A cross-validation framework to extract data features for reducing structural uncertainty in subsurface heterogeneity

Jorge Lopez-Alvis[1,2], Thomas Hermans[2], and Frédéric Nguyen[1]

[1]Urban and Environmental Engineering, Applied Geophysics, University of Liege, Belgium
[2]Department of Geology, Ghent University, Belgium

September, 2019

Abstract

Spatial heterogeneity is a critical issue in the management of water resources. However, most studies do not consider uncertainty at different levels in the conceptualization of the subsurface patterns, for example using one single geological scenario to generate an ensemble of realizations. In this paper, we represent the spatial uncertainty by the use of hierarchical models in which higher-level parameters control the structure. Reduction of uncertainty in such higher-level structural parameters with observation data may be done by updating the complete hierarchical model, but this is, in general, computationally challenging. To address this, methods have been proposed that directly update these structural parameters by means of extracting lower dimensional representations of data called data features that are informative and applying a statistical estimation technique using these features. The difficulty of such methods, however, lies in the choice and design of data features, i.e. their extraction function and their dimensionality, which have been shown to be case-dependent. Therefore, we propose a cross-validation framework to properly assess the robustness of each designed feature and make the choice of the best feature more objective. Such framework aids also in choosing the values for the parameters of the statistical estimation technique, such as the bandwidth for kernel density estimation. We demonstrate the approach on a synthetic case with cross-hole ground penetrating

radar traveltime data and two higher-level structural parameters: discrete geological scenarios and the continuous preferential orientation of channels. With the best performing features selected according to the cross-validation score, we successfully reduce the uncertainty for these structural parameters in a computationally efficient way. While doing so, we also provide guidelines to design features accounting for the level of knowledge of the studied system.

Keywords— Bayesian hierarchical model, prior information, structural uncertainty, feature extraction, dimension reduction, spatial uncertainty

1 Introduction

Modeling subsurface systems requires accounting for uncertainty in many tasks such as reserve estimation, process understanding, decision making or water resources forecasting (Scheidt et al., 2018). To consider explicitly different sources of uncertainty, probabilistic approaches are often used (Tarantola and Valette, 1982; Tarantola, 2005) and allow to easily integrate any types of data or prior knowledge. In the Earth sciences, spatial heterogeneity is of utmost importance but its uncertainty is often not properly represented leading to over-simplifications of subsurface systems (Xu and Valocchi, 2015) and biased predictions made from such systems (Hermans et al., 2018).

Hydrogeological modelling is often hierarchical (Feyen and Caers, 2006; Tsai and Elshall, 2013; Comunian et al., 2016), in the sense that, based on available data, hydrogeologists first speculate on the nature of the depositional system (e.g., marine, deltaic or fluvial) and on global characteristics of the deposits (orientation or size of the structures) leading to the definition of different scenarios that serve as the basis for further modeling. Within each scenario, more specific spatial uncertainty rules can be defined. Each geological scenario might be expressed by its own training image or variogram model depicting the spatial uncertainty. Despite growing efforts made to model realistic prior geological information (see Linde et al., 2015, for a review), a single main structure is often considered which may underestimate the uncertainty or bias models if the structure is wrong (Linde et al., 2006). As an example, Hermans et al. (2015) demonstrated that the posterior distribution of hydrofacies constrained to electrical resistivity tomography and pumping data was dependent on the training image used and that ignoring the uncertainty on the depositional systems led to a biased solution. A possible strategy to avoid these problems is to consider hyperparameters—i.e., higher level parameters having their own prior probability distributions—leading to a so-called Bayesian hierarchical model (Gelman et al., 2014). Such hyperparameters may include the range of a variogram, the choice of training image or even the width of channels in a specific training image. These hierarchical problems have been addressed outside a Bayesian framework (see e.g. Khaninezhad and Jafarpour, 2014;

Golmohammadi and Jafarpour, 2016, in the context of geological scenario identification), but in doing so, the uncertainty in the results is generally not quantified.

Within a Bayesian framework such hierarchical model is then represented by a joint probability distribution involving hyperparameters, parameters and data, increasing the dimensionality of the joint space and making exploration more computationally demanding. Two different general approaches can be used to perform inference (i.e. updating uncertainty given some data) in such hierarchical models: (1) one-step methods where inference on the complete model (i.e., on both hyperparameters and parameters) is done in a single step, and (2) two-step methods where inference is done first for the hyperparameters and then the results are used to obtain the uncertainty on the parameters.

One-step approaches can be formulated by directly applying Markov chain Monte Carlo (MCMC) (e.g. Vrugt et al., 2009) to the complete hierarchical model. MCMC are sampling techniques that can cope with high-dimensional parameters. However, they must be modified to account for the hierarchical structure by changing the equations for the probability of acceptance of the proposal distribution (e.g. Malinverno, 2002) which may not be straightforward for all types of hyperparameters. Modifying one hyperparameter such as the training image, for example, impacts the model in its whole, potentially leading to completely different likelihood, which is not desirable for convergence in MCMC. Only very recent advances have made possible the exploration of such complex joint spaces. In this regard, Arnold et al. (2019) and Demyanov et al. (2019) presented a framework based on the definition of a metric space for the geological scenario and a combination of global optimization and resampling, to approximate a thorough MCMC.

Two-step approaches are based on the factorization of the joint posterior distribution in the product of the posterior of the parameters given the hyperparameter and the (marginal) posterior of the hyperparameter, and perform inference separately for each factor (Neuman, 2003; Khodabakhshi and Jafarpour, 2013; Park et al., 2013). However, the factor corresponding to the (marginal) posterior of the hyperparameter involves a multidimensional integral which may be computationally demanding (Neuman, 2003). The focus of this work is in the computation of the (marginal) posterior of the hyperparameter, i.e. only the first step in a two-step approach while solving the complete inverse problem for the hierarchical model.

Regarding computational demand, it has been argued that two-step may be more efficient than one-step approaches because of their ability to discard or falsify certain values with a relatively cheap method (that does not require inference of the parameters) in the first step (e.g. Park et al., 2013; Hermans et al., 2015). This may be specially advantageous when considering a high number of discrete values or a continuous range of the hyperparameter. However, it is also possible that one-step methods, when designed to be efficient (e.g. Demyanov et al., 2019), could quickly discard values of the hyperparameter that are not consistent with data. Park et al. (2013) make a comparison of their method, a two-step

approach, with rejection sampling (which is used as a one-step method) and show that their method provides similar results with less computations of the forward model. However, rejection sampling is a very expensive method and, to the authors' knowledge, a comparison against more favorable one-step approaches has not been done yet. Such a comparison is, nevertheless, outside of the scope of the paper.

Different ways to handle the hyperparameter factor in a two-step approach have been proposed, especially within the context of Bayesian model averaging (BMA) (Hoeting et al., 1999). When the hyperparameter is discrete, the factorization strategy mentioned above is equivalent to applying BMA for the parameters. In BMA, the aforementioned multidimensional integral is usually approximated by using a Gaussian distribution for the parameter dimensions in the likelihood. This Gaussian distribution is centered on the maximum-likelihood parameters and computed for each value of the hyperparameter (the so-called Laplace approximation), and it is a common approach when using BMA in hydrogeology (Neuman, 2003; Ye et al., 2004; Li and Tsai, 2009). Therefore, the Laplace approximation requires the classical inverse problem to be solved once for each value of the hyperparameter (and would require more involved sampling in the case of continuous hyperparameters). Moreover, to be a higher-order approximation, it requires the evaluation of the Hessian with respect to the parameters. Both the maximum-likelihood estimation and the Hessian may require a significant number of simulations using a computationally expensive numerical model. For this reason, some studies (Li and Tsai, 2009; Tsai and Elshall, 2013) have relied on the fact that, when the number of data becomes large relative to the number of parameters, the Laplace approximation can be simplified and computed using the Bayesian information criterion (BIC) (Raftery, 1995), which does not require evaluation of the Hessian. However, this may not be the case for most problems in Earth sciences, where parameters are usually high-dimensional and data is sparse. Khodabakhshi and Jafarpour (2013) used the same factorization within a sequential Monte Carlo approach, where the hyperparameter factor is first computed with a mixture model and then used for adaptive sampling in an Ensemble Kalman Filter to update the parameters. However, since their method is embedded in the sequential approach, the hyperparameter factor cannot be computed separately, i.e. the hyperparameter factor at the final time cannot be computed without updating the parameters at each time step.

Considering the same factorization of the joint distribution, an alternative method to obtain the (marginal) posterior of the hyperparameter was proposed by Park et al. (2013) that copes with the disadvantages of the Laplace approximation but also computes the hyperparameter factor separately. In other words, their method works for low numbers of data (where BIC does not apply), retains a low computational demand and does not require previous inference on the parameters. Instead of aiming for a point-by-point match of data, a feature match would result in a similar (marginal) posterior distribution for the hyperparameter according to

the authors. Therefore, feature extraction techniques are needed to reduce the dimensions of data low enough so that statistical techniques (e.g. kernel density estimation) may be applied to a low number of Monte Carlo samples to approximate the (marginal) joint distribution of the hyperparameter and the features. This distribution is then evaluated at the features of the observed data to obtain the (marginal) posterior of the hyperparameter. No maximum-likelihood estimation of the parameters for each value of the hyperparameter or Hessian evaluation is needed, as opposed to the Laplace approximation. Feature extraction techniques may incur in some computational time depending on their complexity, but this is generally negligible compared to evaluations of the numerical model. Park et al. (2013) presented an example where they generate Monte Carlo samples of the joint distribution by numerical simulations of reservoir flow data, then disregard parameter dimensions and apply data dimension reduction together with kernel density estimation to approximate the (marginal) posterior distribution of the geological scenario (which is the hyperparameter in their case). As mentioned above, they showed that the method yields results similar to rejection sampling. Hermans et al. (2015) applied it for one discrete hyperparameter but with two different types of data: hydraulic heads and electrical resistivity tomography. Scheidt et al. (2015b) extended the approach to estimate the posterior distribution of a continuous hyperparameter. Scheidt et al. (2015a) follow the same approach but deal with seismic data and a wavelet-based method to reduce dimensions of this data. A major difficulty of this approach is that choosing between the different ways to extract features is not straightforward, and an objective assessment of all the possible choices of features is lacking. Moreover, applying the techniques involves some additional specific parameters whose values are not straightforward to optimize.

In this paper, we define and systematically compare the efficiency of a new range of features for the application of the Park et al. (2013) framework with geophysical data. As part of the features definition, we propose a cross-validation method to select the best feature and the parameters required by the framework that is based on performance scores of the newly designed data features.

We illustrate the proposed approach using near-surface geophysical data to derive posterior probability distributions of one discrete and one continuous hyperparameter.

2 Methodology

2.1 Hierarchical probabilistic model sampling

To deal with multi-level uncertainty problems typically present in Earth sciences we propose to build a Bayesian hierarchical model to explicitly consider the relations between all parameters and data. The probabilistic model considered in this study can be represented as the directed acyclic graph (DAG; Bishop, 2006) shown in

Fig. 1, where each random variable is represented by an open node and relations of conditional dependency are represented by directed arrows.

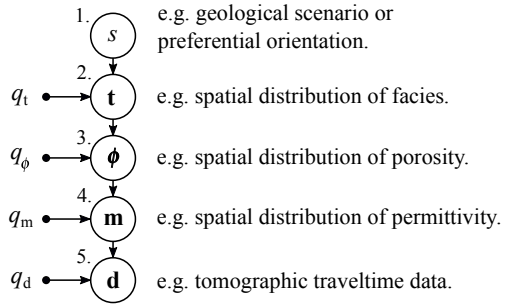

Figure 1: Graphical model for the proposed Bayesian hierarchical model. s stands for the structural parameter, $\mathbf{t}$ is the field of geological facies, $\boldsymbol{\phi}$ is the field of a physical property, $\mathbf{m}$ is the field of a geophysical property, $\mathbf{d}$ is the geophysical data. The q's are the fixed variables required at each step. On the right side, examples of each variable.

In this graphical model, the hyperparameter s at the top controls the structure of spatial parameters in lower levels, therefore we will refer to it as structural parameter. Geophysical data or observations are in the lowest level of this model. Indeed, as implied by the conditional relations of the graph and given its spatially-distributed nature, geophysical data provide a means to reduce uncertainty in spatial heterogeneity. Our objective is to compute the posterior distribution of structural parameters given the geophysical data $p(s|\mathbf{d})$, which can be obtained by considering the corresponding marginal distribution $p(s, \mathbf{d})$ of the joint probability distribution $p(s, \mathbf{t}, \boldsymbol{\phi}, \mathbf{m}, \mathbf{d})$. The DAG implies that the joint probability distribution can be factorized as

$$p(s, \mathbf{t}, \boldsymbol{\phi}, \mathbf{m}, \mathbf{d}) = p(s)p(\mathbf{t}|s)p(\boldsymbol{\phi}|\mathbf{t})p(\mathbf{m}|\boldsymbol{\phi})p(\mathbf{d}|\mathbf{m}) \tag{1}$$

where s stands for the structural parameter, $\mathbf{t}$ is the field of geological facies, $\boldsymbol{\phi}$ is the field of a physical property, $\mathbf{m}$ is the field of a geophysical property, $\mathbf{d}$ is the geophysical data and $p(\cdot|\cdot)$ expresses a conditional probability distribution.

To approximate the joint distribution from Eq. (1) we use Monte Carlo sampling. Since our probabilistic model is represented by a DAG we can obtain samples of the joint distribution by ancestral sampling, i.e., sampling following an order determined by the arrows in Fig. 1. Hence, when sampling a certain node, all nodes pointing to it (termed parent nodes) must be already sampled. Fixed variables that may be required in each sampling step are usually represented as black dots, e.g. the specified noise level in the data used in the last step is included in

q_d (Fig. 1). Once the samples of the joint distribution are obtained we disregard parameters (or dimensions) other than the ones in the marginal of interest, $p(s, \mathbf{d})$. In our implementation, each step is given by (for numbering, refer to Fig. 1):

1. The structural parameter, s, is sampled from either a uniform distribution or a discrete uniform distribution.

2. The geological heterogeneity is represented by a spatially discretized facies field, $\mathbf{t}$. In our case, we consider this field as generated by a stochastic process, such that sampling from $p(\mathbf{t}|s)$ gives a categorical random field defined either by multiple-point statistics or truncated sequential gaussian simulation.

3. The physical property field $\boldsymbol{\phi}$ requires a probability distribution $p(\boldsymbol{\phi}|\mathbf{T} = \mathbf{t})$. If no uncertainty is assumed at this step, then only a relation that assigns a value of the physical property to each facies is used.

4. The geophysical property field $\mathbf{m}$ is obtained by using a petrophysical relation which may also be formulated as a probability distribution $p(\mathbf{m}|\boldsymbol{\Phi} = \boldsymbol{\phi})$.

5. Finally, the geophysical data $\mathbf{d}$ is the result of a geophysical forward operator $g(\mathbf{m})$ and formulated as $p(\mathbf{d}|\mathbf{M} = \mathbf{m})$. Note that this is just the likelihood function defined in the non-hierarchical inverse problem of geophysical data.

Performing ancestral sampling N times according to the DAG of this model (Fig. 1)—i.e., sampling sequentially each conditional probability of the factorized joint distribution in Eq. (1)—will output N samples of the joint probability distribution.

2.2 Designing data features to inform on structural parameters

Given the described process to sample the hierarchical probabilistic model, we notice the data $\mathbf{d}$ are dependent not only on the higher-level structural parameters s but also on intermediate-level parameters. Here, we design features $f(\mathbf{d})$ from the data $\mathbf{d}$ to retain information related only to the structural parameters s and to reduce the dimensionality of the problem. As mentioned in the Introduction, this reduced dimensionality is required to make the use of statistical techniques—such as kernel density estimation (KDE)—computationally tractable. This implies we will approximate the marginal distribution as $p(s|\mathbf{d}) \approx p(s|f(\mathbf{d}))$. We will consider feature extraction as any function $f(\mathbf{d})$ that maps $\mathbf{d}$ from a space of dimension N_d (the number of data points) to a lower dimensional space of dimension N_f—this would also entail function compositions, e.g. $f(\mathbf{d}) = \psi \circ \xi = \psi(\xi(\mathbf{d}))$ where ψ and ξ are functions. The vector of features will be denoted as $\mathbf{f} = f(\mathbf{d})$ and is of dimension N_f. Ideally, feature extraction of data should (1) retain all information

regarding the structural parameter (be informative), and (2) disregard information not related to it (dimension reduction).

A first approach for feature extraction consists in using dimension reduction techniques, or so-called data-driven approaches also referred to as continuous latent variables (Bishop, 2006), which aim to retain as much variability of the original data as possible but with a low dimensional representation of the data. In our study, we consider principal component analysis (PCA) and multidimensional scaling (MDS). PCA is based on the eigendecomposition of the data covariance matrix—the eigenvectors represent orthogonal directions following an order of maximum variability and the corresponding eigenvalues state the magnitude of this variability. By disregarding eigenvectors, PCA can be used as a linear dimension reduction method (it is only based on rotation and scaling operations). MDS takes dissimilarities (or distances) between data samples as input and then maps these samples in a lower-dimensional space by approximating the original distances. This may be achieved by optimizing a so-called stress function—a method which is referred to as Scaling by MAjorizing of a COmplicated Function (SMACOF) (De Leeuw and Heiser, 1980). In this way, MDS works as a non-linear dimension reduction method. When using MDS, one can also choose distance functions that are more suited to state the dissimilarity of interest (Scheidt and Caers, 2009). Note that in practice, mapping back to the original distribution is not exact because we disregard some information by considering only the first components of a decomposition for PCA or by retaining only relative distance between samples for MDS.

A second approach consists in designing $f(\mathbf{d})$ to extract specifically information linked to the structural parameters s using domain knowledge, leading to the so-called insight-driven features (Morzfeld et al., 2018). For instance, Hermans et al. (2015) applied an insight-driven approach favoring a combination of inversion and multidimensional scaling (MDS) to extract relevant features for the geological scenario, while Scheidt et al. (2015a) used a wavelet transform on seismic reflection data in combination with an L^2-norm distance as insight-driven feature to update different uncertain geological parameters. Since these functions depend on the specific combination of structural parameter s and data $\mathbf{d}$, they will be detailed in the following sections. As previously mentioned, in our case $\mathbf{d}$ are ground-penetrating radar (GPR) traveltime data collected in cross-borehole tomographic mode. Note that this could also apply to seismic traveltime.

In this work, when using insight-driven features we always consider their combination with data-driven approaches, i.e. we apply first an insight-driven approach and then use data-driven techniques to further reduce dimensionality while retaining most information (this further reduction in dimensions is to enable the application of kernel density estimation as will be explained below). As a result, all of our features may be considered within the so-called metric space modelling (Park et al., 2013; Scheidt and Caers, 2009). Whether using a data-driven approach or a composition of insight-driven and data-driven approaches, we will refer to the

number of dimensions after feature extraction as N_f.

2.2.1 Extracting data features for discrete geological scenario

While considering the uncertainty of different geological scenarios formulated as a discrete structural parameter s, we propose extracting features from tomographic data in six different ways summarized in Table 1. The first two approaches use dimension reduction techniques (PCA and MDS) on the traveltime directly, and the remaining use dimension reduction techniques but only after an initial insight-driven transformation. The third and fourth approaches transform the data using a histogram and the last two rely on an inverse transform of the data (tomogram).

Case	Insight-driven	Data-driven	distance
PCA_t	-	PCA	-
MDS_t	-	MDS	euclidian
PCA_h	histogram	PCA	-
MDS_h	histogram	MDS	Jensen-Shannon
MDS_v	smooth inversion	MDS	euclidian
MDS_c	smooth inversion and connectivity	MDS	euclidian

Table 1: Different feature extraction cases proposed for geological scenario from top to bottom: PCA on data (PCA_t); MDS with euclidian distance on data (MDS_t); PCA on histograms of traveltime data (PCA_h); MDS with a Jensen-Shannon distance function on histograms of traveltime data (MDS_h); MDS with euclidian distance on geophysical images obtained by regularized inversion of traveltime data (MDS_v); MDS with euclidian distance on connectivity curves obtained from geophysical images (MDS_c).

The targeted discrete structural parameter s is implicitly linked to the connectivity of the medium, i.e. each scenario implies the use of a geostatistical algorithm with defined inputs that is expected to produce different degrees of connectivity (Figs. 2 and 3a). The histogram transformation for cases PCA_h and MDS_h was chosen because differences in connectivity are expected to cause different distributions of traveltimes. For example, if the system is well-connected, the histogram of the traveltimes will show high values for the bins in faster traveltimes and also a multi-modal distribution. This can be observed on Fig. 3c (top and bottom). Indeed, the ray paths follow complex patterns for different source-receiver offsets which may be described as the ray "jumping" from one high velocity to another high velocity object, if a high number of jumps occurs the histogram of traveltimes will tend to be smooth, on the other hand if a low number of jumps occurs

the histogram will display multi-modality. To estimate the distance between two histograms or probability distributions, we used the Jensen-Shannon distance. As suggested by (Scheidt et al., 2015b), the Jensen-Shannon distance (or the square root of the Jensen-Shannon divergence) is an appropriate metric to measure the distance between two probability distributions or, as in our case, their approximations in the form of histograms. We note that the choice of metric must be made in order to better discriminate the parameter of interest by means of the features so far extracted from the data.

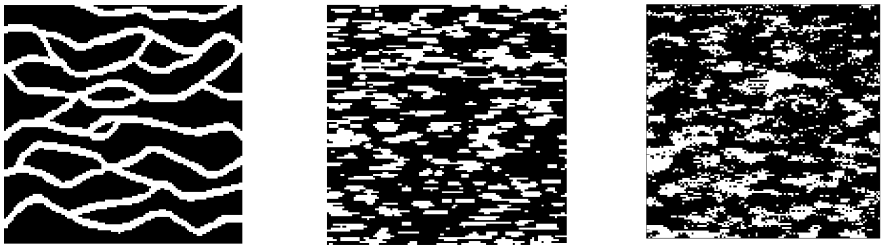

Figure 2: Three different geological scenarios considered: (a) and (b) are training images used for multiple-point geostatistics simulations, and (c) is a realization of a truncated Gaussian simulation with its anisotropic variogram fitted to the training image in (a).

Connectivity may also be quantified if one has access to the knowledge of the spatial distribution of the facies which, in the case of geophysical traveltime data, can be easily approximated using a deterministic inversion (Fig. 3d). To quantify the connectivity, we used the Euler characteristic curve in case MDS_c (Renard and Allard, 2013) by thresholding the inverted velocities in 100 steps (see Fig. 3e). In other words, we obtain the range of velocity values on each inverted "image" and divide it in 100 intervals, then use the upper bound of each interval to get a binary "image" (i.e. all values lower than the upper bound are set to 1 and the remaining to 0) and compute the Euler characteristic for each of these binary images. The result is then a 100-dimensional vector that is a discrete version of the Euler characteristic curve. The Euler characteristic is a topological characteristic and for binary images is equal to the number of objects (or clusters) minus the number of holes in such objects (Renard and Allard, 2013). For comparison, we also used directly the inverted images in case MDS_ν.

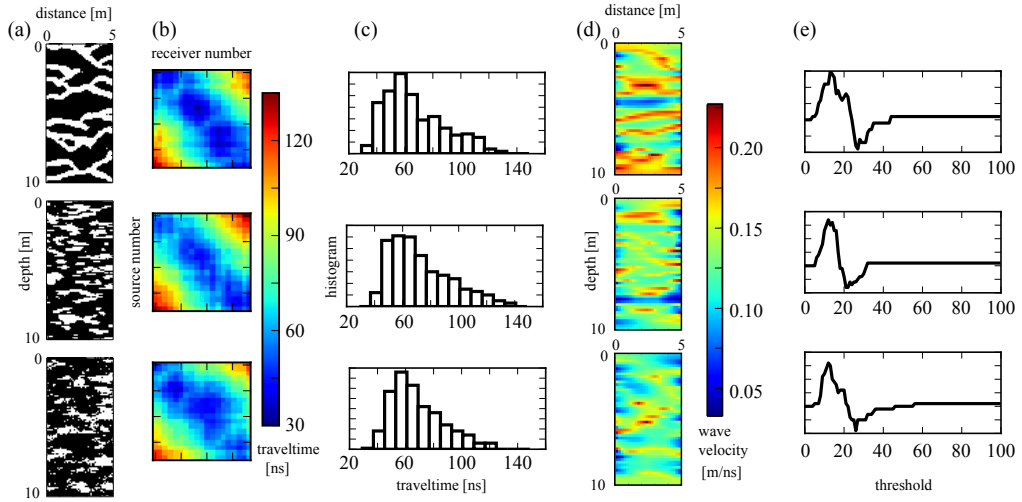

Figure 3: Realizations of geological facies with proposed features for the discrete structural parameter: (a) facies samples, (b) simulated traveltimes (reciprocal data not shown), (c) histogram of simulated traveltimes, (d) deterministic inversion of simulated traveltimes, and (e) Euler characteristic curves.

2.2.2 Extracting data features for continuous channel orientation

The targeted continuous structural parameter s is again linked to the connectivity of the system but here quantified by the orientation of the connectivity rather than the degree of connectivity. In this case, we propose four feature extraction cases (Table 2) in addition to the dimension reduction techniques: two based on what we call "oriented averages" of traveltime and two based on tomograms (inverted velocities).

Case	Insight-driven	Data-driven	distance
PCA_t	-	PCA	-
MDS_t	-	MDS	euclidian
PCA_a	oriented averages	PCA	-
MDS_a	oriented averages	MDS	euclidian
MDS_v	smooth inversion	MDS	euclidian
MDS_R	smooth inversion and Radon transform	MDS	euclidian

Table 2: Different feature extraction cases proposed for preferential orientation. From top to bottom: PCA on data (PCA_t); MDS with euclidian distance on data (MDS_t); PCA on oriented averages of traveltime data (PCA_a); MDS with euclidian distance on oriented averages of traveltime data (MDS_a); MDS with euclidian distance on geophysical images obtained by regularized inversion (MDS_v); MDS with euclidian distance on a Radon transform of geophysical images (MDS_R).

The oriented averages in cases PCA_a and MDS_a were proposed to inform on the orientation of the channel by computing the average of traveltime data in all possible orientations of source-receiver combinations. For the oriented averages in our synthetic setup (described below) we get 37 orientations, hence a vector of 37 insight-driven features.

The Radon transform in the MDS_R case is a line integral transform that is equivalent to a linear tomography taken at constant offsets and in a series of directions (Durrani and Bisset, 1984). It has been used to extract orientation information of images (see e.g. Aydin and Caers, 2013) and we compute it considering eight different directions $\{0, \pi/6, \pi/4, \pi/3, \pi/2, 2\pi/3, 3\pi/2, 4\pi/3\}$ in radians. For comparison, we also used directly the inverted images in case MDS_ν.

2.3 KDE and cross-validation approach

We chose to apply kernel density estimation (KDE) to approximate the marginal distributions $p(s, f(\mathbf{d}_{obs}))$ and $p(\mathbf{d}_{obs})$ using the features samples (obtained by applying the transformations of the previous sections to the Monte Carlo data samples) to compute the posterior $p(s|f(\mathbf{d}_{obs}))$. Heteroscedasticity may arise due to the nature of the structural parameter or be induced by the transformations of the feature extraction. To handle such heteroscedasticity, we use an adaptive version of KDE that is based on clustering of the samples similar to the ones proposed by Park et al. (2013) and Scheidt et al. (2015b). As a result, our implementation takes the following form

$$p(s|f(\mathbf{d}_{obs})) = \frac{p(s, f(\mathbf{d}_{obs}))}{p(\mathbf{d}_{obs})} = \frac{\sum_{i=1}^{N_c} \sum_{s_j, \mathbf{d}_j \in C^{(i)}} K_{H_s}^{(i)}(s - s_j) K_{H_f}^{(i)}(f(\mathbf{d}_{obs}) - f(\mathbf{d}_j))}{\sum_{i=1}^{N_c} \sum_{s_j, \mathbf{d}_j \in C^{(i)}} K_{H_f}^{(i)}(f(\mathbf{d}_{obs}) - f(\mathbf{d}_j))} \tag{2}$$

where N_c is the number of clusters used in the clustering algorithm, $C^{(i)}$ refers to the i-th cluster from the set $\{C^{(i)}|i = 1, ..., N_c\}$, s_j and $\mathbf{d}_j$ are the values for the structural parameter and the data for the j-th sample, therefore the index $j = \{1, ..., N\}$, $K_H^{(i)}(\cdot)$ refers to a scaled kernel function with corresponding bandwidths H_s for the structural parameter and H_f for the data whose values depend on which cluster $C^{(i)}$ they belong to, and $\mathbf{d}_{obs}$ is the observed data. Further details on adaptive kernel density estimation and our particular implementation are presented in the Appendix. What is important to note here is that the bandwidths H_s and H_f are parameters controlling the shape or "smoothing" of the distribution in the joint space $p(s, f(\mathbf{d}))$ and they are implicitly given by N_c.

As previously mentioned regarding the possible heteroscedastic character of the posterior distribution of the structural parameter, adaptive KDE was chosen because it (1) accounts for the degree of uncertainty as a function of the structural parameter s and (2) adjusts the error model in the feature space (i.e. non Gaussian). In the latter case, the noise model for the data is no longer valid for the features. Instead of handling this using "perturbed" observations (Hermans et al., 2016; Morzfeld et al., 2018), adaptive KDE can deal with this directly because it works for heteroscedastic and multimodal distributions.

At this point we should note that our methodology results in three main degrees of freedom, namely the number of Monte Carlo samples N (section 2.1), the number of dimensions N_f after feature extraction and dimension reduction (section 2.2) and the number of clusters N_c (this section and the Appendix) used in the adaptive KDE. Because the evaluation of the numerical model is usually the most computationally demanding step, a low value of samples N should be chosen. Then, N_f and N_c should be chosen so that the method performs optimally. To choose this optimum, we propose a leave-one-out cross-validation approach with two different scores depending on the type of structural parameter being estimated. For discrete parameters, N_f and N_c can be fixed by using the number of correct classifications obtained by assigning the scenario with the highest (marginal) posterior probability at the data sample. In case of equal number of correct classifications, we take the mean of all the (marginal) posterior probabilities of the correctly classified scenarios, termed here as ℓ_d, and pick the one with the highest value (Hermans et al., 2015). For continuous parameters, the proposed cross-validation approach is based on a likelihood score defined by (Habbema et al., 1974) as

$$\ell_c = \frac{1}{n} \sum_{j=1}^{n} \ln p_{-i}(N_f, N_c) \tag{3}$$

where p_{-i} stands for the leave-one-out estimate of the conditional distribution

$p(s|f(\mathbf{d}_i))$, i.e., the probability value computed at the i-th point without considering the same point in the adaptive KDE.

We compare our cross-validation approach to the silhouette index proposed by Scheidt et al. (2015b), in a simple one-dimensional example (Fig. 4) of applying adaptive KDE when the data error model is Gaussian and we aim to estimate its probability density but we can only work with features (e.g. a non-linear feature, like the exponential $f(x) = e^x$ in Fig. 4) as in our approach. Since it is not easy to visually discern which curve gives a better approximation, we used the Jensen-Shannon distance 2.2.1 to measure the distance with respect to the true distribution. Results show that the adaptive KDE would better approximate the original error model in this feature space when the number of clusters, $N_c = 4$, is estimated through cross-validation, which generally produces an optimal bias-variance tradeoff, instead of the result $N_c = 2$ obtained with the silhouette index (Fig. 4). In our case, the probability distribution to be approximated is $p(s, f(d))$ instead of $p(e^x)$ and $p(s, d)$ would be in place $p(x)$. Another advantage of cross-validation is that it can always be applied since we can always generate the necessary Monte Carlo samples. Given high-dimensional and more complex distributions, we expect the use of cross-validation will be more beneficial.

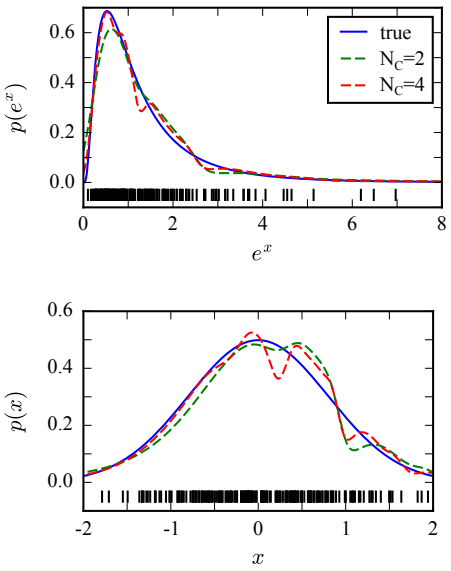

Figure 4: Comparison of the adaptive KDE for a non-linearly transformed space when the number of clusters is chosen by silhouette index (N_c=2) and by cross-validation (N_c=4). Vertical markers in the lower part denote the samples used to approximate the probability distribution. Both plots show the same samples, with the upper one representing an exponential transformation of the values in the lower one.

3 Reducing structural uncertainty using features of GPR traveltime on a synthetic model

3.1 Model set-up

A synthetic case is presented in this section to demonstrate the proposed methodology using GPR traveltime as data $\mathbf{d}$. The spatial domain is a vertical section between two boreholes separated 5 m from each other and whose depth is 20 m (3). As in the usual tomographic survey, data is generated by considering the sources are in one borehole while receivers are located in the other. Afterwards, reciprocal data is simulated by placing sources in the borehole where receivers were firstly placed and vice-versa. Vertical separation of both the receivers and sources is constant and equal to 0.5 m. We consider 19 sources and 19 receivers (and the same number for reciprocal data) where the first position of the receivers/sources is 0.5 m from surface and last is 19.5 m.

In our specific synthetic demonstration we study two cases, one including a discrete structural parameter and the other a continuous one, for which we describe steps 1 to 5 in Fig. 1. An outline of the hierarchical sampling for the discrete structural parameter is presented in Fig. 5. Note that steps 3 to 5 are common for both types of structural parameters.

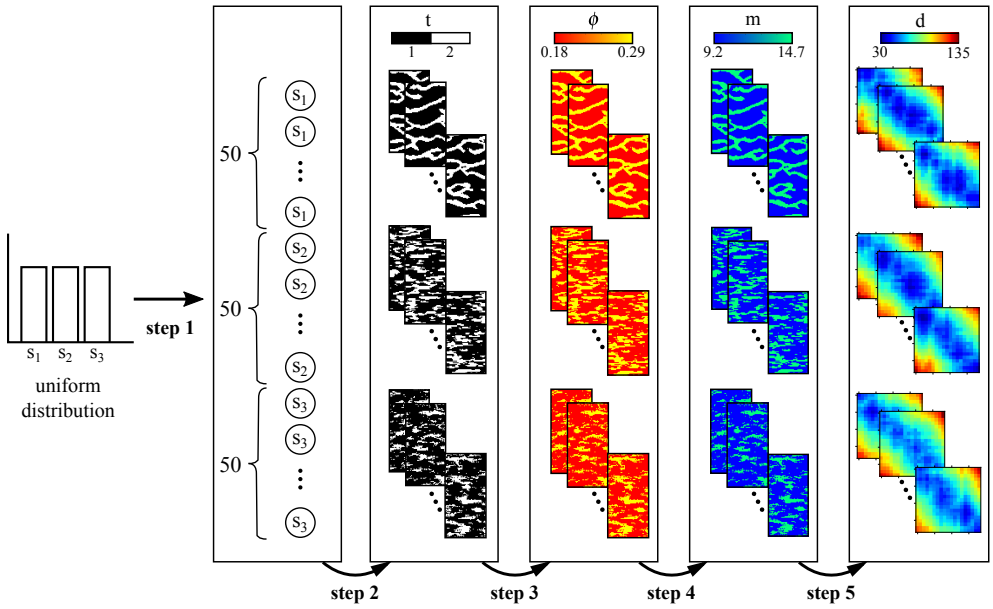

Figure 5: Sketch of the hierarchical sampling process for the geological scenario where s is the geological scenario index, t is the facies index, ϕ is the porosity, m is the relative dielectric permittivity and d are electromagnetic wave traveltimes in nanoseconds plus added Gaussian noise with $\sigma = 1.4\ ns$. Steps 1 through 5 are described in the text.

Step 1. The discrete structural parameter $s \in \{s_1, s_2, s_3\}$ denotes three different geological scenarios, represented by three different geostatistical models: two multiple-point geostatistics models with different training images and one truncated gaussian field model (Fig. 2). Each row in Fig. 5 corresponds to a value of s. The prior $p(s)$ is a discrete uniform distribution and we consider this implicitly by using 50 samples of each value $\{s_1, s_2, s_3\}$ for a total of 150 samples. These 150 samples are used in the following steps.

The continuous structural parameter is the preferential orientation of the geological patterns (channels in our case) and its range is $s \in (0, \pi)$. 200 samples were obtained from a uniform distribution with range $(0, \pi)$. This range was chosen because the training image used (see realizations in Fig. 9a) has a rotational symmetry of order two, i.e. data realizations from s and $s + \pi$ can be considered coming both from s only.

Step 2. In the discrete case, we obtain facies samples $\mathbf{t}$ from s_1 and s_2 by means of multiple-point geostatistics sequential simulations (two first rows for the t column in Fig. 5), considering training images ti_1 and ti_2 (Fig. 2), respectively. The $\mathbf{t}$ samples from s_3 are generated by truncated sequential gaussian simulation, whose anisotropic spherical variogram was obtained by fitting to the training image ti_1 (last row for the t column in Fig. 5). The samples of the continuous s, are

used as input to generate samples of $\mathbf{f}$ by multiple-point geostatistics simulations using training image ti_1.

Step 3. The porosity $\boldsymbol{\phi}$ is given by a constant mapping $\mathbf{q}(\mathbf{t})$ of the facies and the probability distribution can be expressed as

$$p(\boldsymbol{\phi}|\mathbf{T}=\mathbf{t}) = \delta(\boldsymbol{\phi} - \mathbf{q}(\mathbf{t})) \tag{4}$$

where δ is the delta function and

$$\mathbf{q}(\mathbf{t}) = \begin{cases} q_1 & t = 1 \\ q_2 & t = 2 \end{cases} \tag{5}$$

where $q_1 = 0.18$ and $q_2 = 0.29$ are porosity values for two different geological facies. This amounts to assigning a porosity value for each facies (the ϕ column in Fig. 5), but we choose to express it as a conditional probability to be consistent with the Bayesian hierarchical model, where uncertainty may be included at this step to consider e.g. intrafacies variability.

Step 4. We choose a mixing model named CRIM (Birchak et al., 1974) to transform the porosity field into a dielectric permittivity field (the m column in Fig 5). Such transformation is denoted by $\mathbf{r}(\boldsymbol{\phi})$ and the corresponding probability distribution is

$$p(\mathbf{m}|\boldsymbol{\Phi}=\boldsymbol{\phi}) = \delta(\mathbf{m} - \mathbf{r}(\boldsymbol{\phi})) \tag{6}$$

again δ is the delta function and

$$\mathbf{r}(\boldsymbol{\phi}) = ((1-\boldsymbol{\phi})\sqrt{\epsilon_s} + \boldsymbol{\phi}\sqrt{\epsilon_w})^2 \tag{7}$$

where $\epsilon_s = 3$ is the permittivity of the solid grains and $\epsilon_w = 81$ is the permittivity of water. In this way, the facies $t = 1$ will have lower permittivity (therefore, higher electromagnetic wave velocity) than the facies $t = 2$.

Step 5. Numerical modeling of the electromagnetic wave traveltime is done by a ray-path approximation model, as implemented in PyGIMLi's Refraction module (Rücker et al., 2017). Note this approximation reduces computational demand compared to full-waveform simulation. Interestingly, within a feature-based framework, traveltime data can be seen as a first feature extraction step from the full-waveform data. The corresponding probability distribution is

$$p(\mathbf{d}|\mathbf{M}=\mathbf{m}) \sim \mathcal{N}(g(\mathbf{m}), I\sigma^2) \tag{8}$$

where $\mathcal{N}$ stands for a multivariate normal distribution, I is an identity matrix of size N_d, $g(\cdot)$ is the geophysical forward operator given by the numerical model mentioned above, and $\sigma = 1.4\ ns$ states the magnitude of independent normally-distributed noise in the geophysical data. Simulated traveltimes data are shown in data arrays (where columns represent the receiver index and rows the source

index) in the d column of Fig. 5. Note that uncertainty was not considered in steps 4 and 5 here but could easily be included.

We generate samples of the (marginal) joint distribution $p(s, \mathbf{d})$ by following steps 1 to 5 and disregarding the parameter dimensions.

3.2 Results for a discrete structural parameter

We extract features of traveltime data to approximate the posterior distribution of the structural parameters $p(s|\mathbf{d}) \approx p(s|f(\mathbf{d}))$ according to the six different cases mentioned in Section 2.2.1. Fig. 3 shows one realization for each value of the discrete structural parameter, the simulated traveltime data and the corresponding insight-driven features.

Cross-validation was used to select the number of dimensions, N_f, and the number of clusters, N_c, for each one of these cases. We restricted to values $N_f \leq 10$ and $N_c \leq 15$ since the number of samples needed to obtain a good estimate with KDE beyond this bound would be too high. The cross-validation score used was the number of correctly classified realizations, i.e. an integer between 0 and 150, recalling we generated 50 samples for each value of the discrete structural parameter. Fig. 6 shows the cross-validation matrix obtained for the case MDS_h where we can see there is an optimum choice of N_f and N_c that is within our chosen search limits for both parameters. In the cross-validation matrix, we see a counterbalancing of N_f and N_c: within the bound $N_f \leq 8$, the classification maxima for increasing N_f generally correspond to lower values of N_c. Since the same number of samples is considered, this may be explained because lower values of N_c mean the adaptive KDE is using wider bandwidths when going into higher dimensional spaces, effectively covering more space in the density estimation than with a higher N_c. However, the effect of N_f is stronger and leads to better classification, which is also an indication of a properly chosen feature extraction to reduce dimensions. For this reason, in case of the same performance, we rather choose the combination where N_f is lower. Note also that an arbitrary chosen combination of N_c and N_f could easily lead to a significantly lower performance of the approach, highlighting the need to optimize the choice of those degrees of freedom.

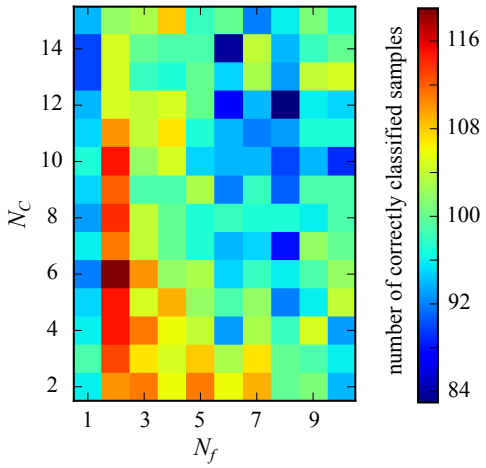

Figure 6: Cross-validation matrix for the case MDS_h of the discrete structural parameter (geological scenario).

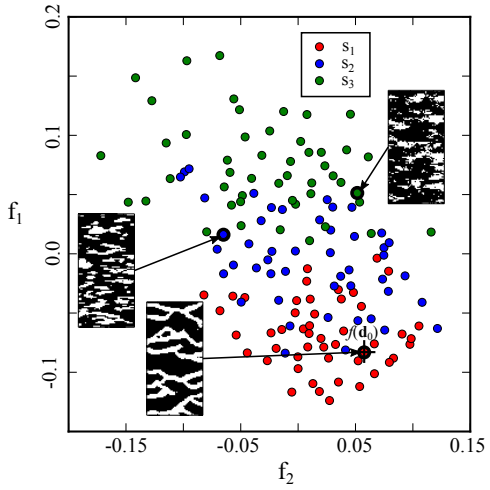

Figure 7: MDS applied on histograms of traveltime data. Examples of realizations for each value of the discrete structural parameter s are shown.

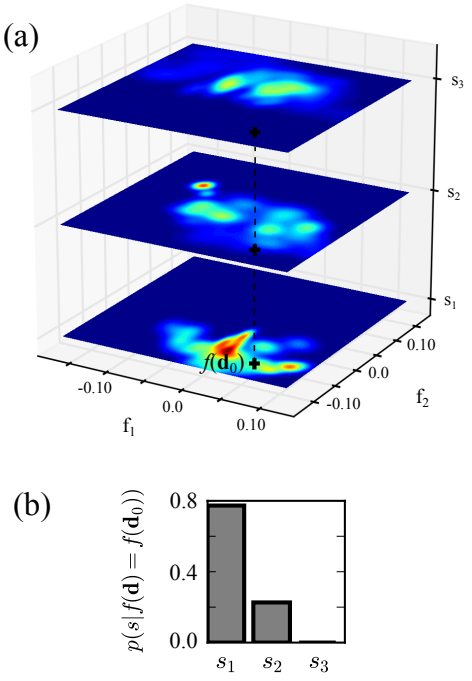

Figure 8: (a) Joint probability distribution $p(s, \mathbf{f})$ for the case of MDS on histograms of traveltime data. The '+' denotes one realization $\mathbf{d_0}$ when the discrete parameter $s = s_1$ and is the same shown in Fig. 7. (b) The posterior probability of the structural parameter s obtained by cross-validation when $\mathbf{d} = \mathbf{d_0}$.

Fig. 7 shows the MDS mapping applied to the histograms of traveltime data in the low dimensional feature space for $N_f = 2$ (the optimum selected by our approach). Points are approximately separated according to the three values of the discrete structural parameter s which means the features are informative on this structural parameter. The joint probability distribution $p(s, f(\mathbf{d})) = p(s, \mathbf{f})$ obtained through adaptive KDE is shown also for this case MDS_h (Fig. 8a). The estimation of the posterior probability of the structural parameter for one data sample d_0 with known true value of $s = s_1$, equivalent to one computation of the leave-one-out cross-validation is shown in Fig. 8b, where we see the method correctly gets the value s_1 as the most likely for d_0. We also note here that the probability of s_3 is very close to zero. If d_0 were measured geophysical data, this geological scenario would be falsified and could be left out of further analysis (e.g. inversion for spatial parameters).

For the other cases, a complete visualization is difficult due to the higher dimensionality of N_f but a summary of the results are shown in Table 3. Some cases

	N_f	N_c	$class$	ℓ_c
PCA_t	8	3	99	0.61
MDS_t	6	14	100	0.66
PCA_h	3	3	108	0.61
MDS_h	2	6	119	0.69
MDS_v	2	7	102	0.62
MDS_c	6	4	109	0.65

Table 3: Cross-validation results for discrete structural parameter s where $class$ refers to the number of correct classifications.

show a higher number of correctly classified samples (66% correctly classified for the worst case and 80% for the best one) but with different values for N_f and N_c. Also, the values of mean updated probability ℓ_d are higher for certain cases but to a lesser degree than for the number of correct classifications. The best performing strategy is the composition of MDS on histograms of traveltime (MDS_h). This means our proposed insight-driven feature has indeed aided to some extent in retaining information only on the structural parameter s. The connectivity-based approach (MDS_c) does not perform better than the data-driven approach. However, it is more discriminating than the tomograms (MDS_v). Those approaches are less effective in terms of computational demand, since they require both a deterministic inversion and computation of the Euler characteristic curves for each realization. This result might appear counter-intuitive as imaging is generally appealing for the human eyes and a common result of geophysical exploration. However, inversion can be considered as a feature extraction of data leading to loss of information related to the regularization operator. We note, however, that these results are related to the type of data (cross-hole GPR traveltime, in our case) and might differ for other data or even other acquisition setups. For instance, surface ERT data has been shown to be extremely sensitive to shallow resistivity structure hence a possible strategy is to extract features from the geophysical image rather than directly from the data or to develop more appropriate insight-driven features (Hermans et al., 2015).

We note a small improvement on the classification scores between PCA, a linear dimension reduction method, and MDS, a non-linear dimension reduction method. This may be explained as MDS being able to account for some non-linearity in the relation of the structural parameter with the data. Also, we see that a higher dimensionality is chosen (through cross-validation) for PCA in comparison with MDS, which may be because both methods are able to retain similar information but with different N_f.

3.3 Results for a continuous structural parameter

As previously mentioned six different cases are considered in which both data-driven and a composition of insight-driven with data-driven features are used (section 2.2.2).

The number of clusters N_c and the number of dimensions N_f was selected according to cross-validation using the minimum value for the score of Eq. (3) (third column in Table 4). Again, we restricted to $N_f \leq 10$ and $N_c \leq 15$. The chosen number of dimensions for the case PCA_t is $N_f = 3$ so, in order to represent the complete space where the method is applied, we would have to use three dimensions. However, for visualization purposes, we use the first two and show the distribution of realizations of features of the data (Fig. 9a). Here, the insets display four samples of the corresponding geological facies for which the simulated data and the PCA features were obtained. We clearly see that points are arranged according to the value of the structural parameter s which means that they are informative of it. Moreover, the distribution of samples reveals that the obtained features are probably linearly related to the structural parameter since they plot close to a circle and orientation is circular (i.e. periodic). Indeed, if we take this into account and plot the orientation versus the angle formed by the two features we see a linear trend (Fig. 9b). The scatter plot reveals a small degree of heteroscedasticity for this specific dataset (higher variance around 0.25π and 0.75π and lower variance around 0.5π and 0) which is also present for the other cases (MDS_t, PCA_a and MDS_a). However, due the small number of samples (200), this may not be statistically significant, therefore the process was repeated with 500 samples where the change in spread as a function of the orientation is clearer (not shown). This means cross-borehole GPR data is more discriminative in angles close to $0°/180°$ and $90°$ and is less discriminative for angles close to $45°$ and $135°$. This could be physically explained by the fact that changes in the length of the wave path through low velocity zones are greater when the angles are close to $0°/180°$ or $90°$. Further analysis is required to validate this conclusion, e.g. prove that the chosen dimension reduction techniques did not affect the results.

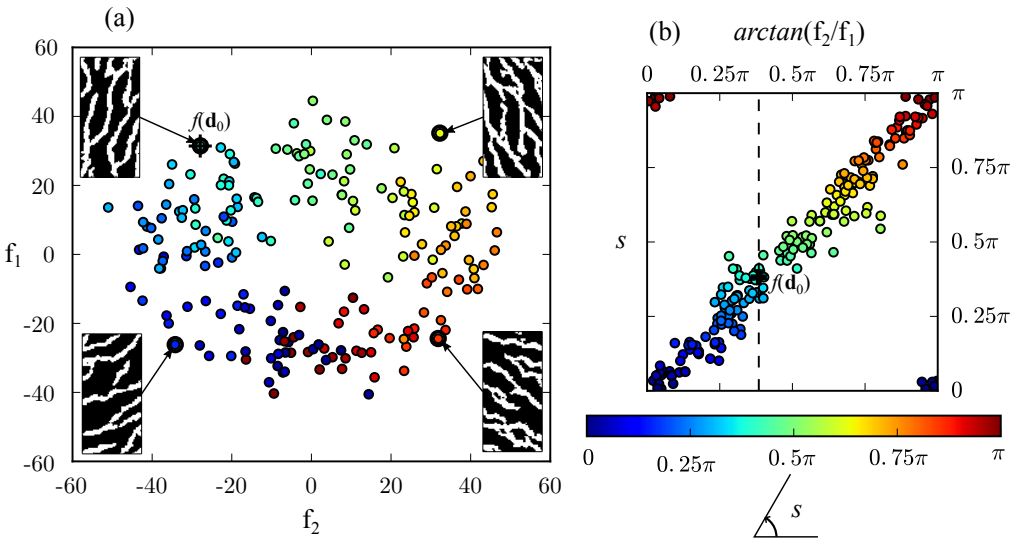

Figure 9: (a) PCA applied on traveltime data for continuous structural parameter s. (b) Same samples as (a) but computing the angle formed by the two features and plotting versus true orientation. Colors are true values for the structural parameter s (preferential orientation). $\mathbf{d_0}$ denotes a particular sample taken out during cross-validation and the dashed line denotes the position in the feature axis for this sample.

For the case PCA_t, Fig. 10a shows the distribution of features of the data together with the continuous structural parameter s and Fig. 10b shows the marginal distributions of the corresponding three-dimensional joint probability distribution $p(s, f_1, f_2)$. In order to apply the adaptive KDE to this circular parameter the bandwidth for the structural parameter dimension was computed in a transformed space (i.e. a two-dimensional space with $x = sin(s)$ and $y = cos(s)$) and the periodicity was accounted for by means of replication of samples in the boundaries (Silverman, 1986). For the other three cases, the method works similarly but its application is harder to visualize given the high number of dimensions N_f selected.

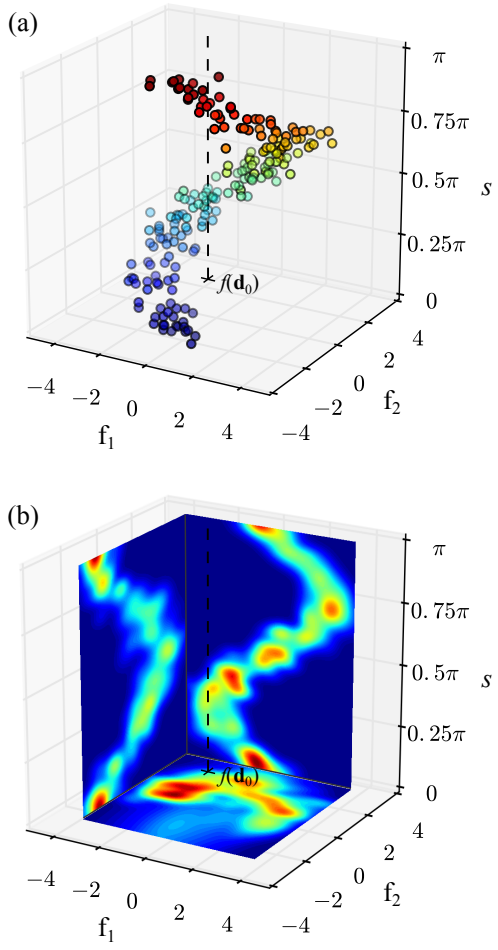

Figure 10: PCA applied on traveltime data showing structural parameter s as third dimension (a) colors are the same as in Fig. 9. Marginal distribution resulting from the application of adaptive KDE (b). The '+' denotes the sample $\mathbf{d_0}$ taken out during cross-validation and is the same as the one referenced in Figs. 9 and 11. The dashed line shows the conditioning to $\mathbf{d_0}$ in $p(s|\mathbf{d_0})$, therefore highlights the direction along which the adaptive KDE is applied.

Since we are dealing with a continuous parameter, the posterior probability distribution is also continuous. The process of building this distribution is depicted in Fig. 10 and the resulting posterior probability distribution for a certain value $\mathbf{d_0}$—taking its value out in the adaptive KDE while performing leave-one-out cross-validation—is shown in Fig. 11. We clearly see that the posterior contains the true value and it is sharply peaked around it which means the method is

correctly estimating the structural parameter s. Given that the prior distribution was uniform, the achieved reduction of uncertainty is on the order of 75%.

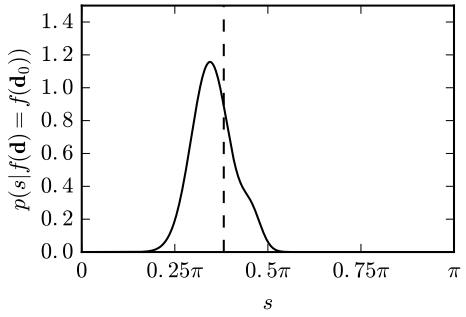

Figure 11: Posterior probability for one sample $\mathbf{d_0}$ taken out during cross-validation computed using features $f(\mathbf{d})$ obtained with PCA directly applied to traveltime data. The vertical dashed line denotes the true value of the sample.

A summary of the obtained results is shown in Table 4 which indicates the best performing case is MDS_t, but it is not far from PCA_t. The similar results of these two cases mean there is no clear advantage in using a non-linear dimension reduction method and may be explained by the mostly linear relation between the structural parameter and the data (as shown by Fig. 9b). We also see that data-driven approaches applied alone perform better than their compositions with insight-driven features, which are used in the last four cases. This means that our chosen insight-driven features provide no better strategy to retain information on the structural parameters s than the data-driven approaches by themselves. This may be explained to some extent by the fact that both PCA and MDS were designed to explicitly search for continuous parameters (also termed continuous latent variables) that explain variability in the data (Bishop, 2006), and not discrete parameters as the ones in the last section. Also, in this case working with the geophysical images gives the worst results and this was not improved by the chosen insight-driven feature (Radon transform).

We must note that our data is highly sensitive to the preferential orientation therefore the dimensions explaining most data variability are indeed related to the chosen structural parameter. When this is not the case, insight-driven features may prove more useful. Finally, it is worth mentioning that insight-driven features are easier to propose when the parameter of interest is discrete, since the expected effect on data can be investigated in a finite number of scenarios.

	N_f	N_c	ℓ_c
PCA_t	3	5	-0.270
MDS_t	4	6	-0.257
PCA_a	3	7	-0.364
MDS_a	5	5	-0.328
MDS_v	2	5	-0.614
MDS_R	4	5	-0.675

Table 4: Cross-validation results for continuous structural parameter s.

4 Conclusions

In this work we provide a novel framework to design and assess data features in the approach proposed by Park et al. (2013)—an approach to reduce the structural parameter uncertainty—making it more objective and readily applicable. Our results show that the design and relative success of data features on which the approach is based is case-dependent, which may therefore challenge the robustness of the approach. Since cross-validation can always be applied, our proposed framework relies on its use to make an objective assessment of the features and the additional degrees of freedom brought by the method.

To illustrate the different choices of feature extraction methods, these were analyzed according to whether they are data-driven only or based on insight about the relation between the data and the structural parameter. In the presented synthetic cases, cross-validation identified the defined insight-driven features as more successful to retrieve the posterior (marginal) probability distribution of a discrete structural parameter (the geological scenario) than for a continuous one (the preferential orientation). Similarly, data-driven approaches performed better for the orientation according to the cross-validation scores and we argue that this is mainly because a significant part of data variability is explained by this structural parameter. We also found that, for the synthetic cases considered in this study, there is not much difference in using a data-driven linear dimension reduction method (such as principal component analysis), in comparison to a nonlinear one (such as multidimensional scaling), other than the former will generally require more dimensions to achieve a similar performance. As an additional result, some useful ways to extract features were proposed when reducing the uncertainty of the geological scenario and the preferential orientation using geophysical tomographic data. All these outcomes may prove useful in the general context of multi-level uncertainty in the Earth sciences. An interesting result of our investigations is that, although geophysical data are often used to produce images of the subsurface through inversion, using the inversion as an insight-driven feature is not necessarily a good approach to reduce the uncertainty on structural parameters. The data themselves can be more informative.

When using data-driven feature extraction techniques, we considered mainly the dimensions that explain most of the variability in the data. It may be interesting for future studies to consider also combinations of different dimensions (maybe excluding the ones explaining most variability) to see if they are more informative on structural parameters, hence provide a better estimation for the structural uncertainty. This may prove especially useful when the structural parameter does not have a major impact on data variability. In the same regard, this suggests using supervised dimension reduction techniques could be beneficial.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement number 722028 (ENIGMA ITN). We thank three anonymous reviewers and the associated editor for their pertinent comments that helped greatly to improve the manuscript. The code used for this work is available at
`https://github.com/jlalvis/CV_AKDE`.

References

D. Arnold, V. Demyanov, T. Rojas, and M. Christie. Uncertainty Quantification in Reservoir Prediction: Part 1—Model Realism in History Matching Using Geological Prior Definitions. *Mathematical Geosciences*, 51(2):209–240, Feb. 2019. ISSN 1874-8961, 1874-8953. doi: 10.1007/s11004-018-9774-6.

O. Aydin and J. Caers. Image transforms for determining fit-for-purpose complexity of geostatistical models in flow modeling. *Computational Geosciences*, 17(2):417–429, Apr. 2013. ISSN 1420-0597, 1573-1499. doi: 10.1007/s10596-013-9340-8.

J. R. Birchak, C. Gardner, J. E. Hipp, and J. M. Victor. High dielectric constant microwave probes for sensing soil moisture. *Proceedings of the IEEE*, 62(1): 93–98, Jan. 1974. ISSN 0018-9219. doi: 10.1109/PROC.1974.9388.

C. M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2.

A. Comunian, L. De Micheli, C. Lazzati, F. Felletti, F. Giacobbo, M. Giudici, and R. Bersezio. Hierarchical simulation of aquifer heterogeneity: implications of different simulation settings on solute-transport modeling. *Hydrogeology Journal*, 24(2):319–334, Mar. 2016. ISSN 1431-2174, 1435-0157. doi: 10.1007/s10040-015-1343-1.

J. De Leeuw and W. J. Heiser. Multidimensional scaling with restrictions on the configuration. In *Multivariate Analysis*, volume V, pages 501–522, Amsterdam, the Netherlands, 1980. North Holland Publishing Company.

V. Demyanov, D. Arnold, T. Rojas, and M. Christie. Uncertainty Quantification in Reservoir Prediction: Part 2—Handling Uncertainty in the Geological Scenario. *Mathematical Geosciences*, 51(2):241–264, Feb. 2019. ISSN 1874-8961, 1874-8953. doi: 10.1007/s11004-018-9755-9.

T. S. Durrani and D. Bisset. The Radon transform and its properties. *Geophysics*, 49(8):1180–1187, Aug. 1984.

L. Feyen and J. Caers. Quantifying geological uncertainty for flow and transport modeling in multi-modal heterogeneous formations. *Advances in Water Resources*, 29(6):912–929, June 2006. ISSN 03091708. doi: 10.1016/j.advwatres.2005.08.002.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science Series. Chapman and Hall/CRC, third edition edition, 2014.

A. Golmohammadi and B. Jafarpour. Simultaneous geologic scenario identification and flow model calibration with group-sparsity formulations. *Advances in Water Resources*, 92:208–227, June 2016. ISSN 03091708. doi: 10.1016/j.advwatres.2016.04.007.

J. D. F. Habbema, J. Hermans, and K. Van den Broek. A stepwise discrimation analysis program using density estimation. In *Proceedings in Computational Statistics*, Vienna, 1974. Physica Verlag.

T. Hermans, F. Nguyen, and J. Caers. Uncertainty in training image-based inversion of hydraulic head data constrained to ERT data: Workflow and case study. *Water Resources Research*, 51(7):5332–5352, July 2015. ISSN 00431397. doi: 10.1002/2014WR016460.

T. Hermans, E. Oware, and J. Caers. Direct prediction of spatially and temporally varying physical properties from time-lapse electrical resistance data. *Water Resources Research*, 52(9):7262–7283, Sept. 2016. ISSN 00431397. doi: 10.1002/2016WR019126.

T. Hermans, F. Nguyen, M. Klepikova, A. Dassargues, and J. Caers. Uncertainty Quantification of Medium-Term Heat Storage From Short-Term Geophysical Experiments Using Bayesian Evidential Learning. *Water Resources Research*, 54(4):2931–2948, Apr. 2018. ISSN 00431397. doi: 10.1002/2017WR022135.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian Model Averaging: A Tutorial. *Statistical Science*, page 22, 1999.

M. M. Khaninezhad and B. Jafarpour. Prior model identification during subsurface flow data integration with adaptive sparse representation techniques. *Computational Geosciences*, 18(1):3–16, Feb. 2014. ISSN 1420-0597, 1573-1499. doi: 10.1007/s10596-013-9378-7.

M. Khodabakhshi and B. Jafarpour. A Bayesian mixture-modeling approach for flow-conditioned multiple-point statistical facies simulation from uncertain training images. *Water Resources Research*, 49(1):328–342, Jan. 2013. ISSN 00431397. doi: 10.1029/2011WR010787.

X. Li and F. T.-C. Tsai. Bayesian model averaging for groundwater head prediction and uncertainty analysis using multimodel and multimethod. *Water Resources Research*, 45(9), Sept. 2009. ISSN 00431397. doi: 10.1029/2008WR007488.

N. Linde, S. Finsterle, and S. Hubbard. Inversion of tracer test data using tomographic constraints. *Water Resources Research*, 42(4), Apr. 2006. ISSN 00431397. doi: 10.1029/2004WR003806.

N. Linde, P. Renard, T. Mukerji, and J. Caers. Geological realism in hydrogeological and geophysical inverse modeling: A review. *Advances in Water Resources*, 86:86–101, Dec. 2015. ISSN 03091708. doi: 10.1016/j.advwatres.2015.09.019.

A. Malinverno. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International*, 151(3):675–688, Dec. 2002. ISSN 0956-540X, 1365-246X. doi: 10.1046/j.1365-246X.2002.01847.x.

M. Morzfeld, J. Adams, S. Lunderman, and R. Orozco. Feature-based data assimilation in geophysics. *Nonlinear Processes in Geophysics*, 25(2):355–374, May 2018. ISSN 1607-7946. doi: 10.5194/npg-25-355-2018.

S. P. Neuman. Maximum likelihood Bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 17(5):291–305, Nov. 2003. ISSN 1436-3240, 1436-3259. doi: 10.1007/s00477-003-0151-7.

H. Park, C. Scheidt, D. Fenwick, A. Boucher, and J. Caers. History matching and uncertainty quantification of facies models with multiple geological interpretations. *Computational Geosciences*, 17(4):609–621, Aug. 2013. ISSN 1420-0597, 1573-1499. doi: 10.1007/s10596-013-9343-5.

A. E. Raftery. Bayesian Model Selection in Social Research. *Sociological Methodology*, 25:111–163, 1995. ISSN 00811750, 14679531. doi: 10.2307/271063.

P. Renard and D. Allard. Connectivity metrics for subsurface flow and transport. *Advances in Water Resources*, 51:168–196, Jan. 2013. ISSN 03091708. doi: 10.1016/j.advwatres.2011.12.001.

C. Rücker, T. Günther, and F. M. Wagner. pyGIMLi: An open-source library for modelling and inversion in geophysics. *Computers & Geosciences*, 109:106–123, Dec. 2017. ISSN 00983004. doi: 10.1016/j.cageo.2017.07.011.

C. Scheidt and J. Caers. Representing Spatial Uncertainty Using Distances and Kernels. *Mathematical Geosciences*, 41(4):397–419, May 2009. ISSN 1874-8961, 1874-8953. doi: 10.1007/s11004-008-9186-0.

C. Scheidt, C. Jeong, T. Mukerji, and J. Caers. Probabilistic falsification of prior geologic uncertainty with seismic amplitude data: Application to a turbidite reservoir case. *Geophysics*, 80(5):M89–M12, Sept. 2015a. ISSN 0016-8033, 1942-2156. doi: 10.1190/geo2015-0084.1.

C. Scheidt, P. Tahmasebi, M. Pontiggia, A. Da Pra, and J. Caers. Updating joint uncertainty in trend and depositional scenario for reservoir exploration and early appraisal. *Computational Geosciences*, 19(4):805–820, Aug. 2015b. ISSN 1420-0597, 1573-1499. doi: 10.1007/s10596-015-9491-x.

C. Scheidt, L. Li, and J. Caers. *Quantifying Uncertainty in Subsurface Systems*. Number 236 in Geophysical Monograph Series. John Wiley and Sons & American Geophysical Union, Hoboken, NJ & Washington D.C., 2018.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1986.

A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, Jan. 2005. ISBN 978-0-89871-572-9 978-0-89871-792-1. doi: 10.1137/1.9780898717921.

A. Tarantola and B. Valette. Inverse problems = quest for information. *Journal of Geophysics*, 50(3):159–170, 1982.

F. T.-C. Tsai and A. S. Elshall. Hierarchical Bayesian model averaging for hydrostratigraphic modeling: Uncertainty segregation and comparative evaluation. *Water Resources Research*, 49(9):5520–5536, Sept. 2013. ISSN 00431397. doi: 10.1002/wrcr.20428.

J. A. Vrugt, C. J. F. ter Braak, H. V. Gupta, and B. A. Robinson. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment*, 23(7):1011–1026, Oct. 2009. ISSN 1436-3240, 1436-3259. doi: 10.1007/s00477-008-0274-y.

T. Xu and A. J. Valocchi. A Bayesian approach to improved calibration and prediction of groundwater models with structural error. *Water Resources Research*, 51(11):9290–9311, Nov. 2015. ISSN 00431397. doi: 10.1002/2015WR017912.

M. Ye, S. P. Neuman, and P. D. Meyer. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resources Research*, 40(5), May 2004. ISSN 00431397. doi: 10.1029/2003WR002557.

Appendix: Adaptive kernel density estimation

The standard (non-adaptive) equation for kernel density estimation that would apply for our case is (Scheidt et al., 2018)

$$p(s|f(\mathbf{d}_{obs})) = \frac{p(s, f(\mathbf{d}_{obs}))}{p(\mathbf{d}_{obs})} = \frac{\sum_{j=1}^{N} K_{H_s}(s - s_j) K_{H_f}(f(\mathbf{d}_{obs}) - f(\mathbf{d}_j))}{\sum_{j=1}^{N} K_{H_f}(f(\mathbf{d}_{obs}) - f(\mathbf{d}_j))} \tag{9}$$

where the involved variables are the same as in Eq. (2) but here no clustering is defined, therefore no separate summation for each cluster is needed and the bandwidths H_s and H_f for the scaled kernel functions are the same for all the N Monte Carlo samples. The expected value of Eq. (9) is also referred to as the Nadaraya-Watson model or kernel regression (Bishop, 2006).

In general, the bandwidth H refers to the width of the kernel that is used to approximate the distributions and for the multivariate case it is a $Q \times Q$ matrix, where Q is the number of dimensions of the variable. Different kernel functions may be used to do this approximation (Silverman, 1986), in our case we chose the multivariate independent Gaussian kernel.

$$K_H(\mathbf{x}) = (2\pi)^{-Q/2} |H|^{-1/2} e^{-\frac{1}{2}\mathbf{x}^T H^{-1} \mathbf{x}} \tag{10}$$

where Q is the number of dimensions of $\mathbf{x}$ and H is a diagonal matrix. As suggested by Park et al. (2013) and Scheidt et al. (2015b), we used clustering in order to make the KDE bandwidth H adaptive. This requires the specification of the number N_c of clusters and results in narrow bandwidths where the density of points is high and wide bandwidths where density is low. We used k-means clustering on the feature space and each sample is assigned a bandwidth H for both the features and the structural parameter according to which cluster it belongs to. The value of the bandwidth H (a diagonal matrix) within each cluster is computed by means of Silverman's rule of thumb (Silverman, 1986) as

$$(H_{ii})^{1/2} = \frac{4}{Q+2}^{\frac{1}{Q+4}} n^{\frac{-1}{Q+4}} \sigma_i \tag{11}$$

where n denotes the number of samples and may be different for each cluster, and σ_i is the standard deviation in the i-th dimension in the same cluster. In this way, the control on the bandwidth is implicit on the number of clusters N_c. Applying KDE with this adaptive approach is expressed in Eq. (2). There H_s and H_f are computed using the same clusters and have dimensions 1×1 and $N_f \times N_f$, respectively.