\,\mathbb{E}\left[\frac{P(X-a_\ell)}{p(X)}\chi^{-\ell}(X,x)\right]\mathbb{E}\left[\Delta^{-\ell}\eta(X)\frac{\bar{P}(X+b_\ell)}{p(X)}\chi^\ell(x,X)\right]\right)\end{aligned} \tag{6.45}$$

In order to lighten the notations, in the sequel we write κ_j for $\kappa_j(h)$, $j = 1, 2$.

Remark 6.2.24. *We remark that the non uniform bounds in* (6.41) *and* (6.44) *are* exactly *the optimal bounds for all Lipschitz-continuous functions h among all bounds involving the factor $\kappa_2(h) = \|h'\|_\infty$, as demonstrated in Döbler (2015, Proposition 3.13). Taking $\ell = 0$ and $c(x) = 1$ leads to (improvements of) the bounds discussed in Chatterjee and Shao (2011) (see their Lemma 4.1).*

Remark 6.2.25. *There exist many papers with bounds on Stein factors. There is often a difference in scaling between our Stein equation and the one used in those papers, that is why we use some function η and the literature rather uses $r\eta$ for some scalar factor $r \neq 0$. Such scaling obviously has an effect on the bounds, which have to be divided by powers of $|r|$ according to the occurrences of η in their expressions. An important reference on Stein factors is Döbler and Peccati (2018) who consider the case of a gamma target. We do not recover their results exactly, because in that paper the equations are extended to the real line. See also Döbler (2012) (i.e. the arXiv version of Döbler, 2015) for an in depth first study of the problem of extending Stein equations outside the support of the target.*

Example 6.2.26 (Standard normal distribution)**.** *Continuing Example 6.2.12, we consider g the solution to*

$$g'(x) - xg(x) = h(x) - \mathbb{E}[h(X)]$$

given in (6.15)*. Applying Proposition 6.2.23, the following holds:*

$$|g(x)| \leq \min\left(\kappa_1 \frac{\Phi(x)(1-\Phi(x))}{\varphi(x)}, \kappa_2\right) \leq \min\left(\kappa_1 \frac{1}{2}\sqrt{\frac{\pi}{2}}, \kappa_2\right)$$
$$|g'(x)| \leq \kappa_1\left(1 + |x|\frac{\Phi(x)(1-\Phi(x))}{\varphi(x)}\right) \leq 2\kappa_1$$
$$|g'(x)| \leq 2\kappa_2 \min\left(|x|, \frac{\int_{-\infty}^x \Phi(u)\mathrm{d}u \int_x^\infty (1-\Phi(u))\mathrm{d}u}{\varphi(x)}\right) \leq 2\kappa_2 \min\left(\sqrt{\frac{2}{\pi}}, |x|\right).$$

To our own surprise, the first bound (both the uniform and the non-uniform one) appears to be a strict improvement on the known bound in this case, from e.g. Chen et al. (2011, Lemma 2.4) or Nourdin and Peccati (2012, Theorem 3.3.1). Each of the uniform bounds are equivalent to the known bound in this case; it is not clear to us whether the non uniform bounds are known (though, once again, we stress that the bounds involving κ_2 are in some sense available in Döbler (2015)).

Example 6.2.27 (Exponential distribution)**.** *Continuing Example 6.2.13, we consider the two different situations. First, g_1 is solution to*

$$g_1'(x) - \lambda g_1(x) = h(x) - \mathbb{E}[h(X)]$$

over the positive real line, given by (6.16)*. Applying Proposition 6.2.23 (with* $c(x) = 1$ *and* $\tau^0_{\exp}(x) = \lambda x$*), the following holds:*

$$|g_1(x)| \leq \frac{1}{\lambda} \min\left(\kappa_1(1 - e^{-\lambda x}), \kappa_2 x\right)$$
$$|g_1'(x)| \leq \min\left(\kappa_1\left(2 - e^{-\lambda x}\right), \kappa_2(|x - \lambda| + x)\right).$$

Note that only items 1 and 2 apply because $c(x) = 1 \notin \mathcal{F}^{(1)}_\ell(\exp)$*. Second,* g_2 *is solution to*

$$\frac{x}{\lambda} g_2'(x) - \left(x - \frac{1}{\lambda}\right) g_2(x) = h(x) - \mathbb{E}[h(X)]$$

over the positive real line, given by (6.17)*. Here all the items of Proposition 6.2.23 apply (with* $c(x) = x/\lambda$*), yielding*

$$|g_2(x)| \leq \min\left(\kappa_1 \frac{1 - e^{-\lambda x}}{x}, \kappa_2\right)$$
$$|g_2'(x)| \leq \kappa_1 \frac{\lambda}{x}\left(1 + \left|x - \frac{1}{\lambda}\right| \frac{1 - e^{-\lambda x}}{x}\right)$$
$$|g_2'(x)| \leq 2\kappa_2 \min\left(\left|\lambda - \frac{1}{x}\right|, \frac{1}{x}\left(1 - \frac{1 - e^{-\lambda x}}{\lambda x}\right)\right).$$

The first bound is uniformly smaller than the bound $1/x$ *of Chatterjee et al. (2011) (bound for* $\lambda = 1$*); the other bounds are of same order than Chatterjee et al.' bound for* $\lambda = 1$*.*

Example 6.2.28 (Poisson distribution)**.** *We continue Example 6.2.14. We consider the solutions* g^+ *and* g^- *to*

$$x\Delta^- g^+(x) - (x - \lambda) g^+(x) = h(x) - \mathbb{E}[h(X)]$$
$$\lambda\Delta^+ g^-(x) - (x - \lambda) g^-(x) = h(x) - \mathbb{E}[h(X)]$$

given in (6.18) *and* (6.19)*, respectively. Recall that* g^- *is the classic solution to the usual equation for the Poisson; also* $g^+(x) = g^-(x + 1)$ *and* $\Delta^+ g^-(x) = \Delta^- g^+(x)$*. Applying Proposition 6.2.23 (with* $\ell = -1$ *and* $c(x) = \lambda$ *or* $\ell = 1$ *and* $c(x) = x$*), the following holds:*

$$|g^-(x)| \leq \min\left(\kappa_1 \frac{P(x-1)\bar{P}(x-1))}{\lambda p(x-1)}, \kappa_2\right) \tag{6.46}$$
$$|\Delta^+ g^-(x)| \leq \kappa_1 \min\left(\frac{1}{\lambda} + \frac{|x - \lambda|}{\lambda^2} \frac{P(x-1)\bar{P}(x-1)}{p(x-1)}, \frac{1}{x} + \frac{|x - \lambda|}{x(x+1)} \frac{P(x)\bar{P}(x)}{p(x+1)}\right) \tag{6.47}$$
$$|\Delta^+ g^-(x)| \leq 2\kappa_2 \min\left(\frac{|x - \lambda|}{\lambda}, \frac{|x - \lambda|}{x}, \frac{\sum_{j=0}^{x-1} P(j) \sum_{j=x}^{\infty} \bar{P}(j)}{\lambda x p(x)}\right) \tag{6.48}$$

(we only give the bounds in terms of g^-; those for g^+ follow trivially). One can see, as illustrated Figure 6.5(a), that the non uniform bound in (6.46) *is strictly smaller than* $1 \wedge \sqrt{2/(e\lambda)}$ *which thus yields an improvement on the classic bound, e.g. in Erhardsson (2005, Theorem 2.3); the constant bound – in terms of* κ_2 *– is already available in Barbour et al. (1992, Remark 1.1.6) (proof in Barbour and Xia, 2006). The bound* (6.47) *is of similar order to the classical* $(1-e^{-\lambda})/\lambda$ *(see Figure 6.5(b)), but does not improve everywhere. Finally the bound* (6.48) *strictly improves on the bound* $1 \wedge 8/(3\sqrt{2e\lambda})$ *from Barbour et al. (1992), as illustrated Figure 6.5(c) for* $\lambda = 10$.

Lemma 6.2.22 also applies to this case, because the Poisson distribution satisfies the conditions (monotonicity of the two ratios for any $\xi \in \mathcal{S}(p)$*). Therefore, the bound* (6.35) *on the solution of equation* (6.34) *becomes:*

$$\|g_\xi\|_\infty \leq \max\left\{\frac{P(\xi-1)}{\xi}, \frac{1-P(\xi)}{\lambda}\right\}, \tag{6.49}$$

as illustrated Figure 6.6(a). Moreover, the bound (6.36) *becomes*

$$||\Delta^+ g_\xi||_\infty = \frac{P(\xi-1)}{\xi} + \frac{1-P(\xi)}{\lambda} \leq \min\left\{\frac{1}{\xi}, \frac{1-e^{-\lambda}}{\lambda}\right\}. \tag{6.50}$$

For any Borel set $A \subset \mathcal{S}(p)$*, the solution is bounded by* (6.37)

$$\|g_A\|_\infty \leq \left(\sum_{j\in A} p(j)\right) \sup_{\xi\in A}\left\{\frac{1}{\xi p(\xi)}, \frac{1}{\lambda p(\xi)}\right\}$$

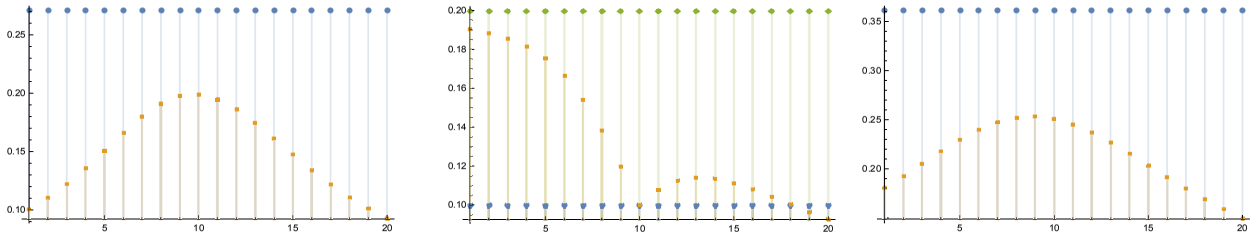
and the bound (6.38) *gives*

$$||\Delta g_A||_\infty \leq \sup_{x\in A}\left(\frac{P(x-1)}{x} + \frac{1-P(x)}{\lambda}\right) \leq \frac{1-e^{-\lambda}}{\lambda}$$

which is the bound given in Barbour et al. (1992, Lemma 1.1.1).

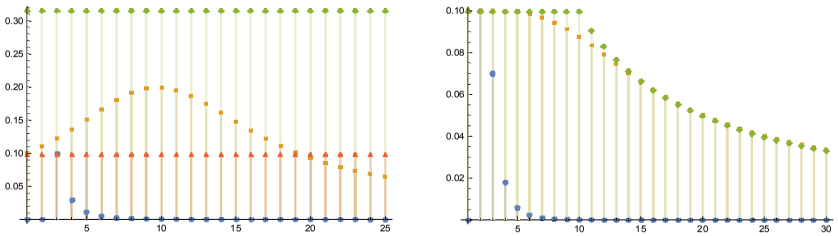
3 Bounds on IPMs and comparison of generators

As described in the introduction of the chapter, one of the purposes of the material of Chapter 3 is to provide quantitative bounds on a distance between an approximating distribution X_n, say, and a target distribution, X_∞. The following very general bound is easily seen to hold.



(a) (b) (c)

Figure 6.5: Figure 6.5(a) gives the non-uniform bound (6.46) (orange curve) as well as the classic bound $1 \wedge \sqrt{2/(e\lambda)}$ (blue curve). Figure 6.5(b) gives the non uniform bound (6.47) (orange curve), the bound $(1-e^{-\lambda})/\lambda$ (blue curve) and $2/\lambda$ (green curve). Figure 6.5(c) gives the non uniform bound (6.48) (orange curve) and the bound $\min(1, 8/(3\sqrt{2e\lambda}))$ (blue curve). All cases correspond to the Poisson distribution of parameter $\lambda = 10$.



(a) (b)

Figure 6.6: Figure 6.6(a) gives the numerical exact value of the function $|g_\xi|$ (blue curve), the bound (6.46) (orange curve), the bound (6.49) (red curve) and $\min(1, 1/|\sqrt{\lambda})$ (green curve). Figure 6.6(b) gives the numerical exact value of the function $|\Delta^+ g_\xi|$ (blue curve), the first bound in (6.50) (orange curve) and the second one (green curve). All cases correspond to the Poisson distribution with parameter $\lambda = 10$ at value $\xi = 2$.

Theorem 6.3.1 (Stein discrepancies). *Let $X_n \sim p_n$ be some random variable and let X_∞ have canonical Stein operators $\mathcal{T}_\infty^{\ell_\infty}$ and $\mathcal{L}_\infty^{\ell_\infty}$ for some $\ell_\infty \in \{-1, 0, 1\}$. Then, for all $\eta \in L^1(p_n)$, $c_1 \in \mathrm{dom}(\mathcal{T}_\infty^{\ell})$ and $h \in L^1(p_\infty) \cap L^1(p_n)$ we have*

$$\begin{aligned}
&\mathbb{E}h(X_n) - \mathbb{E}h(X_\infty) \\
&= \mathbb{E}\left[(\eta_1(X_n) - \mathbb{E}[\eta_1(X_\infty)])\frac{\mathcal{L}_\infty^{\ell_\infty} h(X_n + \ell_\infty)}{\mathcal{L}_\infty^{\ell_\infty}\eta(X_n + \ell_\infty)}\right] && (6.51)\\
&\quad + \mathbb{E}\left[\mathcal{L}_\infty^{\ell_\infty}\eta_1(X_n)\Delta^{-\ell_\infty}\left(\frac{\mathcal{L}_\infty^{\ell_\infty} h(\cdot + \ell_\infty)}{\mathcal{L}_\infty^{\ell_\infty}\eta(\cdot + \ell_\infty)}\right)(X_n)\right] && (6.52)\\
&= \mathbb{E}\left[\big(\mathcal{T}_\infty^{\ell_\infty} c_1(X_n)\big)\frac{\mathcal{L}_\infty^{\ell_\infty} h(X_n + \ell_\infty)}{c_1(X_n + \ell_\infty)}\right] && (6.53)\\
&\quad + \mathbb{E}\left[c_1(X_n)\Delta^{-\ell_\infty}\left(\frac{\mathcal{L}_\infty^{\ell_\infty} h(\cdot + \ell_\infty)}{c_1(\cdot + \ell_\infty)}\right)(X_n)\right]. && (6.54)
\end{aligned}$$

In particular the IPMs (6.5) *can be written as suprema of either of the above.*

There are many ways to exploit Theorem 6.3.1 and we once again refer to the now abundant literature on the topic for whomever needs some convincing. In this section we compare X_n with X_∞ under the additional assumption that both have an accessible Stein operators; for convenience we also impose $\ell_n = \ell_\infty = \ell$. The first step is to associate to X_n its Stein operators $\mathcal{T}_n^\ell$ and $\mathcal{L}_n^\ell$. Then we can withdraw 0 in identities such as (6.51) and (6.53) to obtain

$$\begin{aligned}
\mathbb{E}h(X_n) - \mathbb{E}h(X_\infty) &= \mathbb{E}[(\eta_1(X_n) - \eta_2(X_n))g_h(X_n)] && (6.55)\\
&\quad + \mathbb{E}\left[\big(\mathcal{L}_\infty^\ell\eta_1(X_n) - \mathcal{L}_n^\ell\eta_2(X_n)\big)\Delta^{-\ell}g_h(X_n)\right] + \kappa_{\eta_2}^\ell(h)\\
&= \mathbb{E}[\big(\mathcal{T}_\infty^\ell c_1(X_n) - \mathcal{T}_n^\ell c_2(X_n)\big)g_h^*(X_n)] && (6.56)\\
&\quad + \mathbb{E}[\big(c_1(X_n) - c_2(X_n)\big)\Delta^{-\ell}g_h^*(X_n)] + \kappa_{c_2}^{*\ell}(h)
\end{aligned}$$

with

$$\begin{aligned}
\kappa_{\eta_2}^\ell(h) &:= \mathbb{E}\left[\mathcal{T}_n^\ell(\mathcal{L}_n^\ell\eta_2(\cdot)g_h(\cdot - \ell))(X_n)\right] + (\mathbb{E}[\eta_2(X_n)] - \mathbb{E}[\eta_1(X_\infty)])\mathbb{E}[g_h(X_n)]\\
\kappa_{c_2}^{*\ell}(h) &:= \mathbb{E}\left[\mathcal{T}_n^\ell(c_2(\cdot)g_h^*(\cdot - \ell))(X_n)\right]
\end{aligned}$$

and where the choice of c_1, c_2, η_1 and η_2 are left free up to validation of easily verified technical conditions. If $\mathcal{F}(\mathcal{A}_n^{\ell,\eta_2})$ contains g_h and $\mathbb{E}[\eta_1(X_\infty)] = \mathbb{E}[\eta_2(X_n)]$, then $\kappa_{\eta_2}^\ell(h) = 0$. Similarly, if $\mathcal{F}(\mathcal{A}_n^{\ell,c_2})$ contains g_h^*, then $\kappa_{c_2}^{*\ell}(h) = 0$. In all cases, if the approximation problem is reasonable, these remainder terms should be small. Particularizing to the choice $c_1 = c_2 = 1$ and $\eta_1 = \eta_2 = \mathrm{Id}$, we obtain one of the main results of the chapter (proof in the Appendix).

Theorem 6.3.2. *Suppose that $X_n \sim p_n$ and $X_\infty \sim p_\infty$ are absolutely continuous w.r.t. the same dominating measure. For all $h \in L^1(p_\infty) \cap L^1(p_n)$ we have*

$$\mathbb{E}h(X_n) - \mathbb{E}h(X_\infty) = \mathbb{E}\left[\left(\rho_\infty^\ell(X_n) - \rho_n^\ell(X_n)\right)\mathcal{L}_\infty^\ell h(X_n + \ell)\right] + \kappa_1^{*\ell}(h) \tag{6.57}$$

with

$$\kappa_1^{\star\ell}(h) = \mathbb{E}\left[\mathcal{T}_n^\ell \mathcal{L}_\infty^\ell h(X_n)\right].$$

Furthermore, if $\eta = \mathrm{Id} \in L^1(p_\infty)$, setting $\mu_n = \mathbb{E}[X_n]$ and $\mu_\infty = \mathbb{E}[X_\infty]$ we get

$$\begin{aligned}&\mathbb{E}h(X_n) - \mathbb{E}h(X_\infty)\\ &\quad= \mathbb{E}\left[\left(\tau_n^\ell(X_n) - \tau_\infty^\ell(X_n)\right)\Delta^{-\ell}\left(\frac{-\mathcal{L}_\infty^\ell h(\cdot + \ell)}{\tau_\infty^\ell(\cdot + \ell)}\right)(X_n)\right] + \kappa_{\mathrm{Id}}^\ell(h)\end{aligned} \tag{6.58}$$

with

$$\kappa_{\mathrm{Id}}^\ell(h) = \mathbb{E}\left[\mathcal{T}_n^\ell\left(\frac{\tau_n^\ell(\cdot)}{\tau_\infty^\ell(\cdot)}\mathcal{L}_\infty^\ell h(\cdot)\right)(X_n)\right] + (\mu_n - \mu_\infty)\mathbb{E}\left[\frac{-\mathcal{L}_\infty^\ell h(X_n + \ell)}{\tau_\infty^\ell(X_n + \ell)}\right].$$

Clearly, expressions such as those in Theorem 6.3.1 and 6.3.2 will only be useful if the different functions involved are tractable. We show hereafter that this is the case.

Disclaimer: It is immediate to extend the scope of Theorem 6.3.2 to the comparison of *any* arbitrary distributions without requiring that they share a common dominating measure. Such has already been attempted successfully in Goldstein and Reinert (2013). We do not pursue this here as it would make notations very cumbersome.

We now specialize Theorem 6.3.2 to various situations of interest, that is for Kolmogorov, total variation and Wasserstein metrics, with the added assumption that both the target and the approximating laws are absolutely continuous with respect to the same dominating measure. This is in no way necessary but provides many simplifications; in particular, setting $A_n^\infty = \{x \,|\, p_n(x) \geq p_\infty(x)\}$ and $h_{\mathrm{TV}}(x) = \mathbb{I}_{A_n^\infty}(x) - \mathbb{I}_{(A_n^\infty)^c}(x) = 2\mathbb{I}_{A_n^\infty}(x) - 1$, we reap

$$\begin{aligned}\mathrm{TV}(X_n, X_\infty) &= \sup_B |P_n(B) - P_\infty(B)| = \frac{1}{2}\int |p_n(x) - p_\infty(x)|\,\mu(\mathrm{d}x)\\ &= \frac{1}{2}(\mathbb{E}h_{\mathrm{TV}}(X_n) - \mathbb{E}h_{\mathrm{TV}}(X_\infty)) = \mathbb{E}[\mathbb{I}_{A_n^\infty}(X_n)] - \mathbb{E}[\mathbb{I}_{A_n^\infty}(X_\infty)]\end{aligned}$$

(here and throughout we write $P(B) = \mathbb{E}[\mathbb{I}_B(X)]$ if X has cdf P). Although the set A_n^∞ is intractable, this last rewriting allows to avoid having a supremum in our Stein discrepancy (we work with a single indicator function) and thus leads to improved bounds.

Corollary 6.3.3 (Identity (6.57), score functions and $\ell = 0$)**.** *Suppose that the laws of X_n and X_∞ are absolutely continuous with respect to the Lebesgue measure with densities p_n and p_∞, respectively. Let $\mathcal{S}_n$ (resp., S_∞) be the support of p_n (resp., p_∞); also let $b_n = \sup \mathcal{S}_n$ and $a_n = \inf \mathcal{S}_n$ (resp., $b_\infty = \sup \mathcal{S}_\infty$ and $a_\infty = \inf \mathcal{S}_\infty$). Finally, let $\rho_n(x)$ and $\rho_\infty(x)$ be the scores and $\tau_n(x)$ and $\tau_\infty(x)$ be the Stein kernels of p_n and p_∞.*

1. *The Kolmogorov distance between the random variables X_n and X_∞ is*

$$\begin{aligned}&\mathrm{Kol}(X_n, X_\infty)\\ &= \sup_z \left| \mathbb{E}\left[(\rho_\infty(X_n) - \rho_n(X_n)) \frac{P_\infty(X_n \wedge z)\bar{P}_\infty(X_n \vee z)}{p_\infty(X_n)} \mathbb{I}_{\mathcal{S}_\infty}(X_n) \right] + \kappa_1^\star(z) \right| \qquad (6.59)\\ &\leq \mathbb{E}\left[|\rho_\infty(X_n) - \rho_n(X_n)| \frac{P_\infty(X_n)\bar{P}_\infty(X_n)}{p_\infty(X_n)} \mathbb{I}_{\mathcal{S}_\infty}(X_n) \right] + \sup_z \kappa_1^\star(z) \qquad (6.60)\end{aligned}$$

where

$$\begin{aligned}\kappa_1^\star(z) = &\lim_{x \nearrow b_n \wedge b_\infty} \frac{p_n(x)}{p_\infty(x)} P_\infty(x \wedge z)\bar{P}_\infty(x \vee z)\\ &- \lim_{x \searrow a_n \vee a_\infty} \frac{p_n(x)}{p_\infty(x)} P_\infty(x \wedge z)\bar{P}_\infty(x \vee z).\end{aligned}$$

2. *The total variation distance between X_n and X_∞ is*

$$\begin{aligned}&\mathrm{TV}(X_n, X_\infty) = \kappa_1^\star(\mathbb{I}_{A_n^\infty})\\ &+ \mathbb{E}\left[(\rho_\infty(X_n) - \rho_n(X_n)) \frac{P_\infty(A_n^\infty \cap (-\infty, X_n]) - P_\infty(A_n^\infty)P_\infty(X_n)}{p_\infty(X_n)} \mathbb{I}_{\mathcal{S}_\infty}(X_n) \right] \qquad (6.61)\\ &\leq \mathbb{E}\left[|\rho_\infty(X_n) - \rho_n(X_n)| \frac{P_\infty(X_n)\bar{P}_\infty(X_n)}{p_\infty(X_n)} \mathbb{I}_{\mathcal{S}_\infty}(X_n) \right] + \kappa_1^\star(\mathbb{I}_{A_n^\infty}) \qquad (6.62)\end{aligned}$$

where $A_n^\infty = \{x \,|\, p_n(x) \geq p_\infty(x)\}$, $X_1, X_2 \overset{\text{iid}}{\sim} p_\infty$, and

$$\begin{aligned}\kappa_1^\star(\mathbb{I}_{A_n^\infty}) = &\lim_{x \nearrow b_n \wedge b_\infty} \frac{p_n(x)}{p_\infty(x)} \left(P_\infty(A_n^\infty \cap (-\infty, x]) - P_\infty(A_n^\infty)P_\infty(x) \right)\\ &- \lim_{x \searrow a_n \vee a_\infty} \frac{p_n(x)}{p_\infty(x)} \left(P_\infty(A_n^\infty \cap (-\infty, x]) - P_\infty(A_n^\infty)P_\infty(x) \right).\end{aligned}$$

3. The Wasserstein distance between X_n and X_∞ is

$$\begin{aligned}
&\mathrm{Wass}(X_n, X_\infty) \\
&= \sup_{h\in \mathrm{Lip}(1)} \left| \mathbb{E}\left[\left(\rho_n(X_n) - \rho_\infty(X_n)\right) h'(X_\infty) \tilde{K}_\infty(X_\infty, X_n) \mathbb{I}_{\mathcal{S}_\infty}(X_n) \right] + \kappa_1^\star(h) \right| \qquad (6.63) \\
&\leq \mathbb{E}\left[\left|\rho_n(X_n) - \rho_\infty(X_n)\right| \tau_\infty(X_n) \mathbb{I}_{\mathcal{S}_\infty}(X_n) \right] + \sup_{h\in \mathrm{Lip}(1)} \kappa_1^\star(h) \qquad (6.64)
\end{aligned}$$

where

$$\begin{aligned}
\kappa_1^\star(h) = &\lim_{x \searrow a_n \vee a_\infty} \frac{p_n(x)}{p_\infty(x)} \int_{a_\infty}^{b_\infty} h'(u) P_\infty(x \wedge u) \bar{P}_\infty(x \vee u) \mathrm{d}u \\
&- \lim_{x \nearrow b_n \wedge b_\infty} \frac{p_n(x)}{p_\infty(x)} \int_{a_\infty}^{b_\infty} h'(u) P_\infty(x \wedge u) \bar{P}_\infty(x \vee u) \mathrm{d}u
\end{aligned}$$

Corollary 6.3.4 (Identity (6.58), Stein kernels and $\ell = 0$)**.** *Under the same assumptions and with exactly the same notations as in Corollary 6.3.3, the following results hold true.*

1. The Kolmogorov distance between the random variables X_n and X_∞ is

$$\begin{aligned}
&\mathrm{Kol}(X_n, X_\infty) \\
&= \sup_z \left| \mathbb{E}\left[\frac{\tau_n(X_n) - \tau_\infty(X_n)}{\tau_\infty(X_n)} \mathbb{I}_{\mathcal{S}_\infty}(X_n) \times \right. \right. \qquad (6.65) \\
&\qquad \left. \left. \left(P_\infty(z) - \mathbb{I}[X_n \leq z] + \frac{X_n - \mathbb{E}[X_\infty]}{\tau_\infty(X_n)} \frac{P_\infty(X_n \wedge z) \bar{P}_\infty(X_n \vee z)}{p_\infty(X_n)} \right) \right] + \kappa_{\mathrm{Id}}(z) \right| \\
&\leq \mathbb{E}\left[\left| \frac{\tau_n(X_n)}{\tau_\infty(X_n)} - 1 \right| \left(1 + \frac{|X_n - \mathbb{E}[X_\infty]|}{\tau_\infty(X_n)} \frac{P_\infty(X_n) \bar{P}_\infty(X_n)}{p_\infty(X_n)} \right) \mathbb{I}_{\mathcal{S}_\infty}(X_n) \right] \\
&\qquad + \sup_z |\kappa_{\mathrm{Id}}(z)| \qquad (6.66)
\end{aligned}$$

where

$$\begin{aligned}
\kappa_{\mathrm{Id}}(z) = &(\mu_n - \mu_\infty) \mathbb{E}\left[\frac{P_\infty(X_n \wedge z) \bar{P}_\infty(X_n \vee z)}{\tau_\infty(X_n) p_\infty(X_n)} \right] \\
&+ \lim_{x \nearrow b_n \wedge b_\infty} \frac{\tau_n(x)}{\tau_\infty(x)} \frac{p_n(x)}{p_\infty(x)} P_\infty(x \wedge z) \bar{P}_\infty(x \vee z) \\
&- \lim_{x \searrow a_n \vee a_\infty} \frac{\tau_n(x)}{\tau_\infty(x)} \frac{p_n(x)}{p_\infty(x)} P_\infty(x \wedge z) \bar{P}_\infty(x \vee z).
\end{aligned}$$

2. The total variation distance between X_n and X_∞ is

$$\begin{aligned}
&\mathrm{TV}(X_n, X_\infty) \\
&= \kappa_{\mathrm{Id}}(\mathbb{I}_{A_n^\infty}) + \mathbb{E}\left[\frac{\tau_n(X_n) - \tau_\infty(X_n)}{\tau_\infty(X_n)} \mathbb{I}_{\mathcal{S}_\infty}(X_n) \left(P_\infty(A_n^\infty) - \mathbb{I}_{A_n^\infty}(X_n) \right.\right. \qquad (6.67)\\
&\qquad \left.\left. + \frac{X_n - \mathbb{E}[X_\infty]}{\tau_\infty(X_n)} \frac{P_\infty(A_n^\infty \cap (-\infty, X_n]) - P_\infty(A_n^\infty) P_\infty(X_n)}{p_\infty(X_n)} \right)\right] \\
&\leq \mathbb{E}\left[\left|\frac{\tau_n(X_n)}{\tau_\infty(X_n)} - 1\right| \left(1 + \frac{|X_n - \mathbb{E}[X_\infty]|}{\tau_\infty(X_n)} \frac{P_\infty(X_n)\bar{P}_\infty(X_n)}{p_\infty(X_n)}\right) \mathbb{I}_{\mathcal{S}_\infty}(X_n)\right] \\
&\qquad + \kappa_{\mathrm{Id}}(\mathbb{I}_{A_n^\infty}) \qquad (6.68)
\end{aligned}$$

with

$$\begin{aligned}
\kappa_{\mathrm{Id}}(\mathbb{I}_{A_n^\infty}) = &\lim_{x \nearrow b_n \wedge b_\infty} \frac{\tau_n(x)}{\tau_\infty(x)} \frac{p_n(x)}{p_\infty(x)} \left(P_\infty(A_n^\infty \cap (-\infty, x]) - P_\infty(A_n^\infty) P_\infty(x)\right) \\
&- \lim_{x \searrow a_n \vee b_n} \frac{\tau_n(x)}{\tau_\infty(x)} \frac{p_n(x)}{p_\infty(x)} \left(P_\infty(A_n^\infty \cap (-\infty, x]) - P_\infty(A_n^\infty) P_\infty(x)\right) \\
&+ (\mu_n - \mu_\infty)\mathbb{E}\left[\frac{P_\infty(A_n^\infty \cap (-\infty, X_n]) - P_\infty(A_n^\infty) P_\infty(X_n)}{\tau_\infty(X_n) p_\infty(X_n)}\right].
\end{aligned}$$

3. The Wasserstein distance between X_n and X_∞ is

$$\begin{aligned}
\mathrm{Wass}(X_n, X_\infty) = &\sup_{h \in \mathrm{Lip}(1)} \left| \kappa_{\mathrm{Id}}(h) + \mathbb{E}\left[\frac{\tau_n(X_n) - \tau_\infty(X_n)}{\tau_\infty(X_n)} h'(X_\infty) \times \right.\right. \\
&\left.\left. \left(R_\infty(X_n, X_\infty) + \frac{X_n - \mathbb{E}[X_\infty]}{\tau_\infty(X_n)} \tilde{K}_\infty(X_n, X_\infty)\right) \mathbb{I}_{\mathcal{S}_\infty}(X_n)\right]\right| \qquad (6.69)\\
&\leq 2\mathbb{E}\left[\left|\frac{\tau_n(X_n)}{\tau_\infty(X_n)} - 1\right| |X_n - \mathbb{E}[X_\infty]| \mathbb{I}_{\mathcal{S}_\infty}(X_n)\right] + \sup_{h \in \mathrm{Lip}(1)} \kappa_{\mathrm{Id}}(h) \qquad (6.70)
\end{aligned}$$

where

$$\begin{aligned}
\kappa_{\mathrm{Id}}(h) = &\lim_{x \searrow a_n \vee a_\infty} \frac{\tau_n(x)}{\tau_\infty(x)} \frac{p_n(x)}{p_\infty(x)} \int_{a_\infty}^{b_\infty} h'(u) P_\infty(x \wedge u) \bar{P}_\infty(x \vee u) \mathrm{d}u \\
&- \lim_{x \nearrow b_n \wedge b_\infty} \frac{\tau_n(x)}{\tau_\infty(x)} \frac{p_n(x)}{p_\infty(x)} \int_{a_\infty}^{b_\infty} h'(u) P_\infty(x \wedge u) \bar{P}_\infty(x \vee u) \mathrm{d}u \\
&+ (\mu_n - \mu_\infty)\mathbb{E}\left[\frac{h'(X_\infty)}{\tau_\infty(X_n)} \left(R_\infty(X_n, X_\infty) + \frac{X_n - \mathbb{E}[X_\infty]}{\tau_\infty(X_n)} \tilde{K}_\infty(X_n, X_\infty)\right)\right].
\end{aligned}$$

Corollary 6.3.5 (Identity (6.57), score functions, $\ell = \pm 1$)**.** *Suppose that the laws of X_n and X_∞ are discrete with mass functions p_n on support $\mathcal{S}_n$ and p_∞ on $\mathcal{S}_\infty$ respectively. Let $a_n = \inf \mathcal{S}_n$ and $b_n = \sup \mathcal{S}_n$ (resp. $a_\infty = \inf \mathcal{S}_\infty$ and $b_n = \sup \mathcal{S}_\infty$). Finally, let $\rho_n^\ell(x)$ and $\rho_\infty^\ell(x)$ be the scores and $\tau_n^\ell(x)$ and $\tau_\infty^\ell(x)$ be the Stein kernels of p_n and p_∞. The following results hold true.*

$$\begin{aligned}
\mathrm{TV}(X_n, X_\infty) &= \kappa_1^{\star\ell}(\mathbb{I}_{A_n^\infty}) + \mathbb{E}\Bigg[\left(\rho_\infty^\ell(X_n) - \rho_n^\ell(X_n)\right) \mathbb{I}_{\mathcal{S}_\infty}(X_n + \ell) \times \\
&\qquad\qquad \frac{P_\infty(A_n^\infty \cap (-\infty, X_n - b_\ell]) - P_\infty(A_n^\infty) P_\infty(X_n - b_\ell)}{p_\infty(X_n + \ell)} \Bigg] \\
&\leq \mathbb{E}\left[\left|\rho_\infty^\ell(X_n) - \rho_n^\ell(X_n)\right| \frac{P_\infty(X_n - b_\ell) \bar{P}_\infty(X_n - b_\ell)}{p_\infty(X_n + \ell)} \mathbb{I}_{\mathcal{S}_\infty}(X_n + \ell) \right] + \kappa_1^{\star\ell}(\mathbb{I}_{A_n^\infty})
\end{aligned}$$

with

$$\begin{aligned}
\kappa_1^{\star +}(\mathbb{I}_{A_n^\infty}) &= - \lim_{x \searrow a_n \vee a_\infty} \frac{p_n(x)}{p_\infty(x)} \left(P_\infty(A_n^\infty \cap (-\infty, x-1]) - P_\infty(A_n^\infty) P_\infty(x-1)\right) \\
\kappa_1^{\star -}(\mathbb{I}_{A_n^\infty}) &= \lim_{x \nearrow b_n \wedge b_\infty} \frac{p_n(x)}{p_\infty(x)} \left(P_\infty(A_n^\infty \cap (-\infty, x]) - P_\infty(A_n^\infty) P_\infty(x)\right).
\end{aligned}$$

Corollary 6.3.6 (Identity (6.58), Stein kernels, $\ell = \pm 1$)**.** *Under the same assumptions and with exactly the same notations as in Corollary 6.3.5, the following results hold true.*

$$\begin{aligned}
&\mathrm{TV}(X_n, X_\infty) \\
&= \kappa_{\mathrm{Id}}^\ell(\mathbb{I}_{A_n^\infty}) + \mathbb{E}\Bigg[\frac{\tau_n^\ell(X_n) - \tau_\infty^\ell(X_n)}{\tau_\infty^\ell(X_n)} \mathbb{I}_{\mathcal{S}_\infty}(X_n + \ell) \times \Bigg(P_\infty(A_n^\infty) - \mathbb{I}_{A_n^\infty}(X_n) \\
&\qquad + \frac{X_n - \mathbb{E}[X_\infty]}{\tau_\infty^\ell(X_n + \ell)} \frac{P_\infty(A_n^\infty \cap (-\infty, X_n - b_\ell]) - P_\infty(A_n^\infty) P_\infty(X_n - b_\ell)}{p_\infty(X_n + \ell)} \Bigg) \Bigg] \\
&\leq \mathbb{E}\left[\left| \frac{\tau_n^\ell(X_n)}{\tau_\infty^\ell(X_n)} - 1 \right| \left(1 + \frac{|X_n - \mathbb{E}[X_\infty]|}{\tau_\infty^\ell(X_n + \ell)} \frac{P_\infty(X_n - b_\ell) \bar{P}_\infty(X_n - b_\ell)}{p_\infty(X_n + \ell)} \right) \mathbb{I}_{\mathcal{S}_\infty}(X_n + \ell) \right] \\
&\qquad + \kappa_{\mathrm{Id}}^\ell(\mathbb{I}_{A_n^\infty})
\end{aligned}$$

with

$$\begin{aligned}
&\kappa_{\mathrm{Id}}^+(\mathbb{I}_{A_n^\infty}) \\
&= - \lim_{x \searrow a_n \vee a_\infty} \frac{\tau_n^+(x)}{\tau_\infty^+(x)} \frac{p_n(x)}{p_\infty(x)} \left(P_\infty(A_n^\infty \cap (-\infty, x-1]) - P_\infty(A_n^\infty) P_\infty(x-1)\right) \\
&\qquad + (\mu_\infty - \mu_n) \mathbb{E}\left[\frac{P_\infty(A_n^\infty \cap (-\infty, X_n]) - P_\infty(A_n^\infty) P_\infty(X_n)}{\tau_\infty^+(X_n + 1) p_\infty(X_n + 1)} \mathbb{I}_{\mathcal{S}_\infty}(X_n + 1) \right]
\end{aligned}$$

and

$$\kappa^-_{\mathrm{Id}}(\mathbb{I}_{A_n^\infty}) = \lim_{x \nearrow b_n \wedge b_\infty} \frac{\tau_n^-(x)}{\tau_\infty^-(x)} \frac{p_n(x)}{p_\infty(x)} \left(P_\infty(A_n^\infty \cap (-\infty, x]) - P_\infty(A_n^\infty) P_\infty(x)\right) + (\mu_\infty - \mu_n)\mathbb{E}\left[\frac{P_\infty(A_n^\infty \cap (-\infty, X_n]) - P_\infty(A_n^\infty)P_\infty(X_n)}{\tau_\infty^-(X_n - 1)p_\infty(X_n - 1)} \mathbb{I}_{\mathcal{S}_\infty}(X_n - 1)\right].$$

Example 6.3.7 (Standard normal target)**.** *Let $X_\infty \sim \mathcal{N}(0,1)$ and consider the notation of example 6.2.26. The classic Stein discrepancy between any random variable X_n and X_∞ in this case is*

$$\sup_{h \in \mathcal{H}} |\mathbb{E}\left[g_h'(X_n) - X_n g_h(X_n)\right]| \tag{6.71}$$

with $g_h = \mathcal{L}_\infty h$ the unique bounded solution to the Stein equation $g_h'(x) - xg(x) = h(x) - \mathbb{E}h(X_\infty)$. Applications of (6.71) *are extremely well documented. To illustrate the power of our approach, let X_n be a continuous real random variable. By Corollaries 6.3.3 and 6.3.4 the following bounds hold.*

- Kolmogorov distance.
  Direct computations from (6.59) *yield*

$$\begin{aligned}
\mathrm{Kol}(X_n, X_\infty) &= \sup_z \left| \mathbb{E}\left[(X_n + \rho_n(X_n)) \frac{\Phi(X_n \wedge z)\bar{\Phi}(X_n \vee z)}{\varphi(X_n)}\right] - \kappa_1^\star(z) \right| \\
&\leq \mathbb{E}\left[|X_n + \rho_n(X_n)| \frac{\Phi(X_n)\bar{\Phi}(X_n)}{\varphi(X_n)}\right] + \sup_z \kappa_1^\star(z) \\
&\leq \frac{1}{2}\sqrt{\frac{\pi}{2}} \mathbb{E}\left[|X_n + \rho_n(X_n)|\right] + \sup_z \kappa_1^\star(z)
\end{aligned}$$

  and, from (6.65)*,*

$$\begin{aligned}
&\mathrm{Kol}(X_n, X_\infty) \\
&= \sup_z \left| \mathbb{E}\left[(\tau_n(X_n) - 1)\left(\Phi(z) - \mathbb{I}[X_n \leq z] + X_n \frac{\Phi(X_n \wedge z)\bar{\Phi}(X_n \vee z)}{\varphi(X_n)}\right)\right] + \kappa_{\mathrm{Id}}(z) \right| \\
&\leq \mathbb{E}\left[|\tau_n(X_n) - 1| \left(1 + |X_n| \frac{\Phi(X_n)\bar{\Phi}(X_n)}{\varphi(X_n)}\right)\right] + \sup_z |\kappa_{\mathrm{Id}}(z)| \\
&\leq 2\mathbb{E}\left[|\tau_n(X_n) - 1|\right] + \sup_z |\kappa_{\mathrm{Id}}(z)|
\end{aligned}$$

  For instance, if $X_n \sim t_n$ is Student with n degrees of freedom, then $\kappa_1^\star(z) = \kappa_{\mathrm{Id}}(z) = 0$ for all z, $\rho_n = -(1+n)x/(n+x^2)$ and $\tau_n(x) = (x^2+n)/(n-1)$ (see e.g. Table 4.3) we obtain

$$\begin{aligned}\mathrm{Kol}(X_n, X_\infty) &\leq \mathbb{E}\left[|X_n|\left|\frac{X_n^2-1}{X_n^2+n}\right|\frac{\Phi(X_n)\bar{\Phi}(X_n)}{\varphi(X_n)}\right] \\ &\leq \frac{1}{2}\sqrt{\frac{\pi}{2}}\mathbb{E}\left[|X_n|\left|\frac{X_n^2-1}{X_n^2+n}\right|\right] \leq \frac{2/\sqrt{e}-1/2}{n-1} \approx \frac{0.7130}{n-1}\end{aligned} \tag{6.72}$$

and

$$\begin{aligned}\mathrm{Kol}(X_n, X_\infty) &\leq \mathbb{E}\left[\frac{X_n^2+1}{n-1}\left(1+|X_n|\frac{\Phi(X_n)\bar{\Phi}(X_n)}{\varphi(X_n)}\right)\right] \\ &\leq 2\mathbb{E}\left[\frac{X_n^2+1}{n-1}\right] = \frac{2}{n-2}.\end{aligned}$$

Both our bounds improve e.g. on Cacoullos et al. (1994, Example 1, p1614) but does (of course) not improve on the optimal bound of Pinelis (2015, Theorem 1.2) which is of order $0.158/n$.

- Total variation distance.
  Our upper bounds (6.61) *and* (6.67) *on Total Variation distance are the same as those for the Kolmogorov distance reported above, so that we can compare directly to Duembgen et al. (2019, Lemma 9) who obtain the elegant bound* $\mathrm{TV}(X_n, X_\infty) \leq 2/n$ *in this case. Numerical evaluations of* (6.72) *show that our bound is a (slight) improvement, see Figure 6.7(a).*
- Wasserstein distance.
  Direct computations from (6.63) *yield*

$$\begin{aligned}\mathrm{Wass}(X_n, X_\infty) &= \sup_{h\in\mathrm{Lip}(1)}\left|\mathbb{E}\left[(\rho_n(X_n)+X_n)\,h'(X_\infty)\tilde{K}_\varphi(X_\infty, X_n)\right] + \kappa_1^\star(h)\right| \\ &\leq \mathbb{E}\left[|\rho_n(X_n)+X_n|\right] + \sup_{h\in\mathrm{Lip}(1)}|\kappa_1^\star(h)|.\end{aligned}$$

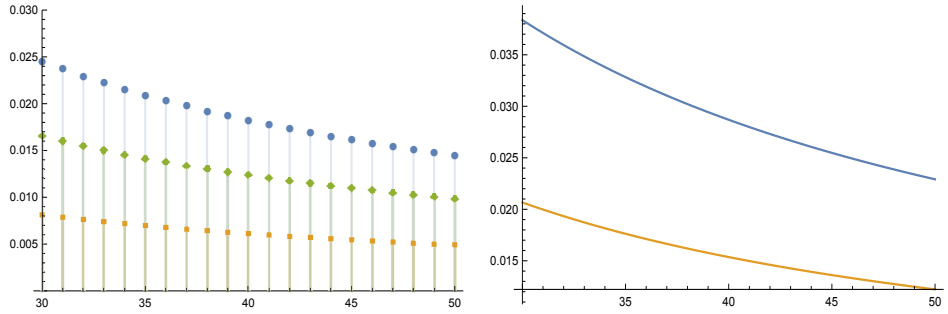
In the particular case of Student t vs standard normal, we obtain

$$\mathrm{Wass}(X_n, X_\infty) \leq \mathbb{E}\left[\left|X_n\frac{1-X_n^2}{n+X_n^2}\right|\right] \leq \frac{3}{\sqrt{2\pi}}\frac{1}{\sqrt{n-1}}.$$

The bounds obtained from (6.69) *are of the same order and not reported here.*

Example 6.3.8 (Beta vs gamma)**.** *Let* $X_B \sim \mathrm{Beta}(\alpha, \beta)$ *with density* $p_B(x) = x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha,\beta)\mathbb{I}_{[0,1]}(x)$ *and cdf* P_B*; also let* $X_G \sim \Gamma(r, s)$ *with density* $p_G(x) = x^{r-1}s^r e^{-sx}/\Gamma(r)\mathbb{I}_{[0,\infty)}(x)$ *and cdf* P_G*. Simple computations yield (see also Table 4.3) the scores and Stein kernels:*

$$\begin{aligned}\rho_B(x) &= \frac{1-\alpha+x(\alpha+\beta-2)}{x(x-1)} \text{ and } \tau_B(x) = \frac{x(1-x)}{\alpha+\beta} \\ \rho_G(x) &= \frac{r-1}{x} - s \text{ and } \tau_G = \frac{x}{s}.\end{aligned}$$



(a) (b)

Figure 6.7: Figure 6.7(a) reports bounds on the total variation distance between t_n and $\mathcal{N}(0,1)$ for $n \in [30,50]$: $2/n$ (green curve), our bound $(2/\sqrt{e}-1/2)/(n-1)$ (blue curve) and numerical evaluation of bound (6.72) (orange curve). Figure 6.7(b) provides our upper bound on the Wasserstein distance (blue curve) as well as the exact value of the Wasserstein distance (computed with the formula $\mathrm{Wass}(X_n, X_\infty) = \int_{-\infty}^{\infty} |P_n(z) - P_\infty(z)| \, \mathrm{d}z$) for the same model and range of n.

In order to facilitate comparison with Duembgen et al. (2019), we consider the same parameter settings as in that paper, namely $r = \alpha$ and $\beta > 1$. Then

$$\rho_B(x) - \rho_G(x) = s + \frac{\beta - 1}{x - 1} \text{ and } \tau_B(x) - \tau_G(x) = x\left(\frac{1-x}{\alpha+\beta} - \frac{1}{s}\right).$$

We apply Corollary 6.57 to obtain

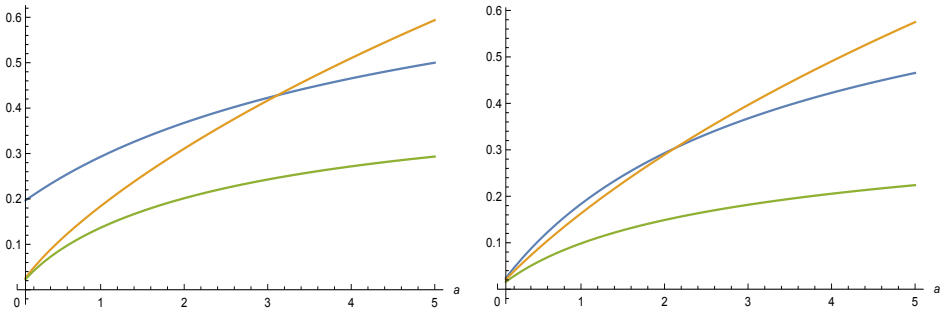
$$\mathrm{TV}(X_B, X_G) \leq \mathbb{E}\left[\left|s + \frac{\beta - 1}{X_B - 1}\right| \frac{P_G(X_B)\bar{P}_G(X_B)}{p_G(X_B)}\right] \tag{6.73}$$

(here we use $\Gamma(\alpha, s)$ as target, i.e. $X_B = X_n$ and $X_G = X_\infty$; $\kappa_1^\star(\mathbb{I}_{A_n^\infty}) = 0$) and

$$\mathrm{TV}(X_G, X_B) \leq \mathbb{E}\left[\left|s + \frac{\beta - 1}{X_G - 1}\right| \frac{P_B(X_G)\bar{P}_B(X_G)}{p_B(X_G)} \mathbb{I}[X_G \in [0,1]]\right] \tag{6.74}$$

(here we use $\mathrm{Beta}(\alpha, \beta)$ as target, i.e. $X_B = X_\infty$ and $X_G = X_n$; $\kappa_1^\star(\mathbb{I}_{A_n^\infty}) = 0$). Numerical evaluations show that our bounds seem to outperform those Duembgen et al. (2019) (see Figure 6.8). More effort needs to be put in the study of the behaviour of the ratio $P_\infty(x)\bar{P}_\infty(x)/p_\infty(x)$. We do not report the corresponding bounds on the total variation distance that can be obtained from Corollary 6.58; we do not either compute the bounds on Kolmogorov or Wasserstein distance.

Example 6.3.9 (Poisson target)**.** *Let $X_\infty \sim \mathrm{Pois}(\lambda)$ and consider the notation of example 6.2.28. The classic Stein discrepancy between any random variable X_n and*



(a) (b)

Figure 6.8: Bounds on $\mathrm{TV}(X_B, X_G)$ obtained through (6.73) (orange curve), (6.74) (green curve) and Duembgen et al. (2019) (blue curve), with $X_B \sim \mathrm{Beta}(a, 3)$ vs $X_G \sim \Gamma(a, a+3)$ (Figure 6.8(a)) and $X_B \sim \mathrm{Beta}(a, 3)$ vs $X_G \sim \Gamma(a, a+2)$ (Figure 6.8(b)).

X_∞ in this case is

$$\sup_{h \in \mathcal{H}} |\mathbb{E}\left[\lambda g_h(X_n + 1) - X_n g_h(X_n)\right]| \tag{6.75}$$

with $g_h(x) = \mathcal{L}_\infty^- h(x-1)$ the unique bounded solution to the Stein equation $\lambda g_h(x+1) - x g(x) = h(x) - \mathbb{E}h(X_\infty)$. Applications of (6.75) are extremely well documented. To illustrate the power of our approach, let X_n be a discrete real random variable with values in $\mathbb{N}$. By Corollaries 6.3.5 and 6.3.6, we get that $\mathrm{TV}(X_n, X_\infty)$ is bounded from above by the following four quantities:

$$B_1(\lambda, X_n) = \mathbb{E}\left[\left|\frac{\lambda}{X_n+1} - 1 - \rho_n^+(X_n)\right| \frac{P_\infty(X_n)\bar{P}_\infty(X_n)}{p_\infty(X_n+1)}\right] + \kappa_1^{\star +}(\mathbb{I}_{A_n^\infty})$$

$$B_2(\lambda, X_n) = \mathbb{E}\left[\left|1 - \frac{X_n}{\lambda} - \rho_n^-(X_n)\right| \frac{P_\infty(X_n-1)\bar{P}_\infty(X_n-1)}{p_\infty(X_n-1)} \mathbb{I}[X_n > 0]\right] + \kappa_1^{\star -}(\mathbb{I}_{A_n^\infty})$$

$$B_3(\lambda, X_n) = \mathbb{E}\left[\left|\frac{\tau_n^+(X_n)}{X_n} - 1\right| \left(1 + \frac{|X_n - \lambda|}{X_n+1} \frac{P_\infty(X_n)\bar{P}_\infty(X_n)}{p_\infty(X_n+1)}\right)\right] + \kappa_{\mathrm{Id}}^+(\mathbb{I}_{A_n^\infty})$$

$$B_4(\lambda, X_n) = \mathbb{E}\left[\left|\frac{\tau_n^-(X_n)}{\lambda} - 1\right| \left(1 + \frac{|X_n - \lambda|}{\lambda} \frac{P_\infty(X_n-1)\bar{P}_\infty(X_n-1)}{p_\infty(X_n-1)}\right) \mathbb{I}[X_n > 0]\right] + \kappa_{\mathrm{Id}}^-(\mathbb{I}_{A_n^\infty})$$

We illustrate the bounds on some easy examples.

Example 6.3.10 (Poisson vs Poisson)**.** *If $X_n \sim \mathrm{Pois}(\lambda_n)$ then $\kappa_1^{\star +}(\mathbb{I}_{A_n^\infty}) = 0$ and $\kappa_1^{\star -}(\mathbb{I}_{A_n^\infty}) = 0$ so that*

$$B_1(\lambda, \lambda_n) = |\lambda - \lambda_n| \mathbb{E}\left[\frac{1}{X_n+1} \frac{P_\infty(X_n)\bar{P}_\infty(X_n)}{p_\infty(X_n+1)}\right] \leq |\lambda - \lambda_n| \frac{\lambda}{\lambda_n}$$

$$B_2(\lambda, \lambda_n) = \left| \frac{1}{\lambda} - \frac{1}{\lambda_n} \right| \mathbb{E}\left[X_n \frac{P_\infty(X_n - 1)\bar{P}_\infty(X_n - 1)}{p_\infty(X_n - 1)} \mathbb{I}[X_n > 0] \right] \leq |\lambda - \lambda_n|.$$

Similar arguments apply for B_3 and B_4 yielding similar results that are not reported here (although it is interesting to note that the first term in B_3 cancels out, and the only non zero term arises through non equality of the means).

Example 6.3.11 (Poisson vs binomial)**.** *If $X_n \sim \text{Bin}(n, \theta)$ and $X_\infty \sim \text{Pois}(n\theta)$ then $\kappa_1^{\star+}(\mathbb{I}_{A_n^\infty}) = 0$ and $\kappa_1^{\star-}(\mathbb{I}_{A_n^\infty}) \leq \sqrt{2\pi} n^{1/2} e^{-n(1-\theta)}$ which is negligible for all values of $\theta \in (0, 1)$. Moreover*

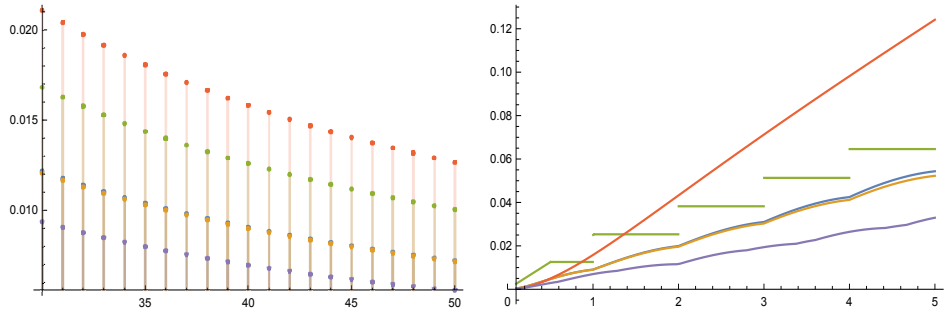
$$\rho_n^+(x) = \frac{\theta}{1-\theta} \frac{n-x}{x+1} - 1 \text{ and } \rho_n^-(x) = 1 - \frac{1-\theta}{\theta} \frac{x}{n-x+1}$$

so that

$$B_1(\lambda, n, \theta) = \mathbb{E}\left[\frac{\theta}{1-\theta} \frac{|X_n - n\theta|}{X_n + 1} \frac{P_\infty(X_n)\bar{P}_\infty(X_n)}{p_\infty(X_n + 1)} \right]$$

$$B_2(\lambda, n, \theta) = \mathbb{E}\left[X_n \frac{|X_n - 1 - n\theta|}{n\theta(n - X_n + 1)} \frac{P_\infty(X_n - 1)\bar{P}_\infty(X_n - 1)}{p_\infty(X_n - 1)} \mathbb{I}[X_n > 0] \right] + \kappa_1^{\star-}(\mathbb{I}_{A_n^\infty})$$

We can also exchange the roles of p_n and p_∞ and compute the same bounds with respect to the Poisson target. Numerical evaluations are reported in Figure 6.9.



(a) (b)

Figure 6.9: Exact value of $\text{TV}(\text{Bin}(n, \lambda/n), \text{Pois}(\lambda))$ (purple curve), bound $B_1(\lambda, n, \lambda/n)$ (blue curve), the same bound when the roles of X_n and X_∞ are reversed (orange curve), the bound $(\lambda/n) \wedge (1 - \sqrt{1 - \lceil \lambda \rceil / n})$ from Duembgen et al. (2019) (green curve) and Chen's classical bound $\lambda(1 - e^{-\lambda})/n$ from Chen (1975) (red curve). Left plot for $\lambda = 1$ and $n \in [30, 50]$; right plot for $n = 40$ and $\lambda \in (0, 5)$.

A Some more proofs

Proof of Lemma 6.2.19. We can use (3.20) to obtain

$$\begin{aligned}\bar{h}(x) &= \mathbb{E}\left[(h(x)-h(X))(\chi^{\ell}(X,x)+\chi^{-\ell}(x,X))\right]\\ &= \mathbb{E}\left[\Delta^{-\ell}h(X_2)\mathbb{E}\left[\Phi_p^{\ell}(X,X_2,x)-\Phi_p^{\ell}(x,X_2,X)|X_2\right]\right]\end{aligned}$$

(we use the fact that $\chi^{\ell}(x,y)+\chi^{-\ell}(y,x) = 1+\mathbb{I}[\ell=0]\mathbb{I}[x=y]$) and it only remains to reorganize the integrand to obtain the claim. To this end we note how, by definition,

$$\begin{aligned}\mathbb{E}\left[\Phi_p^{\ell}(X,y,x)-\Phi_p^{\ell}(x,y,X)\right] &= \frac{\chi^{-\ell}(y,x)}{p(y)}\mathbb{E}[\chi^{\ell}(X,y)] - \frac{\chi^{\ell}(x,y)}{p(y)}\mathbb{E}[\chi^{-\ell}(y,X)]\\ &= \chi^{-\ell}(y,x)\frac{P(y-a_\ell)}{p(y)} - \chi^{\ell}(x,y)\frac{\bar{P}(y-a_\ell)}{p(y)}\end{aligned}$$

where the first identity is immediate by definition of Φ_p^{ℓ} and the last identity follows from the definition of the generalized indicator χ^{ℓ}. □

Proof of Lemma 6.2.21. The expressions (6.27) and (6.28) of the solution g are direct from the definition of $\mathcal{L}_p^{\ell}$ and its representations (3.19) and (6.25). The expressions (6.29) and (6.31) of the derivative are direct from the expression (6.14). For the claim (6.33), we shall first prove the following result:

$$\begin{aligned}&\Delta^{-\ell}g(x)\\ &\quad= \frac{\mathbb{E}\left[\left(\tilde{K}_p^{\ell}(X_1,x+\ell)R_p^{\ell}(x,X_2) - R_p^{\ell}(x,X_1)\tilde{K}_p^{\ell}(X_2,x+\ell)\right)\Delta^{-\ell}h(X_1)\Delta^{-\ell}\eta(X_2)\right]}{\big(-\mathcal{L}_p^{\ell}\eta(x)\big)\big(-\mathcal{L}_p^{\ell}\eta(x+\ell)\big)}\end{aligned}$$

Starting from (6.13) and applying repeatedly (6.25) then (6.26) (once to h and once to η) we obtain

$$\begin{aligned}&\Delta^{-\ell}g(x)\\ &\quad= \frac{\mathbb{E}\left[\tilde{K}_p^{\ell}(X_1,x+\ell)\bar{\eta}(x)\Delta^{-\ell}h(X_1)\right] - \big(-\mathcal{L}_p^{\ell}\eta(x+\ell)\big)\mathbb{E}\left[R_p^{\ell}(x,X_1)\Delta^{-\ell}h(X_1)\right]}{\big(-\mathcal{L}_p^{\ell}\eta(x+\ell)\big)\big(-\mathcal{L}_p^{\ell}\eta(x+\ell)\big)}\\ &\quad= \frac{\mathbb{E}\left[\left(\tilde{K}_p^{\ell}(X_1,x+\ell)\bar{\eta}(x) - R_p^{\ell}(x,X_1)\big(-\mathcal{L}_p^{\ell}\eta(x+\ell)\big)\right)\Delta^{-\ell}h(X_1)\right]}{\big(-\mathcal{L}_p^{\ell}\eta(x)\big)\big(-\mathcal{L}_p^{\ell}\eta(x+\ell)\big)}\\ &\quad= \frac{\mathbb{E}\left[\left(\tilde{K}_p^{\ell}(X_1,x+\ell)R_p^{\ell}(x,X_2) - R_p^{\ell}(x,X_1)\tilde{K}_p^{\ell}(X_2,x+\ell)\right)\Delta^{-\ell}h(X_1)\Delta^{-\ell}\eta(X_2)\right]}{\big(-\mathcal{L}_p^{\ell}\eta(x)\big)\big(-\mathcal{L}_p^{\ell}\eta(x+\ell)\big)}.\end{aligned}$$

To conclude, we decompose the above expectation into four parts with: $X_i < x + a_\ell$ and/or $X_i \geq x + a_\ell$, for $i = 1, 2$ (i.e., using either $\chi^{-\ell}(X_i, x)$ or $\chi^{\ell}(x, X_i)$). Therefore, by considering separately $\ell \in \{0, -1, 1\}$, we can easily verify that

$$\tilde{K}_p^\ell(y, x+\ell) = \begin{cases} \dfrac{P(y - a_\ell)\bar{P}(x + a_\ell)}{p(y)p(x+\ell)} \text{ if } y < x + a_\ell \\ \dfrac{P(x - b_\ell)\bar{P}(y + b_\ell)}{p(y)p(x+\ell)} \text{ if } y \geq x + a_\ell \end{cases}$$

and

$$R_p^\ell(x, y) = \begin{cases} \dfrac{P(y - a_\ell)}{p(y)} \text{ if } y < x + a_\ell \\ \dfrac{-\bar{P}(y + b_\ell)}{p(y)} \text{ if } y \geq x + a_\ell \end{cases}$$

Basic manipulations then give

$$\begin{aligned} &\Delta^{-\ell} g(x)\big(-\mathcal{L}_p^\ell \eta(x)\big)\big(-\mathcal{L}_p^\ell \eta(x+\ell)\big)\frac{\bar{P}(x+a_\ell) + P(x - b_\ell)}{p(x+\ell)} \times \\ &= \left(\mathbb{E}\left[\Delta^{-\ell} h(X_1)\frac{\bar{P}(X_1 + b_\ell)}{p(X_1)}\chi^\ell(x, X_1)\right]\mathbb{E}\left[\Delta^{-\ell}\eta(X_2)\frac{P(X_2 - a_\ell)}{p(X_2)}\chi^{-\ell}(X_2, x)\right] \right. \\ &\quad \left. - \mathbb{E}\left[\Delta^{-\ell} h(X_1)\frac{P(X_1 - a_\ell)}{p(X_1)}\chi^{-\ell}(X_1, x)\right]\mathbb{E}\left[\Delta^{-\ell}\eta(X_2)\frac{\bar{P}(X_2 + b_\ell)}{p(X_2)}\chi^{\ell}(x, X_2)\right] \right) \end{aligned}$$

which leads to the claim as $\bar{P}(x + a_\ell) + P(x - b_\ell) = 1$ and $\ell = a_\ell - b_\ell$. □

Proof of Lemma 6.2.22. The condition implies that g_ξ^- is non decreasing and non negative over $\mathcal{S}(p) \cap (-\infty, \xi]$ and non decreasing and non positive over $\mathcal{S}(p) \cap (\xi, \infty)$. Therefore, the absolute value of the solution for point mass equation (6.34) reaches his supremum at ξ or $\xi + 1$, which gives the bound (6.35). Moreover, the supremum of the difference is observed between ξ and $\xi + 1$. Using the explicit expression (6.23) and the relation $\tau_p^\ell(x+\ell)p(x+\ell) = \tau_p^{-\ell}(x)p(x)$, we have

$$\begin{aligned} \sup_x |\Delta g(x)| = g^-(\xi) - g^-(\xi+1) &= \frac{P(\xi - 1)}{\tau_p^+(\xi)} + \frac{(1 - P(\xi))p(\xi)}{\tau_p^+(\xi+1)p(\xi+1)} \\ &= \frac{P(\xi - 1)}{\tau_p^+(\xi)} + \frac{1 - P(\xi)}{\tau_p^-(\xi)}. \end{aligned}$$

Furthermore, as $x - \mathbb{E}[X] = \tau_p^+(x) - \tau_p^-(x)$, we have $\tau_p^-(\xi) \geq \tau_p^+(\xi)$ if $\xi \leq \mathbb{E}[X]$ (resp. $\tau_p^-(\xi) \leq \tau_p^+(\xi)$ if $\xi \geq \mathbb{E}[X]$). Therefore, the supremum is bounded by $(1 - p(\xi))/\tau_p^+(\xi)$ if $\xi \leq \mathbb{E}[X]$ and otherwise by $(1 - p(\xi))/\tau_p^-(\xi)$.

By remark 6.2.17, the solution $g_A^\ell(x)$ is explicit and defined by g_ξ^ℓ for $\xi \in A$. The sign of g_ξ^ℓ changes according to the relative position of ξ and x. Then, combined with the hypotheses, the maximal value of $|g_A^-(x)|$ is either observed at $x = \min_{\xi\in A}\{\xi\} =: \xi_1$ or $x = \max_{\xi\in A}\{\xi\} + 1 =: \xi_2 + 1$. Then,

$$\begin{aligned}\sup_x |g_A^-(x)| &= \max\left\{\frac{P(\xi_1 - 1)}{p(\xi_1)\tau_p^+(\xi_1)}\sum_{j\in A} p(j), \frac{1 - P(\xi_2)}{p(\xi_2)\tau_p^-(\xi_2)}\sum_{j\in A} p(j)\right\} \\ &\leq \left(\sum_{j\in A} p(j)\right)\sup_{\xi\in A}\left\{\frac{1}{\tau_p^+(\xi)p(\xi)}, \frac{1}{\tau_p^-(\xi)p(\xi)}\right\}.\end{aligned}$$

Finally, due to the monotonicity of each $g_\xi^\ell(x)$ function, the maximal difference $|\Delta g_A^\ell(x)|$ is bounded by the supremum of $|\Delta g_\xi^\ell(x)|$ for $\xi \in A$, which is enough to conclude. □

Proof of Theorem 6.3.2. First take $c_1(x) = c_2(x) = 1$ in (6.56). Without any further assumptions on h, the solution g_h^* of (6.7) with $c(x) = 1$ can be represented as

$$g_h^*(x) = \frac{\mathcal{L}_\infty^\ell h(x+\ell)}{c_1(x+\ell)} = \mathcal{L}_\infty^\ell h(x+\ell)$$

Hence, we obtain (6.57).

Next take $\eta_1 = \eta_2 = \mathrm{Id}$ in (6.55). Then, $-\mathcal{L}_\infty^\ell \eta_1(x) = \tau_\infty^\ell(x)$ and $-\mathcal{L}_n^\ell \eta_2(x) = \tau_n^\ell(x)$, the Stein kernels of p_∞ and p_n. Without any further assumptions on h, the solution $g_h(x)$ of (6.6) with $\eta = \mathrm{Id}$ can be represented as

$$g_h(x) = \frac{-\mathcal{L}_\infty^\ell h(x+\ell)}{\tau_\infty^\ell(x+\ell)}$$

Hence we get (6.58).

□

CHAPTER 7

General conclusions and perspectives

Some conclusions have already been outlined at the end of each chapter. To conclude the second part of the doctoral thesis, we would like to point out here some directions for future research around developments on Stein's method which can also be useful in statistics towards more applied considerations, e.g. the development of goodness-of-fit tests or estimation procedures. In Chapter 6 (Equations (6.3) and (6.4)), we defined a general Stein discrepancy

$$\mathcal{S}_\bullet(X_n, X_\infty; \mathcal{A}_\infty) = \sup_{g \in \mathcal{G}} |\mathbb{E}\left[\mathcal{A}_\infty g(X_n)\right]| \tag{7.1}$$

which serves to measure the dissimilarity between two random variables X_n and X_∞, in terms of the action of $\mathcal{A}_\infty$ over a subclass $\mathcal{G} \subseteq \mathcal{F}(\mathcal{A}_\infty)$ defined by solutions of Stein equations. There is much flexibility in the definition (7.1). Moreover, as explained in Chapter 6, we may rewrite the discrepancy itself in different ways (see Equations (6.3) and (6.4)). These rewriting are reminiscent of the generalized Fisher information distance and the Stein discrepancy (see Ley and Swan, 2013a, for further details). The idea of Liu et al. (2016) and Chwialkowski et al. (2016) is to extend the generalized Fisher information distance in order to construct a goodness of fit test for continuous distributions. To this end, they both, simultaneously and independently, defined an object called *(kernelized) Stein discrepancy* $\mathbb{S}_k(p_\infty, p_n)$ between distributions p_∞ (target distribution) and p_n (observed sample distribution) by

$$\mathbb{S}_k(p_\infty, p_n; \mathcal{A}_\infty) = \mathbb{E}\left[\mathcal{A}^1_\infty \mathcal{A}^2_\infty k(Y_1, Y_2)\right] \tag{7.2}$$

where $\mathcal{A}^i_\infty$ acts on the ith marginal of the function k, and Y_1, Y_2 are two i.i.d. random variables distributed according to p_n. The function k is the *kernel function*. If we choose the operator (6.9) with $c(x) = 1$, i.e.,

$$\mathcal{A}_\infty f(x) = \rho^\ell_\infty(x) f(x) + \Delta^{-\ell} f(x), \tag{7.3}$$

and if the kernel function has its marginals in $\mathcal{G}$, we can rewrite the kernelized Stein discrepancy (7.2) using Theorem 6.3.2:

$$\begin{aligned}\mathbb{S}_k(p_\infty, p_n) &:= \mathbb{S}_k(p_\infty, p_n; \mathcal{A}_\infty) \\ &= \mathbb{E}\left[(\rho_\infty(Y_1) - \rho_n(Y_1))k(Y_1, Y_2)(\rho_\infty(Y_2) - \rho_n(Y_2))\right] + \kappa^*\end{aligned}$$

where $\kappa^* = \mathbb{E}[\mathcal{T}_n^\ell k(Y_1, \cdot - \ell)(Y_2)] + \mathbb{E}[\mathcal{T}_n^\ell k(\cdot - \ell, Y_2)(Y_1)]$ is the remainder term which is equal to zero if $\mathcal{G} \subset \mathcal{F}(\mathcal{A}_n)$.

In the following, we restrict the subclass $\mathcal{G}$ in order to satisfy this assumption. Therefore, clearly, $\mathbb{S}_K(p_\infty, p_n) = 0$ if $p_n = p_\infty$ (this is immediate by construction of the Stein operator). The converse also holds true if the kernel function k is well chosen. Both articles obtained this result for continuous distribution (Liu et al., 2016, Proposition 3.3 and Chwialkowski et al., 2016, Theorem 2.2). However, it can be extended to any univariate distribution. Suppose that $\mathcal{F}$ is a Hilbert space when equipped with the usual inner product $\langle f, g\rangle_{\mathcal{F}} = \mathbb{E}\left[f(X)g(X)\right]$ and let $(e_j)_{j\geq 0}$ form a basis of $\mathcal{F}$. Suppose that the kernel k admits a representation of the form $k(x_1, x_2) = \sum_j \alpha_j e_j(x_1)e_j(x_2)$ where the coefficients α_j are all strictly positive. Then $\mathbb{S}_k(p_\infty, p_n) \geq 0$ with equality if and only if $p_\infty = p_n$. Indeed, under the stated assumptions, we have that

$$\mathbb{S}_k(p_\infty, p_n) = \sum_j \alpha_j \mathbb{E}\left[\mathcal{A}_\infty e_j(Y_1)\mathcal{A}_\infty e_j(Y_2)\right] = \sum_j \alpha_j (\mathbb{E}\big[\mathcal{A}_\infty e_j(Y_1)\big])^2 \geq 0.$$

Moreover, the strict positivity of all α_j ensures that $\mathbb{E}\big[\mathcal{A}_\infty e_j(Y_1)\big] = 0$ for *all* j so that $\mathbb{E}[\mathcal{A}_\infty f(Y_1)] = 0$ for all $f \in \mathcal{F}$.

Note that independence of Y_1, Y_2 is not necessary, and the claim still holds under the weaker assumption that $\mathbb{E}[f(Y_1)f(Y_2)] \geq 0$ for all $f \in \mathcal{F}$. Also, positivity of $\mathbb{S}$ still holds if some of the coefficients α_j cancel. Even the characterizing nature is not a requirement.

In practice, $\mathbb{S}_k(p_\infty, p_n)$ is useless if p_∞ is unknown (p_∞ could also be intractable). Nevertheless, there are many situations in which the right-hand side of (7.2) remains computable. The main example we can think of is when the densities are only known up to a normalizing constant, for instance in the Bayesian context. In this case, the score function does not depend on the normalisation constant. Indeed, if f is a function which is proportional to the desired probability density, we have

$$\rho_p^\ell(x) = \Delta^\ell p(x)/p(x) = \Delta^\ell f(x)/f(x).$$

The kernelized Stein discrepancy (7.2) can be written using the expression of the considered Stein operator $\mathcal{A}_\infty$ given in (7.3):

$$\mathbb{S}_k(p_\infty, p_n) = \mathbb{E}\Big[\rho_\infty^\ell(Y_1)\rho_\infty^\ell(Y_2)k(Y_1,Y_2) + \rho_\infty^\ell(Y_2)\Delta_1^{-\ell}k(Y_1,Y_2) + \rho_\infty^\ell(Y_1)\Delta_2^{-\ell}k(Y_1,Y_2) + \Delta_1^{-\ell}\Delta_2^{-\ell}k(Y_1,Y_2)\Big]. \tag{7.4}$$

Therefore, given a sample $y_1, \ldots, y_N$ drawn from p_n (even without knowing this distribution properly), we propose to empirically estimate the kernelized Stein discrepancy through

$$\begin{aligned}\hat{\mathbb{S}}_k(y_1,\ldots,y_N;p_\infty) &=: \left(\binom{N}{2}\right)^{-1} \sum_{1\le i<j\le N}\sum \hat{K}(y_i,y_j) \\ &= \left(\binom{N}{2}\right)^{-1} \sum_{1\le i<j\le N}\sum \Big[\rho_\infty^\ell(y_i)\rho_\infty^\ell(y_j)k(y_i,y_j) + \rho_\infty^\ell(y_j)\Delta_1^{-\ell}k(y_i,y_j) \\ &\qquad + \rho_\infty^\ell(y_i)\Delta_2^{-\ell}k(y_i,y_j) + \Delta_1^{-\ell}\Delta_2^{-\ell}k(y_i,y_j)\Big]. \end{aligned} \tag{7.5}$$

As noted in Liu et al. (2016, Theorem 4.1 and Proposition 4.2), standard asymptotic properties of U-statistics found e.g. in Serfling (2009, Section 5.5) provide the asymptotic behaviour of (7.5). Assume that $\mathbb{E}\left[(\mathcal{A}_\infty^1\mathcal{A}_\infty^2 k(Y_1,Y_2))^2\right] < \infty$. If $p_\infty \neq p_n$, then $\hat{\mathbb{S}}_k$ is asymptotically normal with

$$\sqrt{N}\left(\hat{\mathbb{S}}_k(Y_1,\ldots,Y_N;p_\infty) - \mathbb{S}_k(p_\infty,p_n)\right) \xrightarrow{d} \mathcal{N}(0,\sigma^2) \tag{7.6}$$

where $\sigma^2 = \mathrm{Var}\left[\mathbb{E}[\mathcal{A}_\infty^1\mathcal{A}_\infty^2 k(Y_1,Y_2)\,|\,Y_1]\right] \neq 0$. Otherwise, if $p_\infty = p_n$, the U-statistics is degenerate, i.e,

$$N\hat{\mathbb{S}}_k(Y_1,\ldots,Y_n;p_\infty) \xrightarrow{d} \sum_j \lambda_j(Z_j^2 - 1) \tag{7.7}$$

where Z_j are i.i.d. standard Gaussian random variables, and $\{\lambda_j\}_{j\ge 0}$ are the eigenvalues of the (self-adjoint) operator $g \mapsto \mathcal{A}g(x) = \mathbb{E}\left[g(Y_2)(\mathcal{A}_\infty^1\mathcal{A}_\infty^2 k(Y_1,Y_2))\,|\,Y_1 = x\right]$.

1 Kernelized Stein Goodness-of-fit tests

These considerations lead to a natural and easily computed *kernelized Stein based* test statistic for $\mathcal{H}_0 : p_\infty = p_n$ vs $\mathcal{H}_1 : p_\infty \neq p_n$, where we reject the null hypothesis if $\hat{\mathbb{S}}_k(y_1,\ldots,y_N;p_\infty)$ is larger than the quantile of the limit distribution under the null. In general (for a general kernel and a general target distribution p_∞), the asymptotic null distribution given in (7.7) is intractable. In such cases, a bootstrap approximation of the distribution provides a useful alternative to the asymptotic distribution. This leads to a generally applicable test statistic.

A naive approach is to consider parametric bootstrap to generate a goodness-of-fit procedure. Indeed, we could generate bootstrap samples from p_∞, estimate the empirical discrepancy (7.5) for each of them, compute the quantile of level $\alpha/2$ and $(1-\alpha/2)$ of these univariate series and then compare the test statistic to these quantiles to reject or not the null hypothesis. The overall goodness-of-fit tests procedure is summarized in Algorithm 1. A weighted bootstrap procedure adapted to degenerate U-statistics could also be used as suggested in Huskova and Janssen (1993) and used in Liu et al. (2016). The alternative bootstrap loop is summarized in Algorithm 2. The asymptotic level of this bootstrap test is demonstrated in Huskova and Janssen (1993).

Algorithm 1: Goodness-of-fit tests using kernelized information Parametric bootstrap

Data: sample $y_1, \ldots, y_N$
Input: target distribution p_∞, kernel function $k(.,.)$, bootstrap sample size, significance level α
Result: reject or not the null hypothesis of the test $\mathcal{H}_0 : Y \sim p_\infty$ vs $\mathcal{H}_1 : Y \not\sim p_\infty$
Compute the statistic $\hat{\mathbb{S}}_k(y_1, \ldots, y_N; p_\infty)$;
for $i = 1, \ldots, nboot$ **do**
  Generate a sample drawn from p_∞;
  Compute the bootstrap statistic $\hat{\mathbb{S}}_k(\text{sample}_i; p_\infty)$;
Compute the critical values c_1 and c_2 given by the $\alpha/2$ and $(1-\alpha/2)$ quantiles of $\left(\hat{\mathbb{S}}_k(\text{sample}_i)\right)_{1 \le i \le nboot}$;
if $c_1 < \hat{\mathbb{S}}_k(y_1, \ldots, y_N; p_\infty)) < c_2$ **then**
  fail to reject the null hypothesis;
else
  reject the null hypothesis;

Algorithm 2: Alternative bootstrap for degenerate U-statistics

for $i = 1, \ldots, nboot$ **do**
  Generate a sample $n(w_1, \ldots, w_N)$ drawn from $\text{Mult}(N; \frac{1}{N}, \ldots, \frac{1}{N})$;
  Compute the bootstrap statistic
  $\hat{\mathbb{S}}_k^*(\text{sample}_i, p_\infty) = \binom{N}{2}^{-1} \sum_{1 \le i < j \le N} \sum \left(w_i - \frac{1}{N}\right)\left(w_j - \frac{1}{N}\right) k(y_i, y_j)$;

Choice of kernel function k

There exist many appropriate k functions. A first kernel is the radial basis function kernel (RBF):

$$k(x,y) = e^{-(x-y)^2/2} = \sum_{n=0}^{\infty} \frac{x^n}{\sqrt{n!}} e^{-x^2/2} \frac{y^n}{\sqrt{n!}} e^{-y^2/2}.$$

We may also consider the Hermite kernel defined with respect to the nth (probabilist) Hermite polynomial, $H_n(x) = (-1)^n(\varphi(x))^{(n)}/\varphi(x)$, by

$$m(x,y) = \frac{1}{\sqrt{2\pi}} \sum_{n=0}^{\infty} \frac{1}{n!} H_n(x) H_n(y).$$

Another example is the Mehler kernel which is also defined with respect to the same polynomials,

$$\ell(x,y) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho^2(x^2+y^2) - 2\rho xy}{2(1-\rho^2)} \right) = \sum_{n=0}^{\infty} \frac{\rho^n}{n!} H_n(x) H_n(y)$$

for $\rho \in (-1,1)$.

These kernel functions are generic and could be applied for any target distribution. An idea is to construct kernel function which are defined according to p_∞. We could choose functions e_i which are naturally associated with the target distribution (for instance, as the Hermite polynomials in the Gaussian case) and then, define a k function by

$$k_N(x,y) = \sum_{i=1}^{N} \alpha_i P_i(x) P_j(y) \tag{7.8}$$

for some $N \in \mathbb{N}$ and constants α_i.

For distributions belonging to the Ord family, Afendras et al. (2011) define orthogonal polynomials of degree at most i which could play the role of P_i in (7.8). If the constant are $\alpha_i = (\mathbb{E}_{p_\infty}[P_i^2(X)])^{-1}$, the k_N function is a sum of N orthonormal polynomials. The Afendras et al.' polynomials are defined in such a way that the following identity holds:

$$\mathbb{E}[P_i(X)g(X)] = \mathbb{E}[(\tau_p^-)^{[k]}(X)\Delta^k g(X)))].$$

Using our IBP relation (3.4) and our notations, we may easily rewrite these polynomials as

$$P_i(x) = (-1)^i \left(\mathcal{T}_p^-\right)^i (\tau_p^-)^{\lceil i \rceil}(x). \tag{7.9}$$

For example, the $n+1$ first polynomials associated with the binomial (n,θ) are non zero. The explicit expressions of the three first ones are $P_0(x) = 1$, $P_1(x) = x - n\theta$, $P_2(x) = x^2 - x2\theta(n-1) - x + \theta^2 n(n-1)$.

Illustration: simulation studies

In order to illustrate the efficiency of our *kernelized Stein based test*, we performed some simulations. We focus here on discrete distributions because, in the literature, there are quite few goodness-of-fit tests applicable to any discrete distribution. The power of the proposed test is compared with existing goodness-of-fit tests which are available for discrete distributions. The most well-known families of goodness-of-fit statistics are the *power-divergence statistics* (see e.g. Read and Cressie, 1988, for details) and the larger family of *ϕ-divergence statistics* defined by Csizar (1967) (see for instance Morales et al., 1995, for details). The power-divergence statistic is defined by

$$2n\mathcal{I}^{\lambda} := \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} O_i \left[\left(\frac{O_i}{E_i} \right)^{\lambda} - 1 \right], \tag{7.10}$$

where O_i (resp. E_i) are observed (resp. expected) frequencies in each cell, k in the number of cells and λ is a real parameter. The ϕ-divergence statistic is determined by

$$D_{\phi} = \sum_{i=1}^{k} E_i \phi(O_i/E_i). \tag{7.11}$$

As described in the review Cressie and Read (1989), the power-divergence statistic is a generalization of well-known tests which are obtained for particular values of the parameter λ: Pearson statistic ($\lambda = 1$), Cressie-Read statistic ($\lambda = 2/3$), log-likelihood ratio ($\lambda = 0$, defined by continuity), Freeman Tukey statistic ($\lambda = -1/2$), modified loglikelihood ratio ($\lambda = -1$, defined by continuity) and Neyman modified chi-squared ($\lambda = -2$). As mentioned in Morales et al. (1995), the particular ϕ function

$$\phi(x) = \frac{1}{\lambda(\lambda+1)}(x^{\lambda+1} - x)$$

illustrates the fact that the power-divergence statistic is a particular case of ϕ-divergence. The asymptotically chi-squared distribution of the ϕ-divergence statistics has been proven (see Morales et al., 1995) and the rate of convergence for some power-divergence statistics is still currently studied (see e.g. Gaunt and Reinert, 2016, Gaunt et al., 2017).

In our simulations, we compare the power-divergence statistics tests defined for particular λ values mentioned hereinabove with the kernelized Stein discrepancy tests defined by an arbitrary kernel (e.g. RBF kernel) and a kernel based on the target Stein operator (e.g. using the polynomials defined in Afendras et al., 2011).

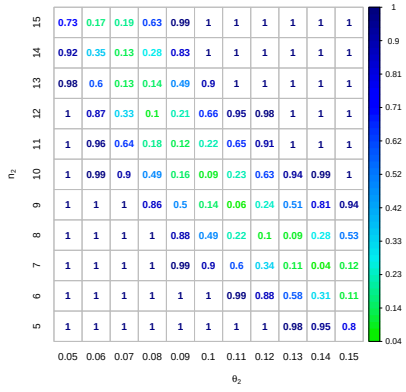
Simulations settings. We concentrate on the case where the target distribution is the binomial with parameters (n, θ). We compare the target distribution to samples

drawn from "close" distributions. We say that two distributions are "close" if the total variation distance between them is "close" to zero. We illustrate hereafter the cases when the samples are drawn from another binomial distribution with parameters (n_2, θ_2), the Beta Binomial distribution with parameters (α, β, n) or Poisson distribution with parameter $n\theta$. Our choices are motivated by results on the total variation distance between the two distributions and more specifically on upper bounds for such distance. Indeed, the total variation distance between two binomial random variables with constant mean ($n\theta = n_2\theta_2$) is smaller than $|\theta - \theta_2| / \max\{1 - \theta, 1 - \theta_2\}$ (by an application of Corollary 6.3.5). In complement, we note that Adell and Jodrá (2006) gives an upper bound for any θ_2 when $n_2 = n$. For the Beta Binomial distribution, Teerapabolarn (2008) obtain the upper bound $n(n-1)/((n+1)(1+\alpha+\beta))$ when $\theta = \alpha/(\alpha+\beta)$. The paper Holmes (2004) provides the upper bound $n\theta^2 \min\{1, 1/n\theta\}$ for the distance between Poisson (np) and binomial (n, θ) distributions. We note that other samples could also be used but are not developed hereafter. For instance, Holmes (2004) gives also an upper bound for the distance between hypergeometric and binomial distribution, Ehm (1991) did it for a Poisson binomial distribution, i.e. a sum of independent but not identically distributed Bernoulli random variables.

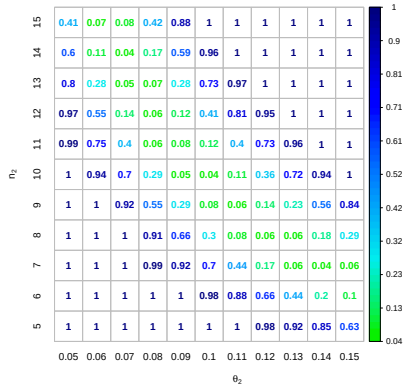
For each sample distribution, we consider 200 samples of different lengths (10, 20, 30, 50, 100 or 150). We compute the kernelized discrepancy according to a Stein operator associated with the target distribution, binomial $(10, 0.1)$, using the RBF kernel and the kernel defined using N Afendras et al's polynomials (for $N \in \{1, \ldots, n+1\}$) and compare the value to bootstrap quantiles. Moreover, the six previously mentioned power-divergence tests are computed as well using the asymptotic chi-squared quantiles.

Results. When the samples are drawn from another binomial distribution, none of the goodness-of-fit tests detect correctly deviations for the parameters when there is equality of the means. Figure 7.1 illustrates the proportion of rejections for each test according to the values (n_2, θ_2) of the sample distribution. The color code is used to facilitate the handling of the tables. For instance, if we compare the target binomial $(10, 0.1)$ with samples drawn from a binomial $(12, 0.06)$, Pearson test rejects in 87% of the cases when the kernelized goodness-of-fit test based on one polynomial has a rejection rate of 96%. Moreover, this figure illustrate that, in this setting, the proposed kernelized goodness-of-fit test has a similar power to Pearson test. Therefore, this method seems to be a serious competitor to existing tests.

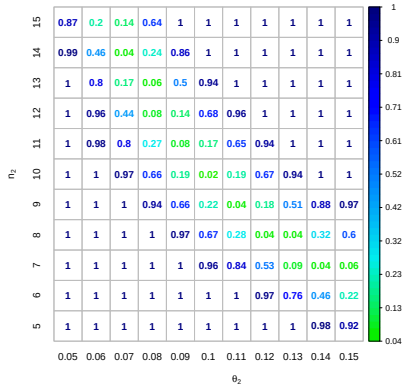
Table 7.1 illustrates the estimated power of kernelized tests compared to the power-divergence tests when the samples are drawn from another family, in this case, the Beta Binomial. The kernelized test using all, i.e. here $n+1$, Afendras polynomials is more powerful than power-divergence tests if the parameter α is quite



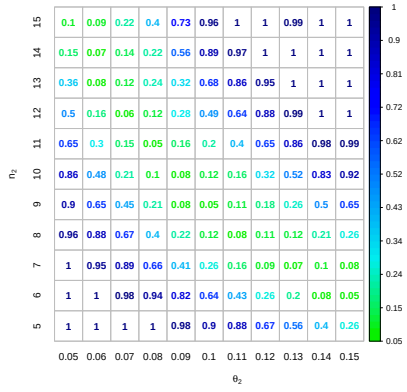
(a) Pearson test



(b) Kernelized GOF with RBF kernel



(c) Kernelized GOF with 1 Af. pol.



(d) Kernelized GOF with 11 Af. pol.

Figure 7.1: Proportion of rejections when 200 samples of size 150 are drawn from a Binomial (n_2, θ_2) and compared to the target distribution Binomial (10,0.1).

large. Moreover, we observed that the power increases with respect to the number of basis functions used in our k-function and the arbitrary RBF kernel is less powerful than the kernel function constructed according to the target distribution. Table 7.2 gives the estimated power of tests based on samples drawn from a Poisson distribution with parameter $n\theta = 1$. We observe that kernelized goodness-of-fit tests have similar power to classic power-divergence tests even if Pearson remains the most powerful.

The preliminary conclusions are promising. However, the study needs to be pursued for instance by considering other target distributions for which appropriate kernel functions are constructed. For instance, we may wonder if iteration of the Stein operator, as (7.9) for Afendras et al.' polynomials, could be extended to other distributions.

2 A generalized MOM estimator

The knowledge of a Stein operator associated with a distribution may also be useful in order to get estimations of the parameters using a "moments-type" method. As illustration, we consider the K-distribution. This is a family of continuous probability distributions on $(0, \infty)$ which is widely used in applications, for example, for modelling radar signals (Watts, 1985), non-normal statistical properties of radiation (Jakeman and Tough, 1987) and in wireless signal processing (Dong, 2012). We consider $Z_2 \sim KD_2(\lambda, c)$ which has a distribution (for example, from Weinberg, 2016) given by

$$KD_2(x; \lambda, c) = \frac{2c}{\Gamma(\lambda)} \left(\frac{cx}{2}\right)^{\lambda} K_{\lambda-1}(cx) \tag{7.12}$$

where the modified Bessel function of the second kind is given, for $x > 0$ and $\nu \in \mathbb{R}$, by $K_\nu(x) = \int_0^\infty e^{-x\cosh(t)} \cosh(\nu t)\, \mathrm{d}t$. The parameters of the K-distribution are not easily estimated. For instance, since there is no closed form for the likelihood function, the maximum likelihood estimator must be numerically computed (Joughin et al., 1993, Bocquet, 2014). An unpublished paper of R. Gaunt, G. Mijoule, Y. Swan and G. Weinberg deduces that

$$\mathcal{A}_2 f(x) = \frac{1}{c^2} x^2 f''(x) + \frac{2\lambda + 3}{c^2} x f'(x) + \left(\frac{4\lambda}{c^2} - x^2\right) f(x) \tag{7.13}$$

is a Stein operator for the considered K-distribution. Then, the random variable X is distributed according to the K-distribution if and only if $\mathbb{E}[\mathcal{A}_2 f(X)] = 0$ for all functions f in the associated Stein class, i.e.,

$$\frac{1}{c^2}\mathbb{E}[X^2 f''(X)] + \frac{2\lambda + 3}{c^2}\mathbb{E}[X f'(X)] + \frac{4\lambda}{c^2}\mathbb{E}[f(X)] - \mathbb{E}[X^2 f(X)] = 0. \tag{7.14}$$

α	1.00	2.00	3.00	4.00	5.00	10.00	20.00	40.00	60.00	80.00	100.00
Pearson	1.00	0.80	0.67	0.40	0.32	0.24	0.10	0.12	0.10	0.12	0.06
Cressie-Read	1.00	0.80	0.64	0.38	0.29	0.20	0.10	0.10	0.08	0.11	0.06
Loglik ratio	1.00	0.73	0.52	0.26	0.18	0.11	0.06	0.05	0.06	0.06	0.02
Freeman-Tukey	0.99	0.64	0.41	0.23	0.15	0.10	0.07	0.07	0.06	0.07	0.06
Mod. loglik. ratio	0.96	0.52	0.29	0.17	0.12	0.07	0.04	0.04	0.04	0.06	0.04
Neyman	0.93	0.47	0.27	0.14	0.08	0.07	0.06	0.04	0.05	0.07	0.04
RBF kernel	0.91	0.46	0.33	0.15	0.12	0.08	0.06	0.05	0.06	0.06	0.04
1 Af. pol.	0.55	0.34	0.28	0.17	0.23	0.07	0.06	0.05	0.06	0.06	0.04
2 Af. pol.	0.76	0.38	0.36	0.21	0.15	0.14	0.10	0.07	0.08	0.10	0.06
3 Af. pol.	0.94	0.64	0.54	0.24	0.20	0.15	0.10	0.07	0.07	0.10	0.06
4 Af. pol.	0.94	0.66	0.54	0.29	0.24	0.18	0.11	0.10	0.12	0.14	0.08
5 Af. pol.	0.93	0.66	0.54	0.30	0.27	0.20	0.14	0.12	0.15	0.14	0.10
6 Af. pol.	0.93	0.68	0.57	0.35	0.28	0.24	0.15	0.12	0.17	0.16	0.12
7 Af. pol.	0.93	0.68	0.57	0.34	0.28	0.26	0.14	0.12	0.17	0.17	0.12
8 Af. pol.	0.93	0.68	0.57	0.34	0.28	0.25	0.14	0.12	0.17	0.17	0.12
9 Af. pol.	0.93	0.68	0.57	0.34	0.28	0.25	0.14	0.12	0.17	0.17	0.12
10 Af. pol.	0.93	0.68	0.57	0.34	0.28	0.25	0.14	0.12	0.17	0.17	0.12
11 Af. pol.	0.93	0.68	0.57	0.34	0.28	0.25	0.14	0.12	0.17	0.17	0.12

Table 7.1: Estimated power of GOF tests based on 200 samples of size 150 drawn from a Beta Binomial$(\alpha, \beta, 10)$ such that $\alpha/(\alpha+\beta) = 0.1$. These samples are compared to the target distribution Bin$(10, 0.1)$.

Sample size	10	20	30	50	100	150
Pearson	0.11	0.14	0.19	0.28	0.18	0.25
Cressie-Read	0.10	0.13	0.17	0.19	0.17	0.24
Loglik ratio	0.07	0.07	0.11	0.10	0.08	0.12
Freeman-Tukey	0.12	0.05	0.13	0.09	0.06	0.08
Mod. loglik. ratio	0.01	0.03	0.08	0.04	0.06	0.05
Neyman	0.07	0.12	0.13	0.08	0.08	0.04
RBF kernel	0.04	0.06	0.04	0.04	0.10	0.10
1 Af. pol.	0.10	0.06	0.10	0.08	0.07	0.10
2 Af. pol.	0.09	0.05	0.07	0.08	0.10	0.12
3 Af. pol.	0.09	0.08	0.10	0.12	0.08	0.18
4 Af. pol.	0.07	0.09	0.10	0.13	0.12	0.17
5 Af. pol.	0.06	0.07	0.12	0.13	0.12	0.18
6 Af. pol.	0.07	0.09	0.10	0.12	0.12	0.15
7 Af. pol.	0.05	0.11	0.10	0.12	0.12	0.14
8 Af. pol.	0.04	0.08	0.10	0.18	0.17	0.16
9 Af. pol.	0.04	0.08	0.12	0.16	0.17	0.17
10 Af. pol.	0.04	0.08	0.10	0.15	0.18	0.16
11 Af. pol.	0.04	0.08	0.10	0.15	0.17	0.16

Table 7.2: Estimated power of GOF tests based on 200 samples drawn from a Poisson with parameter 1. These samples are compared to the target distribution $\mathrm{Bin}(10, 0.1)$.

If we consider two well-chosen functions f_1 and f_2 to plug in (7.14), we obtain a system of two equations which leads to expressions for the parameters (λ, c) of the distribution. More precisely, we have

$$\begin{cases} \lambda = \dfrac{(A_1 + 3B_1)E_2 - E_1(A_2 + 3B_2)}{2E_1(B_2 + 2D_2) - 2E_2(B_1 + 2D_1)} \\ c^2 = \dfrac{-2(B_1 + 2D_1)(A_2 + 3B_2) + 2(A_1 + 3B_1)(B_2 + 2D_2)}{2E_1(B_2 + 2D_2) - 2E_2(B_1 + 2D_1)} \end{cases} \tag{7.15}$$

where the expectations are denoted by $A_i = \mathbb{E}[X^2 f_i''(X)]$, $B_i = \mathbb{E}[X f_i'(X)]$, $D_i = \mathbb{E}[f_i(X)]$ and $E_i = \mathbb{E}[X^2 f_i(X)]$ for $i = 1, 2$. These expectations can easily be empirically estimated. Different choices of functions lead to different estimators. A first trivial choice could be $f_1(x) = 1$ and $f_2(x) = x$ which leads to the estimators

$$\begin{cases} \hat{\lambda} = \dfrac{3m_1 m_2}{4m_3 - 6m_2 m_1} \\ \widehat{c^2} = \dfrac{12m_1}{4m_3 - 6m_2 m_1} \end{cases} \Leftrightarrow \begin{cases} \widehat{c^2} = \dfrac{6m_1}{2m_3 - 3m_2 m_1} \\ \hat{\lambda} = \widehat{c^2} \frac{m_2}{4} \end{cases} \tag{7.16}$$

where m_i is the ith empirical moment, i.e., $m_i = n^{-1}\sum_{j=1}^{n} x_j^i$. A second choice could be $f_1(x) = 1$ and $f_2(x) = 1/x$ which leads to other estimators of λ and c^2:

$$\begin{cases} \lambda = \dfrac{m_2 m_{-1}}{2m_2 m_{-1} - 4m_1} \\ c^2 = \dfrac{4m_{-1}}{2m_2 m_{-1} - 4m_1} \end{cases} \Leftrightarrow \begin{cases} \widehat{c^2} = \dfrac{2m_{-1}}{m_2 m_{-1} - 2m_1} \\ \hat{\lambda} = \widehat{c^2}\frac{m_2}{4} \end{cases} \tag{7.17}$$

Finally, a third choice of $f_1(x) = x$ and $f_2(x) = 1/x$ leads to the estimators

$$\begin{cases} \lambda = \dfrac{m_3 m_{-1} + 3m_1}{2m_3 m_{-1} - 6m_1} \\ c^2 = \dfrac{6m_1 m_{-1}}{m_3 m_{-1} - 3m_1} \end{cases} \tag{7.18}$$

In order to analyse the efficiency of the new estimators, some preliminary simulations are performed using the same simulation settings as Joughin et al. (1993). Three pairs of parameters were tested, $\lambda = 0.5, 1$ and 5. The parameter c is chosen in order to get a unit intensity, i.e., $c = 2\sqrt{\lambda}$ (see Joughin et al., 1993, for details). Different sample sizes n, were considered, i.e., 50, 100, 500 and 1000. For each combination of (λ, c, n), 1000 simulations were performed using Mathematica software. The maximum likelihood estimator is performed on a unique parameter by considering $c = 2\sqrt{\lambda}$. As mentioned in literature (for instance in Abraham and Lyons, 2010), this estimator does not always converge to a solution. Moreover, it is computationally expensive. Our estimators are compared to the estimations of λ developed in Joughin et al. (1993) which are denoted by SFMOM (using second and fourth moments) and FSMOM (using first and second moments). The results of these experiments are available in Table 7.3 where the mean square error is given and graphical representations are available in Figures 7.2 ($\lambda = 0.5$), 7.3 ($\lambda = 1$) and 7.4 ($\lambda = 2$). These simulations allow us to make some preliminary observations. As expected, the variability of each estimator decreases with the sample size but it increases with the parameter λ. As mentioned in Joughin et al. (1993), FSMOM has smaller variability than SFMOM. Moreover, our proposed estimators have a variability which is located between the estimators FSMOM and SFMOM. These preliminary conclusions induce that our method can be competitive with the classic estimators and the properties of such estimators could be further studied.

3 Perspectives

To conclude this thesis, in addition to the two previous sections, we briefly point out some questions that arose during this research project and which may lead to future work on this topic.

		n=50	n=100	n=1000
$\lambda = 0.5$	SFMOM	43.0866	0.229071	0.0144039
	FSMOM	0.0319409	0.0259148	0.0193133
	ML λ	0.0237763	0.0233405	0.0235395
	$\hat{\lambda}_1$	0.393838	0.0664699	0.00475794
	$\hat{\lambda}_2$	0.0415932	0.0192908	0.00577479
	$\hat{\lambda}_3$	635.81	9.22421	0.057265
	ML c^2	0.380421	0.373448	0.376631
	$\widehat{c^2}_1$	430.017	102.677	41.3536
	$\widehat{c^2}_2$	117.409	78.0061	51.7445
	$\widehat{c^2}_3$	131120.	2734.34	64.7791
$\lambda = 1$	SFMOM	563.088	1.14658	0.0551715
	FSMOM	0.0448625	0.0452456	0.0455972
	ML λ	0.404689	0.0679469	0.00487089
	$\hat{\lambda}_1$	3.38022	0.374737	0.0218116
	$\hat{\lambda}_2$	23.5247	0.167231	0.0101334
	$\hat{\lambda}_3$	335.494	21.0152	0.108206
	ML c^2	6.47503	1.08715	0.0779343
	$\widehat{c^2}_1$	115.384	9.85514	0.513627
	$\widehat{c^2}_2$	997.124	4.98162	0.251957
	$\widehat{c^2}_3$	7319.39	457.62	1.31338
$\lambda = 2$	SFMOM	1913.41	154366.	0.271962
	FSMOM	1.22616	1.2345	1.24411
	ML λ	1.30828	1.35742	1.38426
	$\hat{\lambda}_1$	659.615	113.847	0.134791
	$\hat{\lambda}_2$	1964.73	204453.	0.194577
	$\hat{\lambda}_3$	0.921526	0.774253	0.864112
	ML c^2	20.9326	21.7187	22.1481
	$\widehat{c^2}_1$	914.17	157.555	35.0013
	$\widehat{c^2}_2$	2820.84	292327.	34.8935
	$\widehat{c^2}_3$	41.3553	43.438	45.2798

Table 7.3: MSE for the estimation with parameters c^2 and λ based on 1000 simulations of N samples.

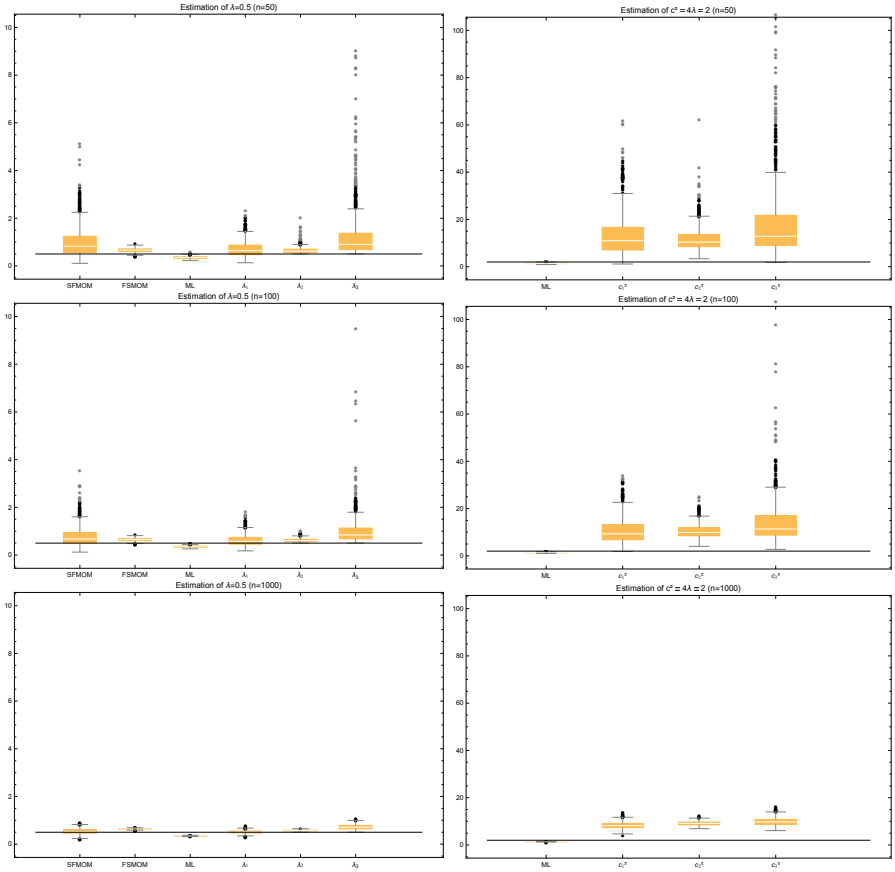


Figure 7.2: Distribution of estimations for $\lambda = 0.5$ (left panels) and $c^2 = 2$ (right panels) when the sample size is $n = 50$ (upper panels), $n = 100$ (middle panels) or $n = 1000$ (lower panels).

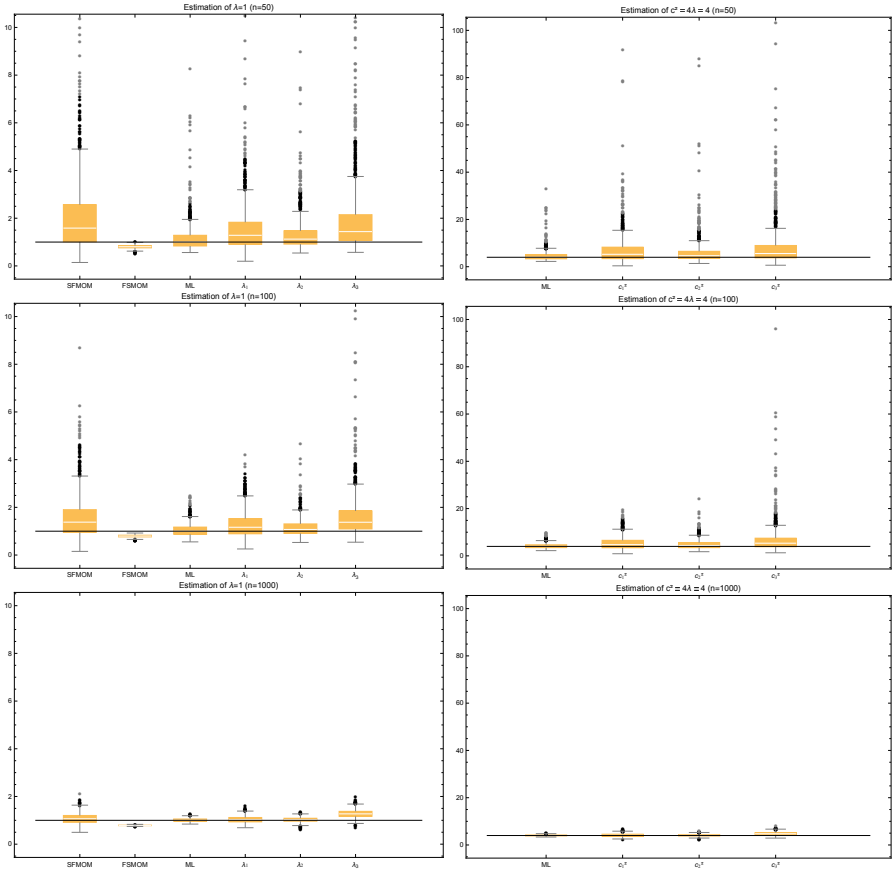


Figure 7.3: Distribution of estimations for $\lambda = 1$ (left panels) and $c^2 = 4$ (right panels) when the sample size is $n = 50$ (upper panels), $n = 100$ (middle panels) or $n = 1000$ (lower panels).

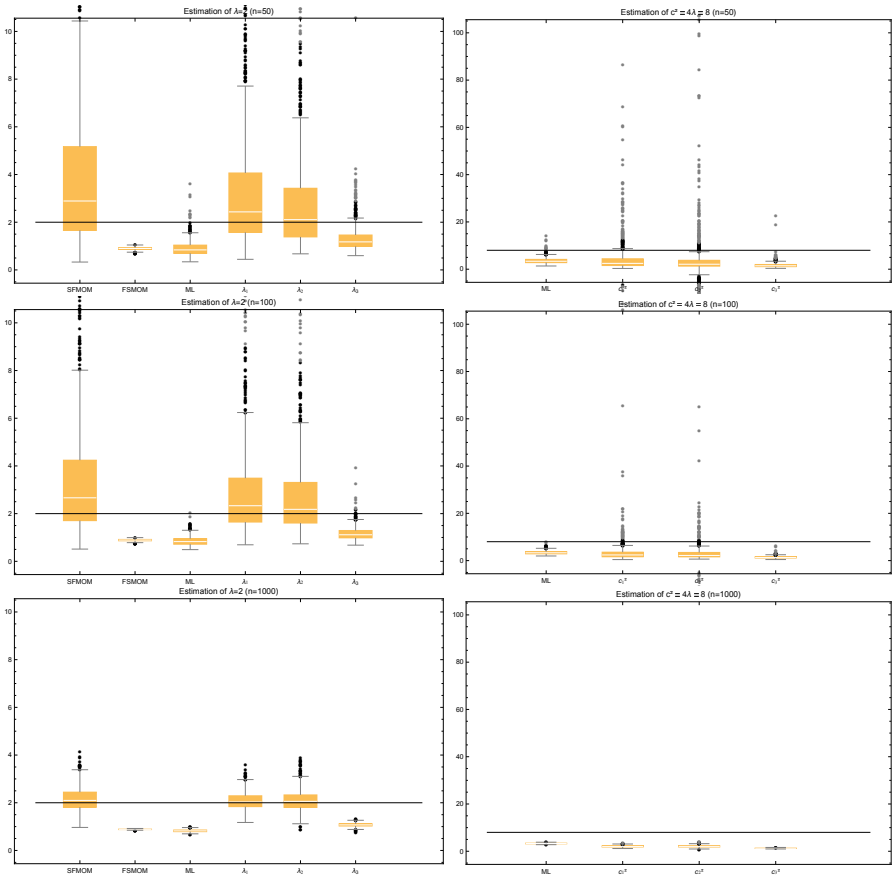


Figure 7.4: Distribution of estimations for $\lambda = 2$ (left panels) and $c^2 = 8$ (right panels) when the sample size is $n = 50$ (upper panels), $n = 100$ (middle panels) or $n = 1000$ (lower panels).

Firstly, as we focus on Stein's method, we did not get the chance to go back to our initial problem about multiple testing. As discussed at the beginning of Part II, we may study the impact of some dependency on the distribution of the number of rejections. Many questions can be deduced from this problematic. For instance, when the data are "lightly" dependent, how the number of rejections behaves under the null hypotheses? Can we quantify the error committed if we consider the binomial distribution instead of the real distribution to determine the cut-off values? What is a "light" dependency and how could we measure it?

At the end of Chapter 4, we highlight the connection between our theory on the one hand and the spectral gap and Poincaré inequalities on the other hand, e.g. based on Bonnefont et al. (2016) and Roustant et al. (2017). It would be of interest to further study the link between those topics.

In Chapter 5, we develop the weight functions in the continuous case for arbitrary function h (see Lemma 5.4.1) but such a generalisation seems difficult for the discrete case. However, we manage to find an expression for $k = 2$ which could be analogous to the expression (5.20) where the factorial coefficients are replaced by the sum of different $\Gamma_2^{\ell_1,\ell_2}$ functions. More precisely, we have $\Gamma_2^{+-}h(x) + \Gamma_2^{-+}h(x) = \frac{1}{p(x)}\mathbb{E}\Big[\big(h(x)-h(X_1)(h(X_2)-h(x)\big)\big(h(X_2)-h(X_1)\big)\mathbb{I}[X_1 < x < X_2]\Big]$. We spent some time trying to generalize this result for larger k but, until now, we have not managed to find a nice generalisation for larger k and arbitrary function h. Of course, the particular choice of $h = \mathrm{Id}$ leads to the simplified expression of Lemma 5.4.3.

Finally, in Example 5.4.11, we mentioned the natural derivative for binomial distribution introduced in Hillion et al. (2014). It could be interesting to generalise such a derivative to other distributions.

In Chapter 6, we develop Stein factors for Gaussian, exponential and Poisson distributions. Other targets could be covered. It will be displayed in the supplementary material related to the article in preparation for publication.

In Section 3, we provide bounds on IPMs for distributions which share a common dominating measure. However, as mentioned in the text, Theorem 6.3.2 could be extended to more general cases which do not satisfy this assumption, as treated in Goldstein and Reinert (2013).

Bibliography

Abraham, D. A. and A. P. Lyons (2010). Reliable methods for estimating the K-distribution shape parameter. *IEEE Journal of Oceanic Engineering 35*(2), 288–302.

Adell, J. A. and P. Jodrá (2006). Exact Kolmogorov and total variation distances between some familiar discrete distributions. *Journal of Inequalities and Applications 2006*(1), 1–8.

Afendras, G. (2013). Unified extension of variance bounds for integrated Pearson family. *Annals of the Institute of Statistical Mathematics 65*(4), 687–702.

Afendras, G., N. Balakrishnan, and N. Papadatos (2018). Orthogonal polynomials in the cumulative Ord family and its application to variance bounds. *Statistics 52*(2), 364–392.

Afendras, G. and N. Papadatos (2011). On matrix variance inequalities. *Journal of Statistical Planning and Inference 141*(11), 3628–3631.

Afendras, G. and N. Papadatos (2014). Strengthened Chernoff-type variance bounds. *Bernoulli 20*(1), 245–264.

Afendras, G., N. Papadatos, and V. Papathanasiou (2007). The discrete Mohr and Noll inequality with applications to variance bounds. *Sankhyā: The Indian Journal of Statistics 69*(2), 162–189.

Afendras, G., N. Papadatos, and V. Papathanasiou (2011). An extended Stein-type covariance identity for the Pearson family with applications to lower variance bounds. *Bernoulli 17*(2), 507–529.

Afendras, G. and V. Papathanasiou (2014). A note on a variance bound for the multinomial and the negative multinomial distribution. *Naval Research Logistics (NRL) 61*(3), 179–183.

Altay, H. and F. Celebioglu (2015). The impacts of political terrorism on gross domestic product in Eurasia: A spatial data analysis. *Eurasian Journal of Business and Economics 8*(15), 21–37.

Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Analysis 27*(2), 93–115.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. *Spatial Analytical Perspectives on GIS 4*, 111–125.

Anselin, L. (2019). A local indicator of multivariate spatial association: extending Geary's c. *Geographical Analysis 51*(2), 133–150.

Anselin, L., A. K. Bera, R. Florax, and M. J. Yoon (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics 26*(1), 77–104.

Anselin, L. and S. Rey (1991). Properties of tests for spatial dependence in linear regression models. *Geographical Analysis 23*(2), 112–131.

Archimbaud, A., K. Nordhausen, and A. Ruiz-Gazen (2018). ICS for multivariate outlier detection with application to quality control. *Computational Statistics & Data Analysis 128*, 184–199.

Arras, B. and C. Houdré (2019a). *On Stein's method for infinitely divisible laws with finite first moment*. Springer.

Arras, B. and C. Houdré (2019b). On Stein's method for multivariate self-decomposable laws with finite first moment. *Electronic Journal of Probability 24*, 1–33.

Barbour, A., M. J. Luczak, and A. Xia (2018). Multivariate approximation in total variation, II: Discrete normal approximation. *The Annals of Probability 46*(3), 1405–1440.

Barbour, A. D. and L. H. Y. Chen (2005a). *An introduction to Stein's method*, Volume 4 of *Lecture Notes Series, Institute of Mathematical Sciences, National University of Singapore*. Singapore University Press.

Barbour, A. D. and L. H. Y. Chen (2005b). *Stein's method and applications*, Volume 5 of *Lecture Notes Series, Institute of Mathematical Sciences, National University of Singapore*. Singapore University Press.

Barbour, A. D. and P. Hall (1984). On the rate of Poisson convergence. *Mathematical Proceedings of the Cambridge Philosophical Society 95*(3), 473–480.

Barbour, A. D., L. Holst, and S. Janson (1992). *Poisson approximation*. Clarendon Press Oxford.

Barbour, A. D. and A. Xia (2006). On Stein's factors for Poisson approximation in Wasserstein distance. *Bernoulli 12*(6), 943–954.

Baricz, Á. (2008). Mills' ratio: monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications 340*(2), 1362–1370.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29*(4), 1165–1188.

Bivand, R., W. G. Müller, and M. Reder (2009). Power calculations for global and local moran's I. *Computational Statistics & Data Analysis 53*(8), 2859–2872.

Bivand, R. S. and D. W. Wong (2018). Comparing implementations of global and local indicators of spatial association. *Test 27*(3), 716–748.

Blázquez, F. L. and B. S. Miño (2014). Maximal correlation in a non-diagonal case. *Journal of Multivariate Analysis 131*, 265–278.

Bobkov, S. G., F. Götze, and C. Houdré (2001). On Gaussian and Bernoulli covariance representations. *Bernoulli 7*(3), 439–451.

Bocquet, S. (2014). Parameter estimation for Pareto and K distributed clutter with noise. *IET Radar, Sonar & Navigation 9*(1), 104–113.

Bonnefont, M. and A. Joulin (2014). Intertwining relations for one-dimensional diffusions and application to functional inequalities. *Potential Analysis 41*(4), 1005–1031.

Bonnefont, M. and A. Joulin (2019). A note on eigenvalues estimates for one-dimensional diffusion operators. *arXiv preprint arXiv:1906.02496*.

Bonnefont, M., A. Joulin, and Y. Ma (2016). A note on spectral gap and weighted Poincaré inequalities for some one-dimensional diffusions. *ESAIM: Probability and Statistics 20*, 18–29.

Borovkov, A. and S. Utev (1984). On an inequality and a related characterization of the normal distribution. *Theory of Probability & its Applications 28*(2), 219–228.

Brascamp, H. J. and E. H. Lieb (1976). On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis 22*(4), 366–389.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, and J. Sander (2000). LOF: identifying density-based local outliers. *Sigmod record 29*(2), 93–104.

Cacoullos, T. (1982). On upper and lower bounds for the variance of a function of a random variable. *The Annals of Probability 10*(3), 799–809.

Cacoullos, T., N. Papadatos, and V. Papathanasiou (1998). Variance inequalities for covariance kernels and applications to central limit theorems. *Theory of Probability & its Applications 42*(1), 149–155.

Cacoullos, T. and V. Papathanasiou (1985). On upper and lower bounds for the variance of functions of a random variable. *Statistics & Probability Letters 3*, 175–184.

Cacoullos, T. and V. Papathanasiou (1986). Bounds for the variance of functions of random variables by orthogonal polynomials and Bhattacharyya bounds. *Statistics & Probability Letters 4*(1), 21–23.

Cacoullos, T. and V. Papathanasiou (1989). Characterizations of distributions by variance bounds. *Statistics & Probability Letters 7*(5), 351–356.

Cacoullos, T. and V. Papathanasiou (1992). Lower variance bounds and a new proof of the central limit theorem. *Journal of Multivariate Analysis 43*(2), 173–184.

Cacoullos, T. and V. Papathanasiou (1995). A generalization of covariance identity and related characterizations. *Mathematical Methods of Statistics 4*(1), 106–113.

Cacoullos, T., V. Papathanasiou, and S. A. Utev (1994). Variational inequalities with examples and an application to the central limit theorem. *The Annals of Probability 22*(3), 1607–1618.

Cai, T. T. and W. Liu (2016). Large-scale multiple testing of correlations. *Journal of the American Statistical Association 111*(513), 229–240.

Campbell, J. and M. Shin (2011). *Essentials of Geographic Information Systems*. Flat world knowledge.

Capéraà, P. and A. I. G. Guillem (1997). Taux de résistance des tests de rang d'indépendance. *Canadian Journal of Statistics 25*(1), 113–124.

Carlen, E. A., D. Cordero-Erausquin, and E. H. Lieb (2013). Asymmetric covariance estimates of Brascamp-Lieb type and related inequalities for log-concave measures. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 49*, 1–12.

Cerioli, A. and A. Farcomeni (2011). Error rates for multivariate outlier detection. *Computational Statistics & Data Analysis 55*(1), 544–553.

Chang, W.-Y. and D. S. P. Richards (1999). Variance inequalities for functions of multivariate random variables. *Advances in Stochastic Inequalities: AMS Special Session on Stochastic Inequalities and Their Applications, October 17-19, 1997, Georgia Institute of Technology 234*, 43.

Chatterjee, S. (2014a). A short survey of Stein's method. *ArXiv preprint arXiv:1404.1392*.

Chatterjee, S. (2014b). *Superconcentration and related topics*. Springer.

Chatterjee, S., J. Fulman, and A. Röllin (2011). Exponential approximation by Stein's method and spectral graph theory. *ALEA Latin American Journal of Probability and Mathematical Statistics 8*, 197–223.

Chatterjee, S. and Q.-M. Shao (2011). Nonnormal approximation by Stein's method of exchangeable pairs with application to the Curie-Weiss model. *The Annals of Applied Probability 21*(2), 464–483.

Chawla, S. and P. Sun (2006). SLOM: a new measure for local spatial outliers. *Knowledge and Information Systems 9*(4), 412–429.

Chen, D., C.-T. Lu, Y. Kou, and F. Chen (2008). On detecting spatial outliers. *Geoinformatica 12*(4), 455–475.

Chen, L. H. (1982). An inequality for the multivariate normal distribution. *Journal of Multivariate Analysis 12*(2), 306–315.

Chen, L. H. (1985). Poincaré-type inequalities via stochastic integrals. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 69*(2), 251–277.

Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *The Annals of Probability 3*(3), 534–545.

Chen, L. H. Y., L. Goldstein, and Q.-M. Shao (2011). *Normal approximation by Stein's method*. Probability and its Applications (New York). Heidelberg: Springer.

Chen, P., I. Nourdin, and L. Xu (2018). Stein's method for asymmetric α-stable distributions, with application to the stable CLT. *arXiv preprint arXiv:1808.02405*.

Chen, Y., X. Dang, H. Peng, and H. L. Bart (2008). Outlier detection with the kernelized spatial depth function. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*(2), 288–305.

Chernoff, H. (1980). The identification of an element of a large population in the presence of noise. *The Annals of Statistics 8*(6), 1179–1197.

Chernoff, H. (1981). A note on an inequality involving the normal distribution. *The Annals of Probability 9*(3), 533–535.

Chwialkowski, K., H. Strathmann, and A. Gretton (2016). A kernel test of goodness of fit. In *International Conference on Machine Learning*, pp. 2606–2615.

Cliff, A. and K. Ord (1972). Testing for spatial autocorrelation among regression residuals. *Geographical Analysis 4*(3), 267–284.

Cliff, A. and K. Ord (1973). *Spatial autocorrelation*, Volume 5. Pion London.

Coakley, C. W. and T. P. Hettmansperger (1992). Breakdown bounds and expected test resistance. *Journal of Nonparametric Statistics 1*(4), 267–276.

Courtade, T. A., M. Fathi, and A. Pananjady (2019). Existence of Stein kernels under a spectral gap, and discrepancy bound. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 55*(2), 777–790.

Cressie, N. and T. R. Read (1989). Pearson's χ^2 and the loglikelihood ratio statistic G^2: a comparative review. *International Statistical Review 57*(1), 19–43.

Cuadras, C. M. (2002). On the covariance between functions. *Journal of Multivariate Analysis 81*(1), 19–27.

Cuadras, C. M. and D. Cuadras (2008). Eigenanalysis on a bivariate covariance kernel. *Journal of Multivariate Analysis 99*(10), 2497–2507.

Cui, X., L. Lin, and G. Yang (2008). An extended projection data depth and its applications to discrimination. *Communications in Statistics – Theory and Methods 37*(14), 2276–2290.

Dang, X. and R. Serfling (2010). Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties. *Journal of Statistical Planning and Inference 140*(1), 198–213.

de Jong, P., C. Sprenger, and F. V. Veen (1984). On extreme values of Moran's I and Geary's c. *Geographical Analysis 16*(1), 17–24.

Diaconis, P. and S. Zabell (1991). Closed form summation for classical distributions: variations on a theme of de Moivre. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics 6*(3), 284–302.

Djenouri, Y., A. Belhadi, J. C.-W. Lin, D. Djenouri, and A. Cano (2019). A survey on urban traffic anomalies detection algorithms. *IEEE Access 7*, 12192–12205.

Djenouri, Y. and A. Zimek (2018). Outlier detection in urban traffic data. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, pp. 3. ACM.

Djenouri, Y., A. Zimek, and M. Chiarandini (2018). Outlier detection in urban traffic flow distributions. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 935–940. IEEE.

Döbler, C. (2012). Stein's method of exchangeable pairs for absolutely continuous, univariate distributions with applications to the Pólya urn model. *arXiv preprint arXiv:1207.0533*.

Döbler, C. (2015). Stein's method of exchangeable pairs for the beta distribution and generalizations. *Electronic Journal of Probability 20*(109), 1–34.

Döbler, C. and G. Peccati (2018). The gamma Stein equation and noncentral de Jong theorems. *Bernoulli 24*(4B), 3384–3421.

Dong, Y. (2012). Optimal coherent radar detection in a K-distributed clutter environment. *IET Radar, Sonar & Navigation 6*(5), 283–292.

Dray, S. (2011). A new perspective about Moran's coefficient: spatial autocorrelation as a linear regression problem. *Geographical Analysis 43*(2), 127–141.

Dray, S. and T. Jombart (2011). Revisiting Guerry's data: introducing spatial constraints in multivariate analysis. *The Annals of Applied Statistics 5*(4), 2278–2299.

Dray, S., P. Legendre, and P. R. Peres-Neto (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling 196*(3-4), 483–493.

Droesbeke, J.-J., M. Lejeune, and G. Saporta (2006). *Analyse statistique des données spatiales*. Editions TECHNIP.

Duembgen, L., R. Samworth, and J. Wellner (2019). Bounding distributional errors via density ratios. *arXiv preprint arXiv:1905.03009*.

Ehm, W. (1991). Binomial approximation to the Poisson binomial distribution. *Statistics & Probability Letters 11*(1), 7–16.

Eichelsbacher, P. and C. Thäle (2015). Malliavin-Stein method for variance-gamma approximation on Wiener space. *Electronic Journal of Probability 20*(123), 1–28.

Erhardsson, T. (2005). Steins method for Poisson and compound Poisson. *An introduction to Stein's method 4*, 61.

Ernst, M. and G. Haesbroeck (2017). Comparison of local outlier detection techniques in spatial multivariate data. *Data Mining and Knowledge Discovery 31*(2), 371–399.

Ernst, M., G. Reinert, and Y. Swan (2019a). First order covariance inequalities via Stein's method. *arXiv preprint arXiv:1906.08372*. To appear in Bernoulli.

Ernst, M., G. Reinert, and Y. Swan (2019b). On infinite covariance expansions. *arXiv preprint arXiv:1906.08376*.

Ernst, M. and Y. Swan (2019). Distances between distributions via Stein's method. *arXiv preprint arXiv:1909.11518*.

Ernst, M. D. (2004). Permutation methods: a basis for exact inference. *Statistical Science 19*(4), 676–685.

Fang, X., Q.-M. Shao, and L. Xu (2018). Multivariate approximations in Wasserstein distance by Stein's method and Bismut's formula. *Probability Theory and Related Fields 174*(3-4), 1–35.

Fathi, M. (2018). Higher-order Stein kernels for Gaussian approximation. *arXiv preprint arXiv:1812.02703*.

Fathi, M. (2019). Stein kernels and moment maps. *The Annals of Probability 47*(4), 2172–2185.

Filzmoser, P., A. Ruiz-Gazen, and C. Thomas-Agnan (2014). Identification of local multivariate outliers. *Statistical Papers 55*(1), 29–47.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Fritsch, V., G. Varoquaux, B. Thyreau, J.-B. Poline, and B. Thirion (2011). Detecting outlying subjects in high-dimensional neuroimaging datasets with regularized minimum covariance determinant. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*, pp. 264–271. Springer.

Fu, W., P. Jiang, G. Zhou, and K. Zhao (2014). Using Moran's I and GIS to study the spatial pattern of forest litter carbon density in a subtropical region of Southeastern China. *Biogeosciences 11*(8), 2401–2409.

Furioli, G., A. Pulvirenti, E. Terraneo, and G. Toscani (2017). Fokker–Planck equations in the modeling of socio-economic phenomena. *Mathematical Models and Methods in Applied Sciences 27*(01), 115–158.

Gaunt, R. E., A. M. Pickett, and G. Reinert (2017). Chi-square approximation by Stein's method with application to Pearson's statistic. *The Annals of Applied Probability 27*(2), 720–756.

Gaunt, R. E. and G. Reinert (2016). The rate of convergence of some asymptotically chi-square distributed statistics by Stein's method. *arXiv preprint arXiv:1603.01889*.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician 5*(3), 115–146.

Genton, M. G. (1998a). Highly robust variogram estimation. *Mathematical Geology 30*(2), 213–221.

Genton, M. G. (1998b). Spatial breakdown point of variogram estimators. *Mathematical Geology 30*(7), 853–871.

Genton, M. G. and A. Lucas (2003). Comprehensive definitions of breakdown points for independent and dependent observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(1), 81–94.

Genton, M. G. and A. Ruiz-Gazen (2010). Visualizing influential observations in dependent data. *Journal of Computational and Graphical Statistics 19*(4), 808–825.

Gervini, D. and V. J. Yohai (2002). A class of robust and fully efficient regression estimators. *The Annals of Statistics 30*(2), 583–616.

Getis, A. and J. K. Ord (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis 24*(3), 189–206.

Getis, A. and J. K. Ord (1993). Erratum: The analysis of spatial association by use of distance statistics. *Geographical Analysis 25*(3), 276–276.

Gneiting, T., W. Kleiber, and M. Schlather (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association 105*(491), 1167–1177.

Goldstein, L. and G. Reinert (2005). Distributional transformations, orthogonal polynomials, and Stein characterizations. *Journal of Theoretical Probability 18*(1), 237–260.

Goldstein, L. and G. Reinert (2013). Stein's method for the beta distribution and the Pólya-Eggenberger urn. *Journal of Applied Probability 50*(4), 1187–1205.

Gorham, J., A. B. Duncan, S. J. Vollmer, L. Mackey, et al. (2019). Measuring sample quality with diffusions. *The Annals of Applied Probability 29*(5), 2884–2928.

Gorham, J. and L. Mackey (2015). Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, pp. 226–234.

Gorham, J. and L. Mackey (2017). Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1292–1301. JMLR.org.

Hampel, F., P. Rousseeuw, E. Ronchetti, and W. Stahel (1986). *Robust Statistics: the approach based on influence functions*. Wiley.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics 42*(6), 1887–1896.

Harris, P., C. Brunsdon, M. Charlton, S. Juggins, and A. Clarke (2014). Multivariate spatial outlier detection using robust geographically weighted methods. *Mathematical Geosciences 46*(1), 1–31.

Harris, P., A. Clarke, S. Juggins, C. Brunsdon, and M. Charlton (2015). Enhancements to a geographically weighted principal component analysis in the context of an application to an environmental data set. *Geographical Analysis 47*(2), 146–172.

Haslett, J., R. Bradley, P. Craig, A. Unwin, and G. Wills (1991). Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *The American Statistician 45*(3), 234–242.

Havard, S., S. Deguen, D. Zmirou-Navier, C. Schillinger, and D. Bard (2009). Traffic-related air pollution and socioeconomic status: a spatial autocorrelation study to assess environmental equity on a small-area scale. *Epidemiology 20*(2), 223–230.

Hettmansperger, T. P. and J. W. McKean (2010). *Robust nonparametric statistical methods*. CRC Press.

Hillion, E., O. Johnson, and Y. Yu (2014). A natural derivative on $[0, n]$ and a binomial Poincaré inequality. *ESAIM: Probability and Statistics 18*, 703–712.

Höffding, W. (1940). Masstabinvariante korrelationstheorie. *Schriften des Mathematischen Instituts und Instituts fur Angewandte Mathematik der Universitat Berlin 5*, 181–233.

Höffding, W. (2012). *The collected works of Wassily Hoeffding*. Springer Science & Business Media.

Holmberg, H. and E. H. Lundevaller (2015). A test for robust detection of residual spatial autocorrelation with application to mortality rates in Sweden. *Spatial Statistics 14*, 365–381.

Holmes, S. (2004). Stein's method for birth and death chains. In *Stein's method: expository lectures and applications*, Volume 46, pp. 45–67. Institute of Mathematical Statistics.

Hope, A. C. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society. Series B (Methodological) 30*(3), 582–598.

Houdré, C. and A. Kagan (1995). Variance inequalities for functions of Gaussian variables. *Journal of Theoretical Probability 8*(1), 23–30.

Houdré, C. and V. Pérez-Abreu (1995). Covariance identities and inequalities for functionals on Wiener and Poisson spaces. *The Annals of Probability 23*(1), 400–419.

Houdré, C., V. Pérez-Abreu, and D. Surgailis (1998). Interpolation, correlation identities, and inequalities for infinitely divisible variables. *Journal of Fourier Analysis and Applications 4*(6), 651–668.

Huber, P. J. (1981). *Robust Statistics*. New-York : Wiley.

Hubert, M. and E. Vandervieren (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis 52*(12), 5186–5201.

Huskova, M. and P. Janssen (1993). Consistency of the generalized bootstrap for degenerate U-statistics. *The Annals of Statistics 21*(4), 1811–1823.

Ibrahim, S., I. Hamisu, and U. Lawal (2015). Spatial pattern of Tuberculosis prevalence in Nigeria: A comparative analysis of spatial autocorrelation indices. *American Journal of Geographic Information System 4*(3), 87–94.

Jakeman, E. and R. J. A. Tough (1987). Generalized K-distribution: a statistical model for weak scattering. *Journal of the Optical Society of America A 4*(9), 1764–1772.

Johnson, R. W. (1993). A note on variance bounds for a function of a Pearson variate. *Statistics & Risk Modeling 11*(3), 273–278.

Joughin, I. R., D. B. Percival, and D. P. Winebrenner (1993). Maximum likelihood estimation of K-distribution parameters for SAR data. *IEEE transactions on Geoscience and Remote Sensing 31*(5), 989–999.

Kamble, B. and K. Doke (2017). Outlier detection approaches in data mining. *International Research Journal of Engineering and Technology (IRJET) 4*, 634–638.

Karlin, S. (1993). A general class of variance inequalities. *Multivariate Analysis: Future Directions, Elsevier Science Publishers, New York*, 279–294.

Kelejian, H. H. and I. R. Prucha (2001). On the asymptotic distribution of the Moran I test statistic with applications. *Journal of Econometrics 104*(2), 219–257.

Klaassen, C. A. J. (1985). On an inequality of Chernoff. *The Annals of Probability 13*(3), 966–974.

Koldobsky, A. and S. J. Montgomery-Smith (1996). Inequalities of correlation type for symmetric stable random vectors. *Statistics & Probability Letters 28*(1), 91–97.

Korwar, R. (1991). On characterizations of distributions by mean absolute deviation and variance bounds. *Annals of the Institute of Statistical Mathematics 43*(2), 287–295.

Kriegel, H.-P., P. Kröger, E. Schubert, and A. Zimek (2011). Interpreting and unifying outlier scores. In *SIAM International Conference on Data Mining*, pp. 13–24. SIAM.

Kusuoka, S. and C. A. Tudor (2012). Stein's method for invariant measures of diffusions via Malliavin calculus. *Stochastic Processes and their Applications 122*(4), 1627–1651.

Lambert, D. (1981). Influence functions for testing. *Journal of the American Statistical Association 76*(375), 649–657.

Lambert, D. and W. Hall (1982). Asymptotic lognormality of p-values. *The Annals of Statistics 10*(1), 44–64.

Landsman, Z., S. Vanduffel, and J. Yao (2013). A note on Stein's lemma for multivariate elliptical distributions. *Journal of Statistical Planning and Inference 143*(11), 2016–2022.

Landsman, Z., S. Vanduffel, and J. Yao (2015). Some Stein-type inequalities for multivariate elliptical distributions and applications. *Statistics & Probability Letters 97*, 54–62.

Lark, R. (2008). Some results on the spatial breakdown point of robust point estimates of the variogram. *Mathematical Geosciences 40*(7), 729–751.

Ledoux, M. (1995). L'algèbre de Lie des gradients itérés d'un générateur markovien–développements de moyennes et entropies. *Annales Scientifiques de l'École Normale Supérieure 28*(4), 435–460.

Ledoux, M., I. Nourdin, and G. Peccati (2015). Stein's method, logarithmic Sobolev and transport inequalities. *Geometric and Functional Analysis 25*(1), 256–306.

Lee, S.-I. (2009). A generalized randomization approach to local measures of spatial association. *Geographical Analysis 41*(2), 221–248.

Leek, J. T. and J. D. Storey (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences 105*(48), 18718–18723.

Ley, C., G. Reinert, and Y. Swan (2017a). Distances between nested densities and a measure of the impact of the prior in bayesian statistics. *The Annals of Applied Probability 27*(1), 216–241.

Ley, C., G. Reinert, and Y. Swan (2017b). Stein's method for comparison of univariate distributions. *Probability Surveys 14*, 1–52.

Ley, C. and Y. Swan (2013a). Local Pinsker inequalities via Stein's discrete density approach. *IEEE Transactions on Information Theory 59*(9), 5584–5591.

Ley, C. and Y. Swan (2013b). Stein's density approach and information inequalities. *Electronic Communications in Probability 18*(7), 1–14.

Ley, C. and Y. Swan (2016). Parametric Stein operators and variance bounds. *Brazilian Journal of Probability and Statistics 30*, 171–195.

Lin, J. (2019). A local model for multivariate analysis: Extending Wartenberg's multivariate spatial correlation. *Geographical Analysis 0*.

Lin, K., Z. Long, and B. Ou (2009). Properties of bootstrap Moran's I for diagnostic testing a spatial autoregressive linear regression model. In *World Congress of the Spatial Econometrics Association, Barcelona*.

Liu, Q., J. Lee, and M. Jordan (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pp. 276–284.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics 18*(1), 405–414.

Lu, C.-T., D. Chen, and Y. Kou (2004). Multivariate spatial outlier detection. *International Journal on Artificial Intelligence Tools 13*(04), 801–811.

Ma, Y. and M. G. Genton (2000). Highly robust estimation of the autocovariance function. *Journal of Time Series Analysis 21*(6), 663–684.

Mackey, L. and J. Gorham (2016). Multivariate Stein factors for a class of strongly log-concave distributions. *Electronic Communications in Probability 21*(56), 1–14.

Maruyama, Y. (2015). An alternative to Moran's I for spatial autocorrelation. *arXiv preprint arXiv:1501.06260*.

McGrath, D. and C. Zhang (2003). Spatial distribution of soil organic carbon concentrations in grassland of Ireland. *Applied Geochemistry 18*(10), 1629–1639.

Melecky, L. (2015). Spatial autocorrelation method for local analysis of the EU. *Procedia Economics and Finance 23*, 1102–1109.

Menz, G. and F. Otto (2013). Uniform logarithmic Sobolev inequalities for conservative spin systems with super-quadratic single-site potential. *The Annals of Probability 41*(3B), 2182–2224.

Miclo, L. (2008). Quand est-ce que des bornes de Hardy permettent de calculer une constante de Poincaré exacte sur la droite ? In *Annales de la Faculté des Sciences de Toulouse*, Volume 17, pp. 121–192.

Miranda, H. and M. S. de Miranda (2011). Combining robustness with efficiency in the estimation of the variogram. *Mathematical Geosciences 43*(3), 363–377.

Mohr, E. and W. Noll (1952). Eine Bemerkung zur Schwarzschen Ungleichheit. *Mathematische Nachrichten 7*(1), 55–59.

Morales, D., L. Pardo, and I. Vajda (1995). Asymptotic divergence of estimates of discrete distributions. *Journal of statistical Planning and Inference 48*(3), 347–369.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika 37*(1/2), 17–23.

Murakami, D. and D. A. Griffith (2019). Eigenvector spatial filtering for large data sets: fixed and random effects approaches. *Geographical Analysis 51*(1), 23–49.

Murdoch, D. J., Y.-L. Tsai, and J. Adcock (2008). P-values are random variables. *The American Statistician 62*(3), 242–245.

Nash, J. (1958). Continuity of solutions of parabolic and elliptic equations. *The American Journal of Mathematics 80*, 931–954.

Nourdin, I. and G. Peccati (2012). *Normal approximations with Malliavin calculus : from Stein's method to universality*. Cambridge Tracts in Mathematics. Cambridge University Press.

Oja, H. (1999). Affine invariant multivariate sign and rank tests and corresponding estimates: a review. *Scandinavian Journal of Statistics 26*(3), 319–343.

O'Leary, B., J. J. Reiners Jr, X. Xu, and L. D. Lemke (2016). Identification and influence of spatio-temporal outliers in urban air quality measurements. *Science of the Total Environment 573*, 55–65.

Olkin, I. and L. Shepp (2005). A matrix variance inequality. *Journal of Statistical Planning and Inference 130*(1-2), 351–358.

Osei, F. B. and A. A. Duker (2008). Spatial and demographic patterns of cholera in Ashanti region-Ghana. *International Journal of Health Geographics 7*(1), 44.

Osland, L., I. S. Thorsen, and I. Thorsen (2016). Accounting for local spatial heterogeneities in housing market studies. *Journal of Regional Science 56*(5), 895–920.

Ou, B., X. Zhao, and M. Wang (2015). Power of moran's I test for spatial dependence in panel data models with time varying spatial weights matrices. *Journal of Systems Science and Information 3*(5), 463–471.

Paindaveine, D. and G. Van Bever (2013). From depth to local depth: a focus on centrality. *Journal of the American Statistical Association 108*(503), 1105–1119.

Papathanasiou, V. (1988). Variance bounds by a generalization of the Cauchy-Schwarz inequality. *Statistics & Probability Letters 7*(1), 29–33.

Papathanasiou, V. (1995). A characterization of the Pearson system of distributions and the associated orthogonal polynomials. *Annals of the Institute of Statistical Mathematics 47*(1), 171–176.

Petri, T. H. (2017). *Expression data analysis and regulatory network inference by means of correlation patterns*. Ph. D. thesis, Ludwig-Maximilians-Universitat Munchen.

Piegorsch, W. W. and A. J. Bailer (2005). *Analyzing environmental data*. John Wiley & Sons.

Pike, J. and H. Ren (2014). Stein's method and the Laplace distribution. *ALEA Latin American Journal of Probability and Mathematical Statistics 11*(2), 571–587.

Pinelis, I. (2015). Exact bounds on the closeness between the Student and standard normal distributions. *ESAIM: Probability and Statistics 19*, 24–27.

Quick, M., G. Li, and J. Law (2019). Spatiotemporal modeling of correlated small-area outcomes: Analyzing the shared and type-specific patterns of crime and disorder. *Geographical Analysis 51*, 221–248.

Rao, B. P. (2006). Matrix variance inequalities for multivariate distributions. *Statistical Methodology 3*(4), 416–430.

Read, T. R. and N. A. Cressie (1988). *Goodness-of-fit statistics for discrete multivariate data*. Springer Series in Statistics.

Reinert, G. (1995). A weak law of large numbers for empirical measures via Stein's method. *The Annals of Probability 23*(1), 334–354.

Reinert, G. (2004). Three general approaches to Stein's method. In *An introduction to Stein's method*, Volume 4. Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore.

Reinert, G., G. Mijoule, and Y. Swan (2018). Stein gradients and divergences for multivariate continuous distributions. *ArXiv preprint arXiv:1806.03478*.

Richardson, S., C. Guihenneuc, and V. Lasserre (1992). Spatial linear models with autocorrelated error structure. *Journal of the Royal Statistical Society: Series D (The Statistician) 41*(5), 539–557.

Ronchetti, E. (1997). Robust inference by influence functions. *Journal of statistical planning and inference 57*(1), 59–72.

Ross, N. (2011). Fundamentals of Stein's method. *Probability Surveys 8*, 210–293.

Rottoli, G. D., H. Merlino, and R. García-Martínez (2018). Knowledge discovery process for detection of spatial outliers. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 57–68. Springer.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications 8*(283-297), 37.

Rousseeuw, P. J. and C. Croux (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association 88*(424), 1273–1283.

Rousseeuw, P. J. and K. V. Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41*(3), 212–223.

Rousseeuw, P. J. and M. Hubert (1999). Regression depth. *Journal of the American Statistical Association 94*(446), 388–402.

Rousseeuw, P. J. and V. Yohai (1984). *Robust and non-linear time series analysis*, Volume 26 of *Lecture notes in Statistics*, Chapter Robust regression by means of S-estimators, pp. 256–272. New York : Springer.

Roustant, O., F. Barthe, and B. Iooss (2017). Poincaré inequalities on intervals–application to sensitivity analysis. *Electronic Journal of Statistics 11*(2), 3081–3119.

Saumard, A. (2019). Weighted Poincaré inequalities, concentration inequalities and tail bounds related to the behavior of the Stein kernel in dimension one. *Bernoulli 25*(4B), 3978–4006.

Saumard, A. and J. A. Wellner (2018). Efron's monotonicity property for measures on $\mathbb{R}^2$. *Journal of Multivariate Analysis 166*, 212–224.

Saumard, A. and J. A. Wellner (2019). On the isoperimetric constant, covariance inequalities and L_p-Poincaré inequalities in dimension one. *Bernoulli 25*(3), 1794–1815.

Schoutens, W. (2001). Orthogonal polynomials in Stein's method. *Journal of Mathematical Analysis and Applications 253*(2), 515–531.

Schubert, E., A. Koos, T. Emrich, A. Züfle, K. A. Schmid, and A. Zimek (2015). A framework for clustering uncertain data. *Proceedings of the Very Large Data Bases Endowment 8*(12), 1976–1979.

Schubert, E., M. Weiler, and A. Zimek (2015). Outlier detection and trend detection: two sides of the same coin. In *2015 IEEE International Conference on Data Mining Workshop*, pp. 40–46. IEEE.

Schubert, E., A. Zimek, and H.-P. Kriegel (2014). Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery 28*(1), 190–237.

Sen, A. (1976). Large sample-size distribution of statistics used in testing for spatial correlation. *Geographical Analysis 8*(2), 175–184.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, Volume 162. John Wiley & Sons.

Sijin, P., H. N. Champa, and K. R. Venugopal (2017). A survey on intent-based diversification for fuzzy keyword search. *International Journal of Computer Science and Information Technologies 8*(6), 602–618.

Singh, A. K. and S. Lalitha (2018). A novel spatial outlier detection technique. *Communications in Statistics-Theory and Methods 47*(1), 247–257.

Smyth, G. K. and B. Phipson (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology 9*(1), 1544–6115.

Soon, S. Y. (1996). Binomial approximation for dependent indicators. *Statistica Sinica 6*(3), 703–714.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California.

Stein, C. (1986). *Approximate Computation of Expectations*, Volume 7. JSTOR.

Su, L., C. Liang, X. Yang, and Y. Liu (2018). Influence factors analysis of provincial divorce rate spatial distribution in China. *Discrete Dynamics in Nature and Society 2018*, 1–11.

Sun, P. and S. Chawla (2004). On local spatial outliers. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*, pp. 209–216. IEEE.

Tanguy, K. (2017). *Quelques inégalités de superconcentration: théorie et applications*. Ph. D. thesis, Université Paul Sabatier-Toulouse III.

Teerapabolarn, K. (2008). A bound on the binomial approximation to the beta binomial distribution. *International Mathematical Forum 3*(28), 1355–1358.

Tiefelsdorf, M. (2002). The saddlepoint approximation of Moran's I's and local Moran's I_i's reference distributions and their numerical evaluation. *Geographical Analysis 34*(3), 187–206.

Toscani, G. (2019). Poincaré-type inequalities for stable densities. *Ricerche di Matematica 68*(1), 225–236.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, Volume 2, pp. 523–531.

Upadhye, N., V. Čekanavičius, and P. Vellaisamy (2017). On Stein operators for discrete approximations. *Bernoulli 23*(4A), 2828–2859.

Wang, R. (2014). Sum of arbitrarily dependent random variables. *Electronic Journal of Probability 19*(84), 1–18.

Watts, S. (1985). Radar detection prediction in sea clutter using the compound K-distribution model. *IEE Proceedings F (Communications, Radar and Signal Processing) 132*(7), 613–620.

Wei, Z. and X. Zhang (2009). Covariance matrix inequalities for functions of beta random variables. *Statistics & Probability Letters 79*(7), 873–879.

Weinberg, G. V. (2016). Error bounds on the Rayleigh approximation of the K-distribution. *IET Signal Processing 10*(3), 284–290.

Witten, D. M. and R. Tibshirani (2009). Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(3), 615–636.

Worboys, M. F. and M. Duckham (2004). *GIS: a computing perspective*. CRC press.

Wu, G., C. Zhang, P. Liu, W. Ren, Y. Zheng, F. Guo, X. Chen, and R. Higgs (2019). Spatial quantitative analysis of garlic price data based on ArcGIS technology. *CMC - Computers Materials & Continua 58*(1), 183–195.

Xu, L. (2019). Approximation of stable law in Wasserstein-1 distance by Stein's method. *The Annals of Applied Probability 29*(1), 458–504.

Ylvisaker, D. (1977). Test resistance. *Journal of the American Statistical Association 72*(359), 551–556.

Yohai, V. (1987). High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics 15*, 642–656.

Yu, C. and W. Yao (2017). Robust linear regression: a review and comparison. *Communications in Statistics-Simulation and Computation 46*(8), 6261–6282.

Zhang, C., L. Luo, W. Xu, and V. Ledwith (2008). Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. *Science of the Total Environment 398*(1), 212–221.

Zhang, T. (2008). Limiting distribution of the G statistics. *Statistics & Probability Letters 78*(12), 1656–1661.

Zimek, A. and P. Filzmoser (2018). There and back again: Outlier detection between statistical reasoning and data mining algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8*(6), e1280.

Zuo, Y. (2019). A new approach for the computation of halfspace depth in high dimensions. *Communications in Statistics-Simulation and Computation 48*(3), 900–921.

Zuo, Y. and R. Serfling (2000a). General notions of statistical depth function. *The Annals of Statistics 8*(2), 461–482.

Zuo, Y. and R. Serfling (2000b). Structural properties and convergence results for contours of sample statistical depth functions. *Annals of Statistics 28*(2), 483–499.

List of Figures

1.1 Illustration of local and global outliers in a univariate setting. 4
1.2 Illustration of Filzmoser et al.'s detection technique and its adaptation. . . . 10
1.3 Illustration of the regularized spatial detection technique. 15
1.4 Social data: detection based on Chen et al. (2008). 17
1.5 Social data: detection based on Harris et al. (2014). 18
1.6 Social data: detection based on Filzmoser et al. (2014). 18
1.7 Social data: detection by the regularized spatial detection technique. 19
1.8 Detection of local outliers on geochemical data measures in Northern Europe. 20
1.9 France with outlying departments colored according to the detection technique. 22
1.10 Gaussian 2nd-order stationary process based on the Matérn model. 26
1.11 Clean and 5%-contaminated 2-dimensional Gaussian process on the grid. . . 27
1.12 Results for simulations ($p = 2$) on a regular grid or Walloon municipalities. . 32
1.13 Results for simulations ($p = 5$) on a regular grid or Walloon municipalities. . 34

2.1 Different schemes for the spatial autocorrelation. 37
2.2 An arrow indicates if two areas are contiguous. 39
2.3 Map of the crude divorce rate in Belgium. 47
2.4 Influence functions on p-value for asymptotic Moran's test. 48
2.5 Influence functions on p-value for asymptotic Moran's tests. 49
2.6 Influence functions on p-value for permutation test. 51
2.7 Reordered weighting matrix to obtain a large or "average" rank Moran index. 53
2.8 Influence on p-value for the rank Moran test 55
2.9 Moran scatterplot. 57
2.10 Histograms of p-values under the null and alternative hypotheses. 64
2.11 Power curves for Moran's tests using simulations on a regular grid. 66
2.12 Influence functions on p-value for Geary's ratio. 70
2.13 Influence functions on p-value for Getis and Ord. 71
2.14 Selection of the subset A for queen contiguity to cover half of the grid 5×5. 76
2.15 Iterative selection of the subset A for rook contiguity to cover half of the grid. 76
2.16 Joint distribution of p-values. 82

3.1 The functions $x \mapsto K_p^\ell(x, x')/p(x)$ for different distributions p. 99
3.2 The functions $x \mapsto K_p^\ell(x, x)/p(x)$ for different distributions p. 100

6.1 Solution and absolute value of its derivative for Gaussian target. 163
6.2 Solution and absolute value of its derivative for exponential target. 163
6.3 Solution and absolute value of its derivative for exponential target (2). . . . 163
6.4 Solution and absolute value of its derivative for Poisson target. 164
6.5 Non-uniform and classic bounds for the Poisson distribution. 171
6.6 Exact value of $|g_\xi|$ and non-uniform bounds for the Poisson distribution. . . 171
6.7 Bounds on TV and Wasserstein distances between t_n and $\mathcal{N}(0,1)$. 180
6.8 Bounds on total variation distance between beta and gamma distributions. . 181
6.9 Bounds for TV distance between Poisson and binomial distributions. 182

7.1 Power of GOF tests to compare binomial distributions. 194
7.2 Distribution of estimations for the K-distribution with parameters $(0.5, \sqrt{2})$. 200
7.3 Distribution of estimations for the K-distribution with parameters $(1, 2)$. . . 201
7.4 Distribution of estimations for the K-distribution with parameters $(2, 2\sqrt{2})$. 202

List of Tables

1.1 Contingency table: real category vs classification resulting. 27
1.2 Error rates and Kappa measures for bivariate simulations. 33
1.3 Error rates and Kappa measures for five-dimensional simulations. 35

2.1 Moments of Moran's index, Geary's ratio and Getis and Ord's statistic. . . . 42
2.2 Considered tests for spatial autocorrelation. 59
2.3 Power for the Gaussian case with rook contiguity on a grid. 62
2.4 Power for the Gaussian case with queen contiguity on a grid. 63
2.5 Power for the simulations on Belgian municipalities. 65
2.6 Power for the Poisson and Bernoulli case on a grid. 67

4.1 Discrete distributions from the cumulative Ord family. 121
4.2 Discrete distributions from the cumulative Ord family (second part). 122
4.3 Continuous distributions from the Pearson family. 123

7.1 Power of GOF tests to compare binomial and beta binomial distributions. . 196
7.2 Power of GOF tests to compare binomial and Poisson distributions. 197
7.3 MSE for estimated parameters of K-distribution. 199