

# Redshift determination through weighted phase correlation: a linearithmic implementation

L. Delchambre<sup>1\*</sup>

<sup>1</sup> *Institut d'Astrophysique et de Géophysique, Université de Liège, Allée du 6 Août 17, B-4000 Sart Tilman (Liège), Belgique*

Accepted ????. Received ???; in original form ???

## ABSTRACT

We present a new algorithm having a time complexity of  $\mathcal{O}(N \log N)$  and designed to retrieve the phase at which an input signal and a set of not necessarily orthogonal templates match at best in a weighted chi-squared sense. The proposed implementation is based on an orthogonalization algorithm and thus also benefits from a high numerical stability. We successfully apply this method to the redshift determination of quasars from the twelfth Sloan Digital Sky Survey (SDSS) quasar catalog and derive the proper spectral reduction and redshift selection methods. Also provided are the derivations of the redshift uncertainty and of the associated confidence. Results of this application are comparable to the performances of the SDSS pipeline while not having a quadratic time dependency.

**Key words:** methods: data analysis – quasars: distances and redshifts.

## 1 INTRODUCTION

The advent of extremely large spectroscopic surveys like the Sloan Digital Sky Survey (SDSS) that includes more than  $2 \times 10^6$  high resolution spectra over  $5200 \text{ deg}^2$  of the sky (Alam et al. 2015) or the Gaia space mission that will provide, by the end of 2018,  $150 \times 10^6$  low resolution spectra (de Bruijne 2012) provide us with unique opportunities to have a statistical view on the kind of objects present in the universe along with some of their fundamental characteristics. These play a key role in the answer to some of the currently most important astrophysical questions like the evolution scenarios of the galaxy; of the universe or its accelerated expansion (Aubourg et al. 2014; Perryman et al. 2001).

Along with these large surveys comes an impressive continuous flow of data that has to be treated right on time through huge dedicated processing centers. One of the most important tasks amongst the spectral reduction processes stands in the objects classification and in their astrophysical parameters (APs) determination. More specifically, in the case of extragalactic objects, these informations critically depend on the availability of reliable redshift estimates.

Redshift determination, even if apparently straightforward, is in practice a challenging problem for which numerous solutions have been proposed:

(i) Visual inspection procedures: a skilled observer can efficiently guess the APs of any object and can deal with

any unexpected cases like corrupted/missing emission lines; spectra superposition or non-physical solutions. Obviously, this choice is unbearable for large surveys though the analysis of any sufficiently large subset is invaluable as it can serve as input to sophisticated computer algorithms that will try to mimic this human expertise. This is the solution undertaken by Pâris et al. (2016) regarding the redshift of quasars and accordingly, it will be used along this paper as the default quasar spectral library.

(ii) Matching of spectral lines: this method consists in extracting some significant patterns out of the input spectra and then trying to match them to known emission/absorption lines. This procedure has been used for a long time but has been shown to be restricted to relatively high signal-to-noise ratio spectra (Machado 2013).

(iii) Computer learning methods: the goal is here to make the algorithm guess the relations that exist between some characteristics of already-reduced objects (e.g. observed wavelengths and fluxes), and the parameters of interest (e.g. redshift coming from a visual inspection procedure). The aim being then to apply these relations to the case of objects whose parameters are still unknown. Interested readers may find in Bishop (2006) the descriptions of many such algorithms. Note that, depending on its complexity, the guessed relation may be non-physical and hard to interpret leading to suboptimal or potentially unrealistic predictions. This is the reason why these should preferably be used for the case of highly non-linear problems for which no other-fast-solution exists.

(iv) Phase correlation: the idea is here to find the optimal correlation of a given observation against one or more

\* E-mail: ldelchambre@ulg.ac.be

templates in order to determine its redshift. Based upon the ability of these templates to match the observations; from the physical nature of this solution and from the shortcomings of the previously mentioned alternatives, we will consider it to be the most trustworthy automated procedure for redshift determination.

Based on the work of [Brault & White \(1971\)](#), [Simkin \(1974\)](#) first suggested the use of the Fast Fourier Transform (FFT) as an efficient way of finding the redshift of galaxies based on their cross-correlation with a single template. [Tonry & Davis \(1979\)](#) later derived the formulation associated with the resulting redshift uncertainties, that were further refined by [Heavens \(1993\)](#). Finally, [Glazebrook \(1997\)](#) generalized the cross-correlation technique to the case of templates coming from the principal components analysis (PCA) decomposition of spectral libraries. Although being currently the most widespread technique for redshift determination, the latter actually suffers from some well-known drawbacks (see section 3.2.1). The solution to these problems comes from the use of a weighting scheme associated with the observed spectrum as implemented in [Bolton et al. \(2012\)](#). Unfortunately, this solution has a quadratic time dependency that makes it fairly time consuming.

The method proposed in the present work overcomes this high numerical complexity and was developed in the framework of the Gaia astrophysical parameters inference system ([Bailer-Jones et al. 2013](#)) and more specifically within the field of the quasar classification module (QSOC) whose goal is to find the APs associated with the quasars that Gaia will detect. In this domain, the time constraints imposed by the Gaia mission restricted us to the use of computer learning methods but in the end, the advent of this new method will allow us to predict fair and fast redshift estimates for the upcoming Gaia data releases.

Section 2 explains the conventions used along this paper. Section 3 makes a brief review of the phase correlation and PCA techniques aimed at better understanding their main limitations. We have developed a fast solution to the problem of the weighted phase correlation in Section 4. Tests against real cases are then performed within Section 5 while extensions of the presented algorithm are discussed in Section 6. Finally, we conclude in Section 7.

## 2 NOTATION

This paper uses the following notations: vectors are in bold italic,  $\mathbf{x}$ ;  $x_i$  being the element  $i$  of the vector  $\mathbf{x}$ . Matrices are in uppercase boldface or are explicitly stated; i.e.  $\mathbf{X}$  from which the  $i$ th column will be denoted by  $\mathbf{X}_i^{\text{col}}$  and the element at row  $i$ , column  $j$  will be denoted by  $\mathbf{X}_{ij}$ . In the following, we will consider the problem of finding the optimal offset between an observed spectrum composed of  $N_s$  samples and  $N_T$  templates of size  $N_p$  by probing various shift estimate,  $Z$ . By considering the zero-padding necessary in order for these to be properly used within the Fourier domain, we will have that the template matrices,  $\mathbf{P}$  and  $\mathbf{T}$ , will be of size  $(N \times N_T)$  with  $N = N_s + N_p$ . Similarly, we will have that the observation vector,  $\mathbf{s}$ , will be of size  $N$  as well. Note that in order for the redshift to turn into a simple offset, we will have to consider a logarithmic wavelength scale. If not stated otherwise, matrices and vectors having a tilde

on top of them (e.g.  $\tilde{\mathbf{T}}$ ) will be specific to a given shift try,  $Z \in 0 \dots N - 1$ . Amongst commonly used operators,  $\mathbf{a} \cdot \mathbf{b}$  denotes the inner-product of  $\mathbf{a}$  and  $\mathbf{b}$ ;  $\mathbf{a} \otimes \mathbf{b}$ , their outer-product;  $\|\mathbf{a}\|$ , the Euclidian norm of  $\mathbf{a}$  and  $\mathbf{a}^*$  its complex conjugate. Finally,  $\mathcal{F}[\mathbf{x}]$  and  $\mathcal{F}^{-1}[\mathbf{x}]$  respectively corresponds to the discrete Fourier transform (hereafter DFT) and inverse DFT of  $\mathbf{x}$ .

## 3 PHASE CORRELATION USING PCA

As already stated, the most commonly used technique for QSO redshift determination consists in finding the best correlation of an observed spectrum against templates coming from the PCA decomposition of a restframe spectral library. More specifically, these are based on spectra sampled on a uniform logarithmic wavelength scale such that the observed wavelength,  $\lambda_{\text{obs}}$ , can be related to the restframe wavelength,  $\lambda_{\text{rest}}$ , through the QSO redshift,  $z$ , as a simple offset:

$$\log \lambda_{\text{obs}} = \log \lambda_{\text{rest}} + \log(z + 1). \quad (1)$$

In the following, we make a brief review of the two above-mentioned techniques that should provide the reader insights about their way of working and aimed at better understanding their main limitations regarding the redshift estimation of QSOs.

### 3.1 Principal components analysis

PCA is a well-known technique designed to extract a set of templates –the principal components– from a typically huge set of data while keeping most of its variance ([Pearson 1901](#)). These principal components will then be those that are the best suited in order to highlight the most important patterns out of the input data set.

Mathematically, the goal of the PCA is to find a decomposition of an input matrix  $\mathbf{X}$ , from which we have subtracted the mean observation, into

$$\mathbf{X} = \mathbf{P}\mathbf{C}, \quad (2)$$

such that

$$\mathbf{D} = \mathbf{P}^T \mathbf{X} \mathbf{X}^T \mathbf{P} = \mathbf{P}^T \boldsymbol{\sigma}^2 \mathbf{P} \quad (3)$$

is diagonal and for which

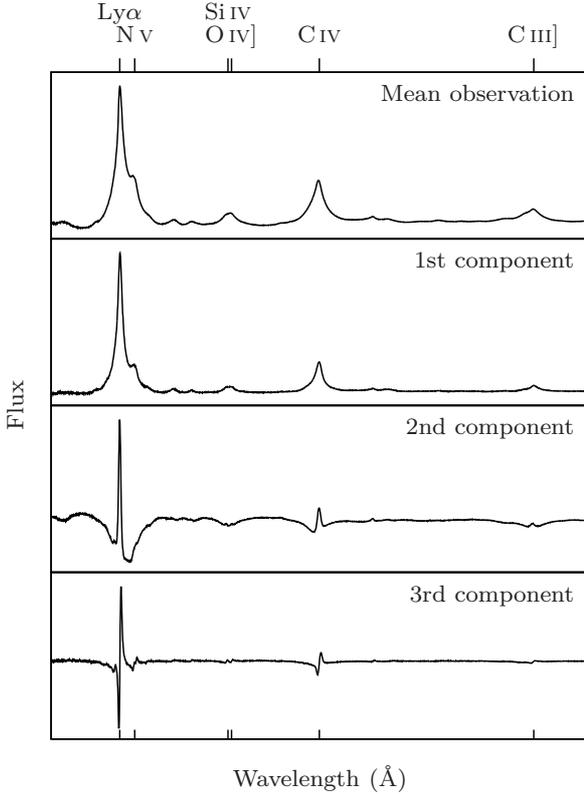
$$\mathbf{D}_i \geq \mathbf{D}_j; \forall i < j. \quad (4)$$

$\mathbf{P}$ , the matrix of the eigenvectors of  $\boldsymbol{\sigma}^2$ , is then called the matrix of principal components;  $\mathbf{C}$  is the associated matrix of principal coefficients and  $\mathbf{D}_i$ 's are the eigenvalues of the covariance matrix,  $\boldsymbol{\sigma}^2$ . Note that according to the spectral theorem<sup>1</sup>,  $\mathbf{P}$  will be orthonormal such that we have

$$\mathbf{C} = \mathbf{P}^T \mathbf{X}. \quad (5)$$

From this orthonormality and from equation 4, we will have that the linear combination of the first principal components of  $\mathbf{P}$  with the associated principal coefficients of  $\mathbf{C}$  will constitute the best linear combination in order to fit  $\mathbf{X}$  in a least squares sense.

<sup>1</sup> Any real symmetric matrix is diagonalized by a matrix of its eigenvectors.



**Figure 1.** Restframe mean observation and first principal components coming from the PCA decomposition of Type I/II SDSS DR12 quasars spectra having  $2.3 \leq z \leq 2.4$  (24939 spectra).

An illustrative example of PCA decomposition is given in figure 1. The latter is based on spectra covering the rest-frame wavelength range 1100–2000Å coming from the SDSS DR12 quasar catalog (Pâris et al. 2016). Notice how the main QSOs emission lines are modelled by the various components as a way to grab the variance coming from the great diversity of shapes encountered within the spectral library. Readers willing more information on the PCA decomposition are invited to read Jolliffe (2002) for a deep analysis of the technique or Schless (2014) for an accessible tutorial.

The application of this technique to the analysis of QSO spectra was first covered by Francis et al. (1992); Yip et al. (2004) later adapted it to the case of the SDSS DR1 quasars classification and redshift determination while Cabanac et al. (2002) did a similar work based upon spectra coming from the Large Zenith Telescope survey whose spectral resolution ( $\lambda/\Delta\lambda \sim 40$ ) is in the same order of magnitude as the one of the red and blue photometers of Gaia (Bailer-Jones et al. 2013).

### 3.1.1 Weighted PCA

One of the main limitations of the classical PCA method is that it does not make any distinction between variance coming from noise and variance coming from a genuine signal. Furthermore, in its naive form, it does not know how to deal with missing data. This last point is particularly cru-

cial in the field of high-redshift surveys where the observed wavelength ranges may not overlap from object to object.

A straightforward approach so as to avoid these shortcomings stands in the use of a weighting scheme that allows each flux within each spectrum to come along with its own uncertainty while performing the PCA decomposition. Such a fully-weighted PCA (WPCA) method was first described in the astronomical literature by Tsalmantza & Hogg (2012) and was later refined by Bailey (2012). In the field of the present study, we will use the implementation described in Delchambre (2015), this choice mainly comes from its high numerical stability. This method is based on the diagonalization of the weighted variance-covariance matrix as defined by

$$\sigma^2 = \frac{(\mathbf{X} \circ \mathbf{W})(\mathbf{X} \circ \mathbf{W})^T}{\mathbf{W}\mathbf{W}^T}, \quad (6)$$

where  $\circ$  represents the element-wise product of two matrices and where  $\mathbf{X}$  is supposed to have a weighted mean observation of zero. The decomposition of  $\sigma^2$  into a diagonal matrix of eigenvalues,  $\mathbf{D}$ , and a matrix of orthonormal principal components,  $\mathbf{P}$ , being then performed using either a combination of two spectral decomposition methods, namely the power iteration method followed by the Rayleigh quotient iteration one, or by the use of the singular value decomposition (SVD). This technique allows us to retrieve the fairest components (i.e. those for which uncertainties are taken into account) without having to worry about missing data: this case being the limiting case of weights equal to zero. Consequently, this method will be used through the rest of this document as the default process in order to retrieve the principal components.

### 3.2 Phase correlation

The goal of the phase correlation algorithm is to find the optimal shift between a set of orthonormal templates –or a sole unit-length template–,  $\mathbf{P}$ , and a given observation,  $\mathbf{s}$ , that has been shifted relatively to  $\mathbf{P}$ . The way to proceed is to compute for each potential shift,  $Z$ , the linear least-squares solution of the shifted templates,  $\tilde{\mathbf{P}}_{ij} \equiv \mathbf{P}_{(i+Z)j}$ , against the observation such as to find the offset having the minimal resulting chi-square. More concisely, this is equivalent to find the minimal shift-dependent chi-square as defined by

$$\chi^2(Z) = \left\| \mathbf{s} - \tilde{\mathbf{P}}\mathbf{a}(Z) \right\|^2, \quad (7)$$

where  $\mathbf{a}(Z)$  contains the optimal linear coefficients in order to fit  $\mathbf{s}$  based on  $\tilde{\mathbf{P}}$ .

Extending the work of Simkin (1974), Glazebrook (1997) noticed that in the case of orthonormal templates, like the PCA principal components, equation 7 becomes

$$\chi^2(Z) = \|\mathbf{s}\|^2 - \|\mathbf{a}(Z)\|^2. \quad (8)$$

Consequently, equation 7 will be minimal for an associated maximal  $\|\mathbf{a}(Z)\|^2$ . Moreover, due to the orthonormality of  $\mathbf{P}$ , we will have that

$$\mathbf{a}(Z) = \tilde{\mathbf{P}}^T \mathbf{s}. \quad (9)$$

More specifically, regarding the  $i$ th linear coefficient,  $a_i(Z)$ , we will have that

$$a_i(Z) = \sum_j \mathbf{P}_{(j+Z)i} s_j = \left( \mathbf{P}_i^{\text{col}} \star \mathbf{s} \right)_Z. \quad (10)$$

We recognize equation 10 as being the correlation of the vector  $\mathbf{P}_i^{\text{col}}$  with  $\mathbf{s}$  that can hence be efficiently computed in the Fourier domain. Interested readers may find in Brault & White (1971) exhaustive hints about the practicalities surrounding the Fourier implementation of equation 10. Let us just point out that both vectors,  $\mathbf{P}_i^{\text{col}}$  and  $\mathbf{s}$  have to be extended and zero-padded such as to deal with the periodic nature of the DFT. Note that in the rest of this document, the curve obtained after evaluating  $\|\mathbf{a}(Z)\|^2$  at each  $Z$  will be termed the *cross-correlation function* (CCF).

A sub-sampling precision on the offset can be gained by considering the fit of a continuous function in the vicinity of the maximal peak of the discrete CCF. Simkin (1974) supposed this peak to be Gaussian profiled, but in the aim of having a model-independent estimate of  $Z$ , we will follow Tonry & Davis (1979) and use a quadratic curve fitting that will allow us to take into account potential asymmetries in the fitted peak.

### 3.2.1 Practicalities

Some issues highlighted in Glazebrook (1997) are the subtraction of the QSO continuum and of the restframe mean spectrum from the observed spectrum. The first issue was here solved by the use of a dedicated method that allows us to fit the QSO continuum in a fast and redshift-independent way. This method will be further described in section 5.1. The second issue is often overcome by omitting the subtraction of the mean spectrum from the input dataset. We have to note that this omission typically degrades the ability of the PCA decomposition to extract the most significant patterns out of this input dataset. Another solution would have been to alter the mean spectrum such as to make it orthonormal to the template components,  $\mathbf{P}$ , –thanks to the use of a Gram–Schmidt orthogonalization process (Press et al. 2002) for example– and to further consider it as being an additional template. This solution will be adopted here for the use of the phase correlation algorithm.

Finally, the major drawback of the implementation of Glazebrook (1997) stands in the fact that the observed spectra typically span only a small part of the template spectra such that the CCF will be computed over a substantial number of unknown points. As a consequence, the fit of the input spectra will be disrupted by the ‘flattening’ of the principal components over the unobserved wavelengths. Figure 2 illustrates the result of the phase correlation algorithm along with the best-fit solution associated with the maximal peak of the CCF. Notice how the solutions are flattened over unobserved wavelengths. More precisely, considering the observation of “SDSS J024008.93-003448.7”, the Ly $\alpha$ , H $\alpha$  and H $\beta$  emission lines are strongly damped despite the fact that the optimal shift was found while for the observation of “SDSS J132218.88+365342.0”, this ‘flattening’ has led to an ambiguity in the CCF that leads to an erroneous shift estimate. Additionally, uncertainties about the observed fluxes are often available and will not be used within this implementation.

## 4 WEIGHTED PHASE CORRELATION

With the aim of dealing efficiently with the previously mentioned problem of unobserved wavelengths and of neglected uncertainties, we will use a  $\chi^2$  formulation similar to equation 7, but whose fluxes are weighted according to the observed spectrum wavelengths. Also, we will drop the orthonormality constraint on the fitted templates since, in any way, the previously mentioned weighting will break it down. We will then have the following objective formula:

$$\chi^2(Z) = \left\| \mathbf{W}\mathbf{s} - \mathbf{W}\tilde{\mathbf{T}}\mathbf{a}(Z) \right\|^2 = \left\| \mathbf{y} - \tilde{\mathbf{X}}\mathbf{a}(Z) \right\|^2, \quad (11)$$

where  $\mathbf{W}$  is the diagonal matrix of weights associated with  $\mathbf{s}$  and  $\tilde{\mathbf{T}}_{ij} \equiv \mathbf{T}_{(i+Z)j}$  is the shifted matrix of –not necessarily orthonormal– template observations. The fastest solution in order to minimize equation 11 for a given  $Z$  stands in the use of a Cholesky decomposition of the design matrix,  $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$ , followed by a forward-backward substitution associated with the image vector  $\tilde{\mathbf{X}}^T\mathbf{y}$  (Press et al. 2002). We have to note that this approach is known to suffer from numerical instabilities (Golub & Van Loan 1996; Press et al. 2002) and is solely provided here as a comparison point regarding its computational performances. Practically, slower but more stable methods based on the orthogonalization of  $\tilde{\mathbf{X}}$  should be preferred.

In a computational point of view, the evaluation of equation 11 for each  $Z$  will require  $\mathcal{O}(N^2)$  flops<sup>23</sup>, the latter being mainly dedicated to the building of the design matrices. This relatively high complexity constitutes the main limitation of this implementation and makes it unaffordable for the tight processing of a large survey like Gaia. Nonetheless, it has proven to provide fair redshift estimates and is currently being effectively used in the SDSS-III spectral classification redshift measurement pipeline with a singular value decomposition (SVD) of  $\tilde{\mathbf{X}}$  advantageously replacing the Cholesky decomposition of the design matrix (Bolton et al. 2012).

### 4.1 Orthogonal decomposition approach

The previous section points out the risks encountered while using a naive approach for solving the normal equations associated with equation 11. In this optics, let us explore the effect of the orthogonalization of  $\tilde{\mathbf{X}}$  on the latter equation. For this purpose, let us detail the QR decomposition of  $\tilde{\mathbf{X}} = \mathbf{Q}\mathbf{R}$ <sup>4</sup>, that is such that

$$\mathbf{Q}^T\tilde{\mathbf{X}} = \mathbf{Q}_{N_T-1} \cdots \mathbf{Q}_1\tilde{\mathbf{X}} = \mathbf{Q}_{N_T-1} \cdots \mathbf{Q}_1\tilde{\mathbf{X}}_i = \mathbf{R}, \quad (12)$$

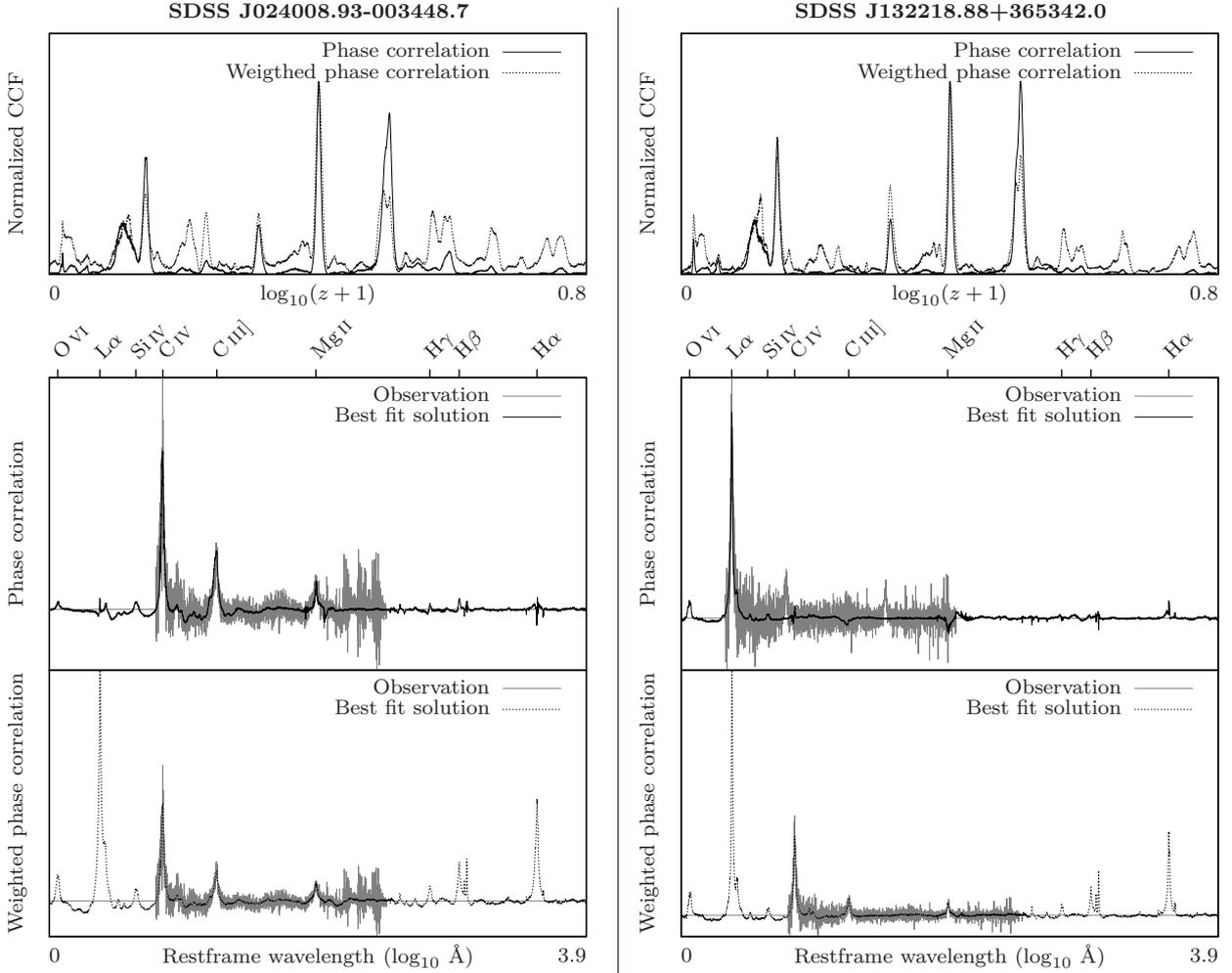
where  $\mathbf{R}$  is an upper triangular matrix of size  $(N \times N_T)$  and where each  $\mathbf{Q}_i$  is an Householder reflection matrix designed to annihilate the elements below the  $i$ th row of the  $i$ th column of  $\tilde{\mathbf{X}}_i$  (Press et al. 2002). More precisely, given  $\tilde{\mathbf{X}}'_i$ , the not-already upper-triangular part of  $\tilde{\mathbf{X}}_i$ , we will have

$$\mathbf{Q}_i = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - 2\mathbf{v}_i \otimes \mathbf{v}_i / \|\mathbf{v}_i\|^2 \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}'_i \end{pmatrix} \quad (13)$$

<sup>2</sup> Floating operations

<sup>3</sup> The interested reader may find in Golub & Van Loan (1996) informations and references about the various algorithmic complexities used along this document.

<sup>4</sup> Note that we dropped the upper tilde for clarity purpose



**Figure 2.** Examples of cross-correlation functions coming from the phase correlation and weighted phase correlation of two quasar spectra against the ten first SDSS DR12 principal components plus mean observation (top). The best fit solutions associated with the maximal peak of the phase correlation (middle) and of the weighted phase correlation (bottom) are also provided.

with

$$\mathbf{v}_i = \mathbf{x}_i \pm \|\mathbf{x}_i\| \mathbf{e}_1; \quad (14)$$

$\mathbf{x}_i$  being the first column of  $\tilde{\mathbf{X}}'_i$  and  $\mathbf{e}_1$  being the first row of the identity matrix. For numerical stability reasons, the choice between subtraction and addition in equation 14 should be matched to the sign of the first element of  $\mathbf{x}_i$  (Press et al. 2002).

By using such a decomposition, we will have that equation 11 becomes

$$\chi^2(Z) = \|\mathbf{y} - \mathbf{Q}\mathbf{R}\mathbf{a}(Z)\|^2 = \|\mathbf{y} - \mathbf{Q}\mathbf{b}(Z)\|^2, \quad (15)$$

with the last  $N - N_T$  elements of  $\mathbf{b}(Z)$  being zeros. The point is now to recognize equation 15 as being the weighted counterpart of equation 7 such that the first  $N_T$  elements of  $\mathbf{b}(Z)$  will be equal to the first  $N_T$  elements of  $\mathbf{Q}^T \mathbf{y}$ , whose computation can be efficiently performed by successive multiplication of each of the  $\mathbf{Q}_i$  with the associated  $\mathbf{y}_i \equiv \mathbf{Q}_{i-1} \cdots \mathbf{Q}_1 \mathbf{y} = (b_1(Z) \cdots b_{i-1}(Z) \mathbf{y}'_i)$ , rather than by

explicitly computing the general  $\mathbf{Q}$  matrix. This efficiency mainly comes from the fact that:

(i) We do not need to explicitly compute any  $\mathbf{Q}'_i$ , since we will have that the  $j$ th column of the product  $\mathbf{Q}'_i \tilde{\mathbf{X}}'_i$  will be given by

$$\left(\mathbf{Q}'_i \tilde{\mathbf{X}}'_i\right)_j^{\text{col}} = \left(\tilde{\mathbf{X}}'_i\right)_j^{\text{col}} - 2 \frac{\mathbf{v}_i \cdot \left(\tilde{\mathbf{X}}'_i\right)_j^{\text{col}}}{\mathbf{v}_i \cdot \mathbf{v}_i} \mathbf{v}_i, \quad (16)$$

and similarly,

$$\mathbf{Q}'_i \mathbf{y}'_i = \mathbf{y}'_i - 2 \frac{\mathbf{v}_i \cdot \mathbf{y}'_i}{\mathbf{v}_i \cdot \mathbf{v}_i} \mathbf{v}_i. \quad (17)$$

That is: the computation of  $\mathbf{Q}'_i \mathbf{y}'_i$  and of any column of the products  $\mathbf{Q}'_i \tilde{\mathbf{X}}'_i$  is now reduced to a single inner product (the product  $\mathbf{v}_i \cdot \mathbf{v}_i$  being common to all multiplications, it can be pre-computed) and to a single vector subtraction.

(ii) We do not need to compute any  $\mathbf{R}_{i,j}$ . Differently stated, we do not need to compute the first row nor the first column of any  $\mathbf{Q}'_i \tilde{\mathbf{X}}'_i$ .

This implementation, termed ‘factorized QR algorithm’, has a total complexity which can compete with the Cholesky solution of the normal equations while gaining in numerical stability. But practically it is of low interest for us since it remains a quadratic problem that is consequently out of the time processing required by the Gaia tight data reduction.

Let us note that the equation 15 still provides us with a weighted formulation of the CCF, that is  $\|\mathbf{b}(Z)\|^2$ , such that we can already investigate the effects of the weighting on the best fit solutions at its maximal peak and on the CCF itself. As illustrated in figure 2, the fitted spectra do no longer exhibit border flattening and thanks to this, the maximal peaks are now clearly identified. More particularly, regarding the observation of “SDSS J132218.88+365342.0”, the optimal peak of the CCF turns out to be unambiguously identified thanks to the use of this weighted formulation of the phase correlation.

#### 4.1.1 Factorized QR algorithm with lookup tables

The quadratic nature of the factorized QR algorithm comes from the large amount of inner products involved in the computation of the first  $N_T$  elements of each  $\mathbf{b}(Z)$ . More specifically, by developing each inner product coming from equations 16 and 17 in the case of the initial reduction  $\tilde{\mathbf{X}}_2 \equiv \mathbf{Q}_1 \tilde{\mathbf{X}}$  and associated image production  $\mathbf{y}_2 \equiv \mathbf{Q}_1 \mathbf{y}$ , we get

$$\mathbf{v}_1 \cdot \mathbf{v}_1 = 2\alpha \left( \alpha + \tilde{\mathbf{X}}_{11} \right), \quad (18)$$

$$\mathbf{v}_1 \cdot \mathbf{y} = \alpha y_1 + \tilde{\mathbf{X}}_1^{\text{col}} \cdot \mathbf{y}, \quad (19)$$

$$\mathbf{v}_1 \cdot \tilde{\mathbf{X}}_j^{\text{col}} = \alpha \tilde{\mathbf{X}}_{1j} + \tilde{\mathbf{X}}_1^{\text{col}} \cdot \tilde{\mathbf{X}}_j^{\text{col}} \quad (20)$$

with  $\alpha = \text{sgn}(\tilde{\mathbf{X}}_{11}) \left( \tilde{\mathbf{X}}_1^{\text{col}} \cdot \tilde{\mathbf{X}}_1^{\text{col}} \right)^{\frac{1}{2}}$ . At this point, it should be noted that

$$\tilde{\mathbf{X}}_i^{\text{col}} \cdot \tilde{\mathbf{X}}_j^{\text{col}} = \mathbf{w}^2 \cdot \left( \tilde{\mathbf{T}}_i^{\text{col}} \circ \tilde{\mathbf{T}}_j^{\text{col}} \right) = \sum_{k=1}^N w_k^2 \left( \mathbf{T}_i^{\text{col}} \circ \mathbf{T}_j^{\text{col}} \right)_{k+Z} \quad (21)$$

and that

$$\tilde{\mathbf{X}}_i^{\text{col}} \cdot \mathbf{y} = (\mathbf{w}^2 \circ \mathbf{s}) \cdot \tilde{\mathbf{T}}_i^{\text{col}} = \sum_{k=1}^N w_k^2 s_k \mathbf{T}_{(k+Z)i}, \quad (22)$$

with  $\mathbf{w} \equiv \text{diag}(\mathbf{W})$ . We can readily see that equations 21 and 22 can be efficiently computed in the Fourier domain. In order to take benefits from it, let us define the *lookup table* of the inner products of  $\tilde{\mathbf{X}}$  with itself as

$$\tilde{\mathbf{L}}_{ij} = \tilde{\mathbf{X}}_i^{\text{col}} \cdot \tilde{\mathbf{X}}_j^{\text{col}} = \mathcal{F}^{-1} \left[ \mathcal{F} \left[ \mathbf{T}_i^{\text{col}} \circ \mathbf{T}_j^{\text{col}} \right]^* \circ \mathcal{F} \left[ \mathbf{W}^2 \right] \right]_Z, \quad (23)$$

and the one containing the inner products of  $\tilde{\mathbf{X}}$  with  $\mathbf{y}$  as

$$\tilde{\mathbf{l}}_i = \tilde{\mathbf{X}}_i^{\text{col}} \cdot \mathbf{y} = \mathcal{F}^{-1} \left[ \mathcal{F} \left[ \mathbf{T}_i^{\text{col}} \right]^* \circ \mathcal{F} \left[ \mathbf{W}^2 \mathbf{s} \right] \right]_Z. \quad (24)$$

Note that in the latter equations,  $\mathcal{F} \left[ \mathbf{T}_i^{\text{col}} \circ \mathbf{T}_j^{\text{col}} \right]^*$  and  $\mathcal{F} \left[ \mathbf{T}_i^{\text{col}} \right]^*$  are template-specific and can hence be computed in advance.

Explicitly stated, these lookup tables allow us to have for any shift estimates,  $Z$ , an instantaneous evaluation of all the inner products associated with the initial reduction process. Furthermore, thanks to the  $\mathbf{Q}_1$  orthonormality, we will have that the lookup tables associated with  $\tilde{\mathbf{X}}_2$  will be

also given by  $\tilde{\mathbf{L}}$  and  $\tilde{\mathbf{l}}$ . Consequently, we can easily compute the inner product of  $\tilde{\mathbf{X}}_2^{\text{col}}$  with itself based on  $\tilde{\mathbf{L}}$  as

$$\left( \tilde{\mathbf{X}}_2^{\text{col}} \right)_i \cdot \left( \tilde{\mathbf{X}}_2^{\text{col}} \right)_j = \tilde{\mathbf{L}}_{ij} - \mathbf{R}_{1i} \mathbf{R}_{1j}; \quad \forall i, j \quad (25)$$

and in the same way, we can compute the inner products of  $\tilde{\mathbf{X}}_2^{\text{col}}$  with  $\mathbf{y}_2^{\text{col}}$  based on  $\tilde{\mathbf{l}}$  as

$$\left( \tilde{\mathbf{X}}_2^{\text{col}} \right)_i \cdot \mathbf{y}_2^{\text{col}} = \tilde{\mathbf{l}}_i - \mathbf{R}_{1i} b_1(Z); \quad \forall i. \quad (26)$$

Equations 25 and 26 will allow us to recursively process each subsequent  $\tilde{\mathbf{X}}_i^{\text{col}}$  and  $\mathbf{y}_i^{\text{col}}$  in a way similar to the one used to produce  $\tilde{\mathbf{X}}_2$  and  $\mathbf{y}_2$  and will be referred to as the *lookup tables update equations*. Finally, let us note that once these lookup tables have been computed, only the first  $N_T$  rows of  $\tilde{\mathbf{X}}$  and  $\mathbf{y}$  are now needed for the algorithm to run.

If we suppose now that  $N_T \ll N$ , then we will have that most of the computation time will be spent in the building of the initial values of the lookup tables (equations 23 and 24). More precisely these will crudely correspond to the DFT of  $\mathbf{w}^2$  and of  $\mathbf{w}^2 \circ \mathbf{s}$ ; their vector multiplication with each combination of the templates plus the inverse transforms leading to these initial values. Despite the fact that the previous derivation is a bit coarse, it still assesses the linearithmic (i.e.  $\mathcal{O}(N \log N)$ ) behaviour of the presented algorithm. Regarding now the specific problem of the QSO redshift determination within the Gaia mission (expected to be  $N = 10^4$ ,  $N_T = 10$ ), tests performed on a 2,4Ghz CPU provide execution times of  $180.35 \pm 6.76$  seconds for the normal equations solution compared to  $0.173 \pm 0.002$  second for our implementation; these become respectively  $4.95 \pm 0.19$  hours compared to  $1.81 \pm 0.02$  second for the case of  $N = 10^5$  and  $N_T = 10$ . Finally, let us note that the proposed algorithm can be easily implemented in parallel given the fact that the estimation of each  $\chi^2(Z)$  can be separately performed. As a consequence, the execution time can be scaled by an arbitrary factor that is inversely proportional to the number of running processes.

## 5 APPLICATION

Unsurprisingly, the performance of the presented method was assessed on type I/II QSOs coming from the SDSS DR12 quasar catalog (P aris et al. 2016). The choice of this catalogue comes from the fact that all spectra contained therein were visually inspected and can hence be considered as being extremely reliable regarding their redshift. Additionally, it is also interesting to note that the latter contains a non-negligible number of 297 301 QSOs that is adequate in order to derive strong statistics.

Due to time constraints and to the need for the WPCA algorithm to have a well covered input space of parameters (i.e. numerous observations), we used a two-fold cross-validation in order to test our method. That is: we split our input catalog into two randomly drawn parts out of which we extract the principal components; then we compute the redshift of spectra belonging to each part based on both the weighted and classical phase correlation algorithms whose inputs are the principal components built on the alternative part. Following is a detailed description of the processes leading to this cross-validation.

## 5.1 Procedure description

Raw spectra are generally not readily exploitable. Rather, we have to reduce them such as to get rid of most of the contaminating signals that encompass, for the specific case of this study: deviant points (amongst which night sky emission lines and spectrograph edge effects) and QSOs continuum. Note that since the SDSS DR12 spectra are already sampled on a uniform logarithmic scale, nothing has to be done in order for equation 1 to be fulfilled but usually spectra have to be resampled.

The estimation of the QSO continuum turns out to be a challenging problem upon which the quality of the principal components and of the redshift prediction strongly depend (Machado 2013). Four broad kinds of approaches have been investigated so far in order to estimate this continuum: (1) the fit of a ‘damped’ power-law function to the observed spectra (Ferland 1996); (2) the use of PCA such as to predict the shape of the Ly $\alpha$  forest continuum based on the red part of the spectrum (Suzuki et al. 2005; Pâris et al. 2011; Lee et al. 2012); (3) the modelling of the dependency between the intrinsic QSO continuum and the absorption that it encounters as a mean to extrapolate it (Bernardi et al. 2003) and (4) through the use of techniques related to the multiresolution analysis (Dall’Aglio et al. 2008; Machado 2013). We choose to use this last alternative based on the fact that we do not require the resulting continuum to have a physical basis (i.e. the continuum subtraction being rather used as a normalization) and on the fact that we would like to have the most empirical estimation of this continuum. Following Machado (2013), we found that the signature of the continuum clearly stands within the low frequency components of the pyramidal median transform (Starck 1996, hereafter PMT) of the input spectrum. In practice, the PMT is computed on a flipped version of the spectrum concatenated with the original version and another flipped version such as to ensure continuity at the border. After taking the inverse transform through a third degree fitting polynomial, we enforce the smoothness of the solution by convolving it with a thousand points-wide Savitzky-Golay filter such as to provide the final continuum. Besides its accuracy, we have to note that the PMT, from its pyramidal nature, has an algorithmic complexity of  $\mathcal{O}(N \log N)$  and will consequently not degrade the performances of the global process.

After having subtracted the derived continuum, we discard border regions for which either  $\lambda < 3800\text{\AA}$  or  $\lambda > 9250\text{\AA}$ ; we reject  $4\text{\AA}$  regions around each significant night sky emission lines and finally we perform a  $k$ -sigma clipping ( $k = 3$ ,  $\sigma = 4$ ) on the two first scales of the PMT such as to remove extremely deviant points. Finally, we get an estimate of the signal-to-noise ratio (hereafter SNR) of each continuum-subtracted spectrum through the computation of a ‘noiseless’ spectrum coming from the hypothesis that the noise within these spectra is entirely contained within the five first scales of the biorthogonal spline stationary wavelet transform of each spectrum (Cohen & Daubechies 1992; Burrus 1997). Practically, a spline of third degree was used for both analysis and synthesis. Figure 3 illustrates the result of the initial reduction process.

Spectra having an estimated SNR greater than 1 are then set on a common logarithmic wavelength scale with a uniform sampling of  $\Delta \log_{10} \lambda = 10^{-4}$ , equal to the origi-

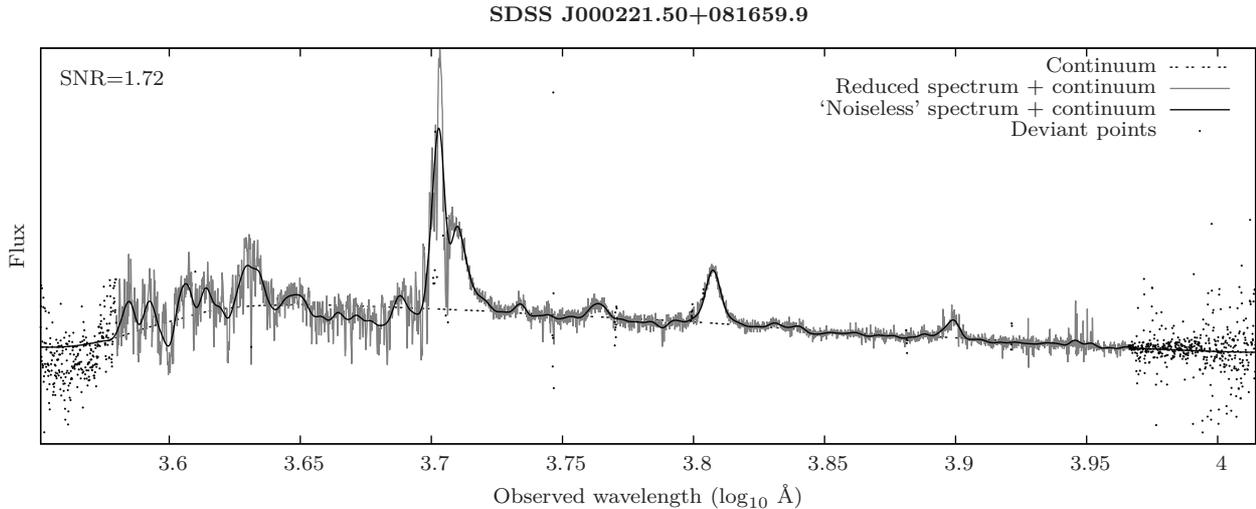
nal sampling of the spectra. The 116 374 resulting spectra are then divided into two equal parts –called learning sets– each of which is being used to produce the principal components and mean observations associated with each part of the cross-validation process. Resulting from this subdivision, we will have that the input catalogue will be split into two parts –the test sets– each consisting in 133 860 observations. Note that given the fact that the broad absorption line QSOs are discarded, both sets do not sum up to 293 301 QSOs.

We then compute the classical and weighted CCF of each spectrum contained within the two test sets based on the mean observation and ten first principal components coming from the alternative learning set. Out of these CCF we extract the five most significant peaks –having a separation of at least  $15,000\text{km s}^{-1}$ – and we fit them with a second order polynomial such as to gain a sub-sampling precision on the predicted peak position. Note that we choose to consider multiple solutions based on the fact that the most significant peaks may not always lead to a physical basis. For example, we might have deep absorption lines either coming from the host galaxy of the quasars or from extragalactic objects being located along the line-of-sight during acquisition and leading to ‘negative’ fitted emission lines. These can definitely prevent the highest peak –the one with the associated minimal  $\chi^2$ – from being the effective one. In order to discriminate between these five selected solutions, we define two score measures:  $\chi_r^2(z)$ , defined as the ratio of the value of the peak associated with the redshift  $z$  to the value of the maximal peak and  $Z_{\text{score}}(z)$ , defined as

$$Z_{\text{score}}(z) = \prod \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{e_\lambda}{\sigma(e_\lambda) \sqrt{2}} \right) \right], \quad (27)$$

where  $e_\lambda$  are the mean values of the emission lines covered by the observed spectrum if we consider it to be at redshift  $z$  and where  $\sigma(e_\lambda)$  are the associated uncertainties. Note that both  $e_\lambda$  and  $\sigma(e_\lambda)$  are computed over a range of eleven points surrounding each emission line. We can recognize each term of equation 27 as being the cumulative distribution function of a normal distribution of mean zero and variance  $\sigma^2(e_\lambda)$  evaluated at  $e_\lambda$ . The use of equation 27 allows us to have a numerical estimate of the ability for a given redshift,  $z$ , to grab the following chosen QSO emission lines: O VI  $\lambda 1033$ ; Ly $\alpha$   $\lambda 1215$ ; N V  $\lambda 1240$ ; Si IV  $\lambda 1396$ ; C IV  $\lambda 1549$ ; C III]  $\lambda 1908$ ; Mg II  $\lambda 2797$ ; H  $\gamma$   $\lambda 4340$ ; H  $\beta$   $\lambda 4861$  and H  $\alpha$   $\lambda 6562\text{\AA}$ . Typical values of  $Z_{\text{score}}(z)$  range from  $\sim 1$  for a solution with a clear match of all positive emission lines while it voluntarily penalizes solutions with a match of at least one ‘negative’ emission line by giving them a  $Z_{\text{score}}(z) \sim 0$ , values in between often occur in low SNR spectra or spectra with strongly damped emission lines. Finally, an error on each estimated peak position is derived and will be further described in section 6.1.

For each spectra, the selection of the optimal redshift out of the five potential ones,  $z_1, \dots, z_5$  for which  $1 = \chi_r^2(z_1) \geq \chi_r^2(z_2) \geq \dots \geq \chi_r^2(z_5)$  and coming either from the classical CCF or from the weighted CCF is done in the following way: if  $Z_{\text{score}}(z_1) > 0.8$ , then select  $z_1$ ; otherwise choose the shift having the highest  $\chi_r^2$  and for which both  $Z_{\text{score}}(z_i) > 1 - 10^{-6}$  and  $\chi_r^2(z_i) > 0.8$ ; otherwise choose the shift having the highest  $Z_{\text{score}}$  and for which  $\chi_r^2(z_i) > 0.9$ . Note that the previous selection and constants therein are purely empirical and based on an iterative visual inspection



**Figure 3.** Result of the initial reduction process designed to decompose any input spectrum into: a continuum spectrum through the smoothing of the low-frequency components of the PMT; a ‘noiseless’ spectrum through the removal of high-frequency components of the biorthogonal spline stationary wavelet transform and a set of deviants points through a  $k$ -sigma clipping of the first two scales of the PMT as well as through the removal of night sky emission lines and border regions. Also computed is the SNR of the continuum-subtracted spectrum. The continuum drop occurring at wavelengths shorter than the Ly $\alpha$  limit can be explained by the absorption induced by the intergalactic medium –mainly composed of hydrogen atoms– located along the line-of-sight towards the QSO under study (Petitjean et al. 1993).

of misclassified spectrum. This final step provides us with what we thought to be the most probable redshift estimate for a given input spectrum along with the associated uncertainty and a warning flag notifying a failure and/or imprecision in the CCF computation; in the peak identification or in the redshift selection (e.g. all fluxes to zero, low  $Z_{\text{score}}$  or less precise uncertainties).

## 5.2 Results

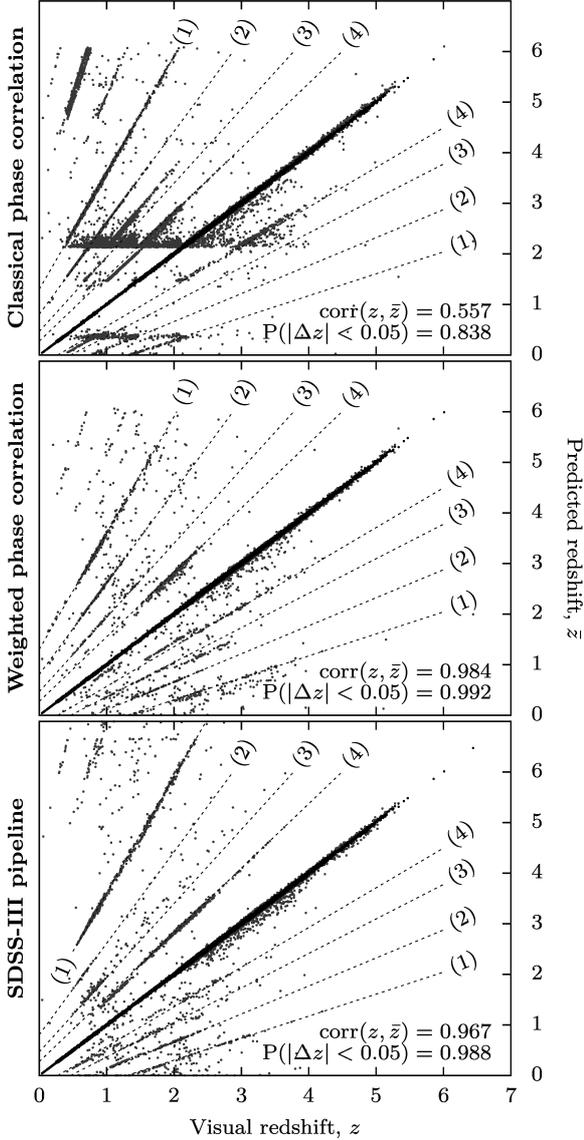
Figure 4 illustrates the result of the cross-validation process for both the classical phase correlation and weighted phase correlation algorithms and further illustrates a comparison with the redshift predicted by the SDSS-III pipeline. We can readily see that the performances of the classical phase correlation algorithm are strongly degraded compared to the weighted version with a correlation factor of 0.557 compared to 0.984 and a ratio of observations having  $|\Delta z| < 0.05$  of 0.838 compared to 0.992, respectively. These differences mainly come from the previously mentioned problem of border flattening that translates into frequent emission line mismatches and into errors coming from the difficulty that the algorithm has in order to extrapolate the regions surrounding the Ly $\alpha$  and H $\alpha$  emission lines. This difficulty arise because of the prominence of these lines as well as because of the high correlation they have with the other emission lines (Yip et al. 2004). As a consequence, the algorithm is often constrained to consider the Ly $\alpha$  or H $\alpha$  lines to be embedded within the observed spectra which graphically results in a gap around  $0.4 < \bar{z} < 2.12$ . Note that the systematic errors occurring at  $\bar{z} \sim 0.4$  and at  $\bar{z} \sim 2.12$  can be attributed to the fitting of these specific emissions lines to the residual spectrograph edge effects –particularly significant within the low

SNR spectra– and that these errors account for  $\sim 2\%$  of the observations having  $|\Delta z| \geq 0.05$ .

Investigation of the most significant errors coming from the emission lines mismatch, illustrated in figure 4, shows that the latter can be modelled as a linear relation between the predicted redshift and the effective redshift. Indeed, if we consider an emission line observed at wavelength  $\lambda$  and falsely considered to stand at a restframe wavelength  $\lambda_f$  instead of  $\lambda_t$ , we will have that the predicted redshift,  $z_f$ , can be related to the effective redshift,  $z_t$ , through

$$\frac{z_f + 1}{z_t + 1} = \frac{\lambda_t}{\lambda_f}. \quad (28)$$

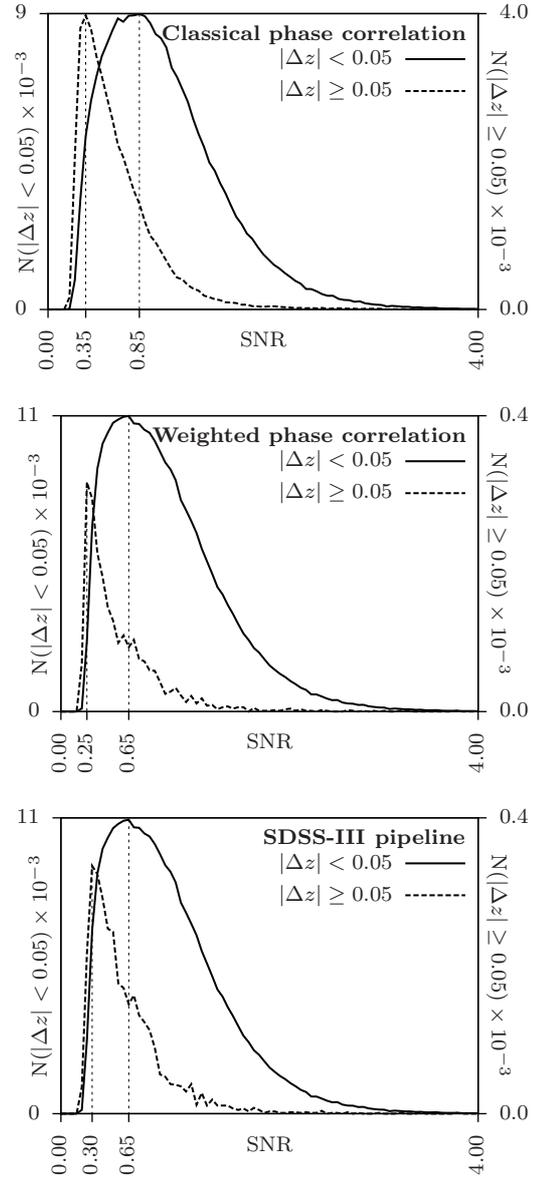
These mismatches do not constitute, in themselves, real cases of degeneracy regarding our  $\chi_r^2$  and  $Z_{\text{score}}$  selection criteria. Indeed, each of the configuration mentioned within figure 4 has unconfused emission lines that make the resulting redshift unambiguous. Rather, the observed degeneracies also come from the low SNR of the observed spectra. Figure 5 illustrates the distribution of the SNR of both the observation having  $|\Delta z| \geq 0.05$  and those having  $|\Delta z| < 0.05$  for our three cases of study. We notice that for all three cases, the SNR of the maximal peak of the fair redshifts estimate is approximately twice the one of the erroneous ones, this is especially significant in the cases of the weighted phase correlation and of the SDSS-III pipeline where the errors come nearly exclusively from this line mismatch problem. Furthermore, a visual inspection of these degenerated spectra shows both potential redshifts to be undistinguishable from one another in most of the cases and thus constituting *in fine* effective cases of degeneracy. Consequently, some of the low SNR spectra will unavoidably have ambiguous redshift estimates that will stand in well specific regions defined by equation 28. Nevertheless, these will be easily identified



**Figure 4.** Results of the classical phase correlation and weighted phase correlation algorithms based on the cross-validation of observations coming from the SDSS DR12Q quasar catalog plus predictions coming from the SDSS-III pipeline. Provided information are the correlation factor between  $z$  and  $\bar{z}$  and  $P(|\Delta z| < 0.05)$ , the ratio of observations having an absolute error lower than 0.05 (depicted as black dots). Dotted numbered lines correspond to known mismatches between common emission lines: (1) Mg II with Ly $\alpha$ ; (2) Mg II with C IV; (3) Mg II with C III] and (4) C IV with Ly $\alpha$  (see section 5.2).

as having a low  $Z_{\text{score}}$  and/or a low redshift confidence (see section 6.1).

Finally, we may notice that our implementation seems to have a better tolerance to noise compared to the SDSS-III implementation (i.e. see within figure 5 the lower peak of the erroneous curve as well as its globally smaller width). This higher tolerance does not come from differences in the algorithms since both implementations are based on the sole solution to equation 11, but either: (1) from the higher num-



**Figure 5.** Histograms of the SNR distribution amongst the observations having  $|\Delta z| < 0.05$  and those having  $|\Delta z| \geq 0.05$  for the cases of the classical phase correlation; of the weighted phase correlation and of the SDSS pipeline output. A bin width of  $\Delta \text{SNR} = 0.1$  was used in each of these graphs. Also highlighted is the position of the maximal peak of each histogram.

ber of PCA components we used (11 compared to 4); (2) from the fact that the components we used were more suitable in order to represent the observed spectra or (3) from the fact that the redshifts coming from the visual inspection procedure are also subject to errors, especially since we are concerned with low SNR spectra where degeneracy may occur.

In order to reject the fact that this higher tolerance comes from the larger number of components we used, we repeat the described cross-validation procedure by using only three components (plus mean observation) instead of ten. The results of this configuration lead us to the same conclu-

sions with a correlation factor of 0.976 (compared to 0.967) and a ratio of observations having  $|\Delta z| < 0.05$  of 0.989 (compared to 0.988). Although the differences in the erroneous SNR curves are less perceptible, it still remains globally sharper. Furthermore, we have to mention that within the SDSS-III pipeline, no more than four principal components were used because any larger number of components would make the error higher. In regard to this point and to the fact that we succeed in getting good predictions using 11 components, we might suppose, in anyway, that the components we used were of higher quality in order to model this specific dataset. Nevertheless, let us mention that we cannot totally reject the hypothesis according to which this better tolerance comes from a fortuitous statistical fluctuation itself produced by the degeneracy occurring during the visual inspection of some low SNR spectra.

## 6 DISCUSSION

### 6.1 Redshift confidence & uncertainty estimation

In order for the derived redshift to be effectively used within subsequent scientific applications it is mandatory for it to come along with an estimation on its uncertainty and to have a confidence level that the chosen redshift is indeed in the vicinity of the real redshift. To make it clear, we may have a redshift estimation with a reasonable uncertainty (e.g.  $z = 2.31 \pm 10^{-3}$ ) but being degenerated such that we are not sure that it stands in the neighbourhood of the effective redshift. Fortunately, the computed CCF offers us simple and efficient ways to evaluate both the redshift uncertainty as well as the confidence we can set on it.

Generally speaking, we know that for a sufficiently large sample of observed points, the  $\chi^2$  map defined in the parameters  $\{a_1, \dots, a_n\}$  can be approximated in the neighbourhood of the global minimum,  $\{a_1^*, \dots, a_n^*\}$ , as

$$\chi^2(a_i) \approx \frac{(a_i - a_i^*)^2}{\sigma^2(a_i^*)} + C, \quad (29)$$

where  $C$  is a function depending on  $a_j$ ,  $j \neq i$  and thus considered here as a constant. In other words, the approximation of the  $\chi^2$  map near a global minimum can be evaluated for each of the parameters independently from the others as a simple quadratic curve whose curvature depends on the uncertainty of the varying parameter. As a consequence, if  $\chi^2(a_i)$  increases by one compared to the optimal  $\chi^2$ , then we will have that  $\sigma^2(a_i) = (a_i - a_i^*)^2$ . The reader may find in [Bevington & Robinson \(2003, section 8.1\)](#) more informations about the variation of the  $\chi^2$  near the optimum and more particularly about the derivation of equation 29. Regarding the uncertainty on the predicted redshifts, we used a second order polynomial such as to fit the optimal peak of the CCF,  $Z$ , and derived its associated uncertainty<sup>5</sup>,  $\sigma(Z)$ . We then use the propagation of the uncertainty such as to get the error on the estimated redshift

$$\sigma(z) = (z + 1)\sigma(Z)s \log b, \quad (30)$$

<sup>5</sup> Beware that the shift value corresponding to the uncertainty will have an associated decrease by one compared to the maximal peak of the CCF.

where  $b$  is the base of the logarithmic scale we used (in our case  $b = 10$ ) and  $s$  is the sampling of the spectra on this logarithmic scale (in our case  $s = \Delta \log_{10} \lambda = 10^{-4}$ ).

Secondly, we have to evaluate the confidence we can have on the predicted redshift. First estimators of this confidence are the already mentioned  $Z_{\text{score}}(z)$  and  $\chi_r^2(z)$  (see section 5.1). Indeed, a secure estimate will typically have  $Z_{\text{score}}(z) \approx \chi_r^2(z) \approx 1$ . Unfortunately, these estimators do not take into account the potential ambiguity that might be present during the selection of the CCF peak associated with the predicted redshift. In order to tackle this lack, we defined the chi-squared difference associated with a redshift estimate,  $z_i$ , as

$$\Delta\chi_r^2(z_i) = \min |\chi_r^2(z_i) - \chi_r^2(z_j)|, \quad \forall j \neq i, \quad (31)$$

where each  $z_j$  corresponds to a redshift associated with a peak selected within the CCF. We may notice that any redshift being unsure due to the ambiguity in the CCF peak selection will now be marked as having  $\Delta\chi_r^2(z) \approx 0$ . Also note that compared to [Bolton et al. \(2012\)](#), we use the distance between all  $\chi_r^2$  and not only those for which  $\chi_r^2(z_i) \geq \chi_r^2(z_j)$  because we adopt the hypothesis that any solution having  $\chi_r^2(z_i) < \chi_r^2(z_j)$  might have been falsely rejected while being a valid solution.

### 6.2 Dealing with zero weights

It commonly happens for the weight matrix,  $\mathbf{W}$  from equation 11, to have a lot of successive weights set to zero, this is especially true if we consider that the observation can be padded such as either to match the size of the templates or to deal with the periodic nature of the phase correlation. Additionally, nothing prevents us from shifting the observation in a circular way such as to have this set of successive zeroes being in the first rows of the weights matrix and to later get rid of this artificial shift by sliding back the CCF. This is particularly interesting if we have a number of zeroed weights equal to –or greater than– the number of components we used,  $N_T$ . In this case, the  $N_T$  first rows of the matrix  $\tilde{\mathbf{X}}$  and of the vector  $\mathbf{y}$  used within the factorized QR algorithm with lookup tables will all be equal to zero and as a consequence, none of the  $\tilde{\mathbf{X}}_i$ 's, as well as none of the  $\mathbf{y}_i$ 's have to be computed. Differently stated, in addition to the building of the lookup tables and their updates, we solely have to compute  $\mathbf{R}_{ij}$  and  $b_i(Z)$  through equations 16 and 17:

$$\mathbf{R}_{ij} = -\frac{\tilde{\mathbf{l}}_{ij}}{\sqrt{\tilde{\mathbf{l}}_{ii}}} \quad (32)$$

and

$$b_i(Z) = -\frac{\tilde{\mathbf{l}}_i}{\sqrt{\tilde{\mathbf{l}}_{ii}}}. \quad (33)$$

This allows us to greatly simplify our algorithm and leads to execution times of  $0.082 \pm 0.001$ s for the case  $N = 10^4$ ,  $N_T = 10$  and of  $0.912 \pm 0.026$ s for the case  $N = 10^5$ ,  $N_T = 10$ . A rough comparison shows these execution times to be twice faster than those presented at the end of section 4.1.1.

### 6.3 Templates weighting

Although, the weighting of the observed spectra is the most important regarding the redshift determination of QSOs, one might also want to have a template weighting such as, for example, to highlight some patterns or to reflect the fact that these templates often come along with their own uncertainties. To this aim, we plug into equation 11, the diagonal matrix of weights associated with the template observations,  $\mathbf{W}_T$ , that is

$$\chi^2(Z) = \left\| \tilde{\mathbf{W}}_T \mathbf{W}_T \mathbf{s} - \mathbf{W} \tilde{\mathbf{W}}_T \tilde{\mathbf{T}} \mathbf{a}(Z) \right\|^2 = \left\| \tilde{\mathbf{y}} - \tilde{\mathbf{X}} \mathbf{a}(Z) \right\|^2. \quad (34)$$

After orthogonalization of the matrix  $\tilde{\mathbf{X}} = \tilde{\mathbf{Q}} \tilde{\mathbf{R}}$ , we get to

$$\chi^2(Z) = \|\tilde{\mathbf{y}}\|^2 - \|\mathbf{b}(Z)\|^2, \quad (35)$$

with the first  $N_T$  elements of  $\mathbf{b}(Z)$  being equal to the first  $N_T$  elements of  $\tilde{\mathbf{Q}}^T \tilde{\mathbf{y}}$ . We can already note that since  $\tilde{\mathbf{y}}$  is now shift-dependent, the knowledge of  $\mathbf{b}(Z)$  alone is no more sufficient in order to find the optimal shift such that  $\chi^2(Z)$  must be explicitly evaluated through equation 35.

Computation of the first  $N_T$  elements of  $\mathbf{b}(Z)$  is straightforwardly done using the procedure described in section 4.1.1 with  $\tilde{\mathbf{y}}$  replacing  $\mathbf{y}$  and both lookup tables given by

$$\tilde{\mathbf{l}}_{ij} = \mathcal{F}^{-1} \left[ \mathcal{F} \left[ \mathbf{W}_T^2 \left( \mathbf{T}_i^{\text{col}} \circ \mathbf{T}_j^{\text{col}} \right) \right]^* \circ \mathcal{F} \left[ \mathbf{W}^2 \right] \right]_Z \quad (36)$$

and

$$\tilde{\mathbf{l}}_i = \mathcal{F}^{-1} \left[ \mathcal{F} \left[ \mathbf{W}_T^2 \mathbf{T}_i^{\text{col}} \right]^* \circ \mathcal{F} \left[ \mathbf{W}^2 \mathbf{s} \right] \right]_Z. \quad (37)$$

Finally, we will have that each  $\|\tilde{\mathbf{y}}\|^2$  will be given by

$$\|\tilde{\mathbf{y}}\|^2 = \mathcal{F}^{-1} \left[ \mathcal{F} \left[ \mathbf{W}_T^2 \right] \circ \mathcal{F} \left[ \mathbf{W}^2 \mathbf{s}^2 \right] \right]_Z. \quad (38)$$

## 7 CONCLUSIONS

We have presented a new method for computing the weighted phase correlation of an observed input signal against not necessarily orthogonal templates. This method is found to be the preferred alternative to the classical phase correlation in the case of input observations having a limited coverage and/or having very distinct weights. The implementation of this method is based on a weighted chi-squared problem solved through a highly modified version of the QR orthogonalization algorithm designed to take benefit of the performances of the fast Fourier transform such as to compute the numerous inner products present within the original QR algorithm. This implementation provides us with a numerically stable algorithm having a linearithmic time complexity that makes it affordable for the tight spectral processing of QSOs within the Gaia mission.

We have presented a complete application of this method to the case of the redshift determination of type I/II QSOs coming from the SDSS DR12 quasar catalog through a two-fold cross-validation procedure. This application is based on templates coming from the weighted principal components analysis decomposition of independent spectra coming from the same catalog. We described in detail the reduction of those input spectra as well as the method we used in order to select the most probable redshift amongst the set of possible ones. Results of this cross-validation show

our method to be the one of predilection for QSO redshift determination and is comparable to the SDSS-III pipeline output while not being a  $\mathcal{O}(N^2)$  process.

Finally, we showed how we can get both the uncertainty on the predicted redshift as well as the confidence we can set on it. We further discuss two extensions of our method, namely: the time saving we can get if having a sufficient number of successive zeroed weights and the weighting of the template observations.

A free implementation of the described algorithm has been released under the GNU Public License<sup>6</sup> and can be freely downloaded at <https://github.com/ldelchambre/wcorrQRL>.

## ACKNOWLEDGEMENTS

The author acknowledges support from the ESA PRODEX Programme ‘Gaia-DPAC QSOs’ and from the Belgian Federal Science Policy Office.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation and the U.S. Department of Energy Office of Science. The SDSS-III web site is <http://www.sdss3.org/>.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofísica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington and Yale University.

## REFERENCES

- Alam, S., Albareti, F.D., Allende Prieto, C., et al. 2015, ApJS, 219, 12
- Aubourg, É., Bailey, S., Bautista, J. E., et al. 2014, arXiv:1411.1074
- Bailer-Jones C.A.L. et al., 2013, A&A, 559, A74
- Bailey S., 2012, PASP, 124, 1015
- Bernardi, M., Sheth, R. K., SubbaRao, M., et al. 2003, AJ, 125, 32
- Bevington, P.R. Robinson, D.K., 2003, Data reduction and error analysis for the physical sciences, 3rd edn., McGraw-Hill
- Bishop, C., 2006, Pattern Recognition and Machine Learning, 1st edn. Springer-Verlag, New York
- Brault J.W., White O.R., 1971, A&A, 13, 169
- Bolton, A. S., Schlegel, D. J., Aubourg, É., et al. 2012, AJ, 144, 144

<sup>6</sup> <http://www.gnu.org/licenses/gpl-3.0.txt>

- Burrus C.S., Gopinath R.A., Guo H., 1997, Introduction to Wavelets and Wavelet Transforms: A Primer, 1st edn. Prentice Hall, London
- Cabanac, R. A., de Lapparent, V., & Hickson, P. 2002, *A&A*, 389, 1090
- Cohen A., Daubechies I., Feauveau J.C. 1992, *Communications on Pure and Applied Mathematics*, 45, 485
- Dall’Aglia, A., Wisotzki, L., & Worseck, G. 2008, *A&A*, 491, 465
- de Bruijne, J. H. J. 2012, *Ap&SS*, 341, 31
- Delchambre L., 2015, *MNRAS*, 446, 3545
- Ferland, G. J. 1996, University of Kentucky Internal Report, 565 pages,
- Francis, P. J., Hewett, P. C., Foltz, C. B., & Chaffee, F. H. 1992, *ApJ*, 398, 476
- Glazebrook K., Offer A.R., Deeley K., 1997, *ApJ*, 492, 98
- Golub G.H., Van Loan C.F., 1996, *Matrix Computations*, 3rd edn. The Johns Hopkins Univ. Press, London
- Heavens, A. F. 1993, *MNRAS*, 263, 735
- Jolliffe I.T., 2002, *Principal Component Analysis*, 2nd edn. Springer, New York
- Lee, K.-G., Suzuki, N., & Spergel, D. N. 2012, *AJ*, 143, 51
- Machado, D. P., Leonard, A., Starck, J.-L., Abdalla, F. B., & Jouvel, S. 2013, *A&A*, 560, A83
- Pâris I. et al., 2016, *A&A*, in prep.
- Pâris, I., Petitjean, P., Rollinde, E., et al. 2011, *A&A*, 530, A50
- Pearson K., 1901, *Phil. Mag.*, 2, 559
- Perryman, M. A. C., de Boer, K. S., Gilmore, G., et al. 2001, *A&A*, 369, 339
- Petitjean, P., Webb, J. K., Rauch, M., Carswell, R. F., & Lanzetta, K. 1993, *MNRAS*, 262, 499
- Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P., 2002, *Numerical recipes in C++: The Art of Scientific Computing*, 2nd edn. Cambridge Univ. Press, New York
- Schless J., 2014, preprint ([arXiv:1404.1100](https://arxiv.org/abs/1404.1100))
- Simkin S.M., 1974, *A&A*, 31, 129
- Starck J.L., Murtagh F., Pirenne B., Albrecht M., 1996, *PASP*, 108, 446
- Suzuki, N., Tytler, D., Kirkman, D., O’Meara, J. M., & Lubin, D. 2005, *ApJ*, 618, 592
- Tonry J., Davis M., 1979, *ApJ*, 84, 1511
- Tsalmantza P., Hogg D.W., 2012, *ApJ*, 753, 122
- Yip, C. W., Connolly, A. J., Vanden Berk, D. E., et al. 2004, *AJ*, 128, 2603

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.