Multidim Syst Sign Process (2010) 21:25–43
DOI 10.1007/s11045-009-0086-3

Interpolation of signals with missing data using Principal Component Analysis

P. Oliveira · L. Gomes

Received: 9 November 2006 / Revised: 20 March 2009 / Accepted: 26 March 2009 /
Published online: 10 April 2009

Abstract A non-iterative methodology for the interpolation and regularization of multidimensional sampled signals with missing data resorting to Principal Component Analysis (PCA) is introduced. Based on unbiased sub-optimal estimators for the mean and covariance of signals corrupted by zero-mean noise, the PCA is performed and the signals are interpolated and regularized. The optimal solution is obtained from a weighted least mean square minimization problem, and upper and lower bounds are provided for the mean square interpolation error. This solution is a refinement to a previously introduced method proposed by the author Oliveira (Proceedings of the IEEE international conference on acoustics, speech, and signal processing—ICASSP06, Toulouse, France, 2006), where three extensions are exploited: (i) mean substitution for covariance estimation, (ii) Tikhonov regularization method and, (iii) dynamic principal components selection. Performance assessment benchmarks relative to averaging, Papoulis-Gerchberg, and Power Factorization methods are included, given the results obtained from a series of Monte Carlo experiments with 1-D audio and 2-D image signals. Tight upper and lower bounds were observed, and improved performance was attained for the refined method. The generalization to multidimensional signals is immediate.

Keywords Signal reconstruction · Missing data · Principal Component Analysis · Non bandlimited signals

1 Introduction

The problem of interpolation of multidimensional sampled signals with missing data is central in a series of engineering problems. Autonomous robotic surveying (Pascoal et al. 1997), underwater positioning, remote sensing, digital communications (subject to bursts of destructive interferences), estimation and control in networked systems, and computer vision (when

P. Oliveira (✉) · L. Gomes
Instituto Superior Técnico and Institute for Systems and Robotics, Av. Rovisco Pais, 1,
1049-001 Lisbon, Portugal
e-mail: pjcro@isr.ist.utl.pt

occlusions occurs) are a few of a multitude of examples where data is not available at uniform temporal/spatial rates.

The scientific community has been active for a long time in solving interpolation and reconstruction problems, see Benedetto and Ferreira (2000), Choi and Munson (1998), and Marvasti (2001), and Yen (1956) and the references therein for an in-depth repository of available techniques. Iterative methods such as Papoulis–Gerchberg algorithm (P-G) (Gerchberg 1974; Papoulis 1975, 1973), the Expectation/Maximization (EM) algorithm (Roweis 1998), and the Power Factorization method (Hartley 2003) are the most commonly used. However, the iterative characteristics of these methods, with the correspondent computational burden, the restricted domain of application to bandlimited signals, and the low convergence rates verified, preclude its use in a number of relevant applications.

Primarily motivated by a terrain based navigation problem for underwater autonomous robotic activities (Pascoal et al. 1997; Oliveira 2007), this paper extends previous work of the authors presented in Oliveira (2006), where a new methodology was proposed for the interpolation of signals with missing data, that departed from the aforementioned approaches. In this work a non-iterative methodology for the regularized interpolation of multidimensional sampled signals with missing data, based on Principal Component Analysis (PCA) is proposed. Resorting to unbiased sub-optimal estimators for the mean and covariance of multidimensional signals, corrupted by zero-mean noise, the Principal Component Analysis is straightforward to be computed. The signal interpolation is tackled in the components space, formulating a weighted least squares minimization problem with known optimal solution. Moreover, based on PCA properties, corrected upper and lower bounds (relative to Oliveira 2006) for the mean square interpolation error and the interval of validity of the proposed method are provided. Moreover, relative to the basic solution previously proposed by the authors, three refinements are exploited: (i) mean substitution, (ii) Tikhonov regularization and, (iii) dynamic principal components selection. It is important to remark that not only the intervals of validity of the resulting methods are extended but these methods also outperform the basic one.

Principal Component Analysis has already been used in interpolation problems with sampled signals with incomplete data. In Shum et al. (1995), PCA is applied to sparse data from segmented images (not directly on the complete signal, as in the present work). Also in Blanz and Vetter (2002), PCA is computed from signals in a database (without missing data), and is then used to perform a convex mixture of the base signals.

The structure of the paper is the following: Sect. 2 introduces unbiased estimators for the mean and covariance of discrete time multidimensional signals and the PCA computation procedure. Section 3 describes an efficient and unbiased estimator for the mean and covariance accounting for missing data and an optimal solution for the interpolation of signals. No assumption on the stochastic distribution of the noise present is required except that it is a zero-mean process. Lower and upper bounds for the interpolation error variance are deduced, exploiting the properties of the optimal transformation at hand, and the interval of validity of the basic methodology is discussed. Based on the work developed in the previous section, in Sect. 4, a number of refinements to the covariance estimator and for the optimal minimization problem are introduced and discussed. A classic technique to deal with missing samples for the covariance estimation is considered and a dynamic choice of parameters, together with a regularization method, are presented. Results from a series of Monte Carlo experiments with 1-D audio and 2-D image signals are summarized in Sect. 5, allowing the performance assessment of the proposed method. Additionally, alternative methods for interpolation of signals with missing data are tested for an in-depth benchmark of the proposed interpolator. Finally, some conclusions are drawn and future work is unveiled in Sect. 6.

2 PCA for signals

Principal Component Analysis was developed independently by Karhunen in statistical theory and generalized by Loève, based on a method previously introduced by Pearson and applied to psychometry by Hotelling, as detailed in Jolliffe (2002), Mertins (1999), and in the references therein.

Considering all linear transformations, PCA, based on the Karhunen-Loève (KL) transform, allows for the optimal approximation to a stochastic signal in the least square sense. It is a widely used signal expansion technique, featuring uncorrelated coefficients, with superior performance in dimensionality reduction. These features make PCA an interesting methodology for many multidimensional signal processing applications such as data compression, image and voice processing, data mining, exploratory data analysis, pattern recognition, and time series prediction (Jolliffe 2002).

2.1 Mean and covariance

Consider a set of M multidimensional signals $\mathbf{x}_i \in l_2$, i.e. with finite energy, where $i = 1, \ldots, M$, from a discrete time real-valued stochastic process corrupted by zero mean noise, represented as column vectors of length N after a trivial stacking operation.

When computing PCA for a set of multidimensional signals, unbiased and efficient estimators for the mean and covariance of those signals are required. Basic results will now be introduced.

Proposition 1 *For a set of signals* $\mathbf{x}_i$*, where* $i = 1, \ldots, M$

(i) the estimator for the j^{th} *component of the ensemble mean* $\mathbf{m}_x(j)$, $j = 1, \ldots, N$ *is given by:*

$$\mathbf{m}_x(j) = \frac{1}{M} \sum_{i=1}^{M} \mathbf{x}_i(j), \quad j = 1, \ldots, N;$$

(ii) the estimator for the covariance element $\mathbf{R}_{xx}(j,k)$, $\{j,k\} = 1, \ldots, N$, *where* $\mathbf{y}_i = \mathbf{x}_i - \mathbf{m}_x$,

$$\mathbf{R}_{xx}(j,k) = \frac{1}{M-1} \sum_{i=1}^{M} \mathbf{y}_i(j)\mathbf{y}_i(k),$$

are unbiased and efficient. Moreover, $\mathbf{m}_x \in l_2$ *and* $\|\mathbf{R}_{xx}\|$ *is finite.*

The proof of this proposition resorts to basic statistical signal processing theory that can be found for instance in Kay (1993).

2.2 Principal Component Analysis

The Principal Component Analysis can be carried out resorting to the KL transform, following the classical approach. The objective is to find an orthogonal basis to decompose a stochastic signal $\mathbf{r} \in l_2$, from the same original space, to be computed as $\mathbf{r} = \mathbf{U}\mathbf{v} + \mathbf{m}_x$, where the vector $\mathbf{v} \in l_2$ is the projection of $\mathbf{r}$ in the basis $\mathbf{v} = \mathbf{U}^T(\mathbf{r} - \mathbf{m}_x)$. The matrix $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_N]$ is composed by the N orthogonal column vectors of the basis, verifying the eigenvalue problem

$$\mathbf{R}_{xx}\mathbf{u}_j = \lambda_j \mathbf{u}_j, \quad j = 1, \ldots, N, \ \ \mathbf{u}_j \in l_2. \tag{1}$$

Assuming that the eigenvalues are ordered, i.e. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$, the choice of the first $n << N$ principal components, leads to an approximation to the stochastic signals given by the ratio on the covariances associated with the energy of the components, i.e. $\sum_{j=1}^{n} \lambda_j / \sum_{j=1}^{N} \lambda_j$.

Departing from the perfect interpolation setup (Unser 2000), the matrix $\tilde{\mathbf{U}} = [\mathbf{u}_1\, \mathbf{u}_2 \ldots \mathbf{u}_n]$ with dimensions $R^{N \times n}$ is used as the approximate PCA associated transformation, thus obtaining a sub-optimal solution. In many applications where stochastic multidimensional signals are the key to tackle the problem at hand, this approximation can lead to large dimensional reduction and thus to a computational complexity optimization.

The advantages of PCA are threefold: (i) it is an optimal (in terms of mean square error) linear scheme for compressing a set of high dimensional vectors into a set of lower dimensional vectors; (ii) the model parameters can be computed directly from the ensemble covariance; (iii) given the model parameters, projection into and from the bases are computationally inexpensive operations $\mathcal{O}(nN)$.

3 Interpolation using PCA

The purpose of this section is to describe a methodology, supported on PCA, allowing the interpolation of multidimensional sampled signals with missing data, corrupted by zero mean noise, based on the following assumption, central to the rest of this work:

Assumption 3.1 The missing information on the multidimensional sampled signals are negligible and the available samples, in a number greater than the selected number of principal components, are representative of the original signal.

The underlying process can result for instance from a non-homogeneous spatial survey, due to physical or kinematic constrains, or associated with the reception of a signal in a communication channel corrupted by bursts of noise that destroy completely the information contained in some samples.

It is important to remark that no assumption on the stochastic characteristics of the noise corrupting the underlying signal is required, except for the null mean, departing from the commonly used Gaussian noise characteristics found in the literature (see examples in Blanz and Vetter (2002) and Roweis (1998)). It is impossible to distinguish between a non null noise mean and the signal itself, however, given the approach for the principal components computation, this disturbance is automatically rejected.

3.1 Mean and covariance estimators with missing data

The estimators for the mean and covariance from Proposition 1 do not take into account possible missing values. Hence, new estimators are proposed and an indicator index $\mathbf{l}$ is introduced, on which is applied the same stacking operation as in the multidimensional signals.

The index $\mathbf{l}_i(j)$, $j = 1, \ldots, N$ is set to 1 if the j^{th} component of signal $\mathbf{x}_i(j)$ is available and zero otherwise. In the latter, the component $\mathbf{x}(j)$ is also set to zero, without loss of generality. Considering the required adjustments to Proposition 1, estimators for the mean and covariance of multidimensional signals with missing data are now presented.

Lemma 1 *Given a set of M signals* $\mathbf{x}_i$, $i = 1, \ldots, M$, *with associated indices* $\mathbf{l}_i$, *the auxiliary vector of counters* $\mathbf{c} = \sum_{i=1}^{M} \mathbf{l}_i$, *and* $\mathbf{C} = \sum_{i=1}^{M} \mathbf{l}_i \mathbf{l}_i^T$:

(i) the estimator for the j^{th} component of the ensemble mean is

$$\mathbf{m}_x(j) = \frac{1}{\mathbf{c}(j)} \sum_{i=1}^{M} \mathbf{l}_i(j)\mathbf{x}_i(j), \quad j = 1, \ldots, N;$$

(ii) the estimator for the covariance element $\mathbf{R}_{xx}(j,k)$, $j,k = 1,\ldots,N$, given $\mathbf{y}_i = \mathbf{x}_i - \mathbf{m}_x$, can be computed from

$$\mathbf{R}_{xx}(j,k) = \frac{1}{\mathbf{C}(j,k) - 1} \sum_{i=1}^{M} \mathbf{l}_i(j)\mathbf{l}_i(k)\mathbf{y}_i(j)\mathbf{y}_i(k).$$

These estimators share the same properties as the ones introduced in Proposition 1.

The proof resorts only to basic signal processing tools and is omitted here (see Jolliffe (2002) and Kay (1993) for details).

3.2 Solution to the interpolation problem

To solve the interpolation problem central to this paper, consider that each signal $\mathbf{x}_i$ is obtained from the original signal $\mathbf{r}_i$ due to missing data, verifying the relation $\mathbf{x}_i = \mathbf{L}_i\mathbf{r}_i$, where $\mathbf{L}_i \in \mathcal{R}^{N\times N}$ is a diagonal matrix, filled with the indicator index $\mathbf{l}_i$. The interpolation operation is formulated as finding $\tilde{\mathbf{r}}_i$ that minimizes the weighted l_2 norm of the error. However, due to the existence of missed samples, it is only possible to compute the estimation error on the components of the signal which are known. Thus, the correct form of formulating the problem is to consider only the interpolation error for the available elements.

Lemma 2 *Considering the original signal $\mathbf{r}_i$, from which there is only available a signal with samples indexed by $\mathbf{L}_i$, the optimal interpolated signal $\tilde{\mathbf{r}}_i$ (in the minimum error energy sense) can be obtained solving the weighted least mean square problem*

$$\min_{\tilde{\mathbf{r}}_i \in \mathcal{R}^N} \|\mathbf{L}_i(\tilde{\mathbf{r}}_i - \mathbf{r}_i)\|^2_{2,\mathbf{W}},$$

given the symmetric positive semi-definite weight $\mathbf{W} \in \mathcal{R}^{N\times N}$, where the solution based on PCA is given by

$$\tilde{\mathbf{v}}_i = (\tilde{\mathbf{U}}^T\mathbf{L}_i\mathbf{W}\mathbf{L}_i\tilde{\mathbf{U}})^{-1}\tilde{\mathbf{U}}^T\mathbf{L}_i\mathbf{W}\mathbf{y}_i. \tag{2}$$

Proof Given that a minimum energy estimation error problem can be formulated and solved as a weighted least mean square error optimization, the problem at hand is written as

$$\min_{\tilde{\mathbf{v}}_i \in \mathcal{R}^n} \|\mathbf{L}_i(\tilde{\mathbf{U}}\tilde{\mathbf{v}}_i + \mathbf{m}_x - \mathbf{r}_i)\|^2_{2,\mathbf{W}} = \|\mathbf{L}_i\tilde{\mathbf{U}}\tilde{\mathbf{v}}_i + \mathbf{L}_i\mathbf{m}_x - \mathbf{L}_i\mathbf{r}_i\|^2_{2,\mathbf{W}},$$

resorting to the approximated (sub-optimal) PCA projection $\tilde{\mathbf{r}}_i = \tilde{\mathbf{U}}\tilde{\mathbf{v}}_i + \mathbf{m}_x$. Through the relations $\mathbf{x}_i = \mathbf{L}_i\mathbf{r}_i$ and $\mathbf{y}_i = \mathbf{x}_i - \mathbf{L}_i\mathbf{m}_x$, the following minimization is then obtained

$$\min_{\tilde{\mathbf{v}}_i \in \mathcal{R}^n} \|\mathbf{L}_i\tilde{\mathbf{U}}\tilde{\mathbf{v}}_i - \mathbf{y}_i\|^2_{2,\mathbf{W}}.$$

This is a weighted version of a linear least square problem, for which a well-known solution exists, resulting in (2), where the relations $\mathbf{L}\mathbf{L}^T = \mathbf{L}$ and $\mathbf{L}^T = \mathbf{L}$ were used. □

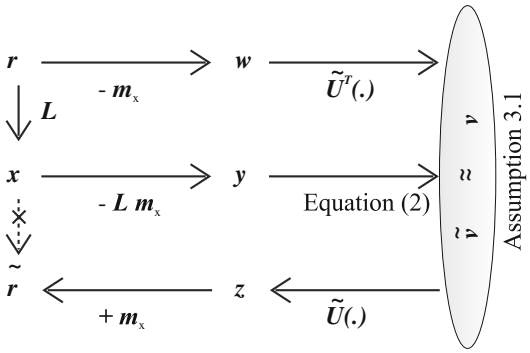

Fig. 1 Diagram describing the interpolation of sampling signals with missing data

From the previous assumption, the principal components can be computed with negligible degradation, and the signal can finally be reconstructed using the relation $\tilde{\mathbf{r}}_i = \tilde{\mathbf{U}}\tilde{\mathbf{v}}_i + \mathbf{m}_x$. The relations among the underlying signals are depicted in Fig. 1. Note that the aforementioned assumption can be interpreted as a change on the focus of the data from sample rates to the amount of information available.

According to optimal stochastic minimization techniques (Kailath et al. 2000), the knowledge on the stochastic process characteristics allows for the optimal choice of the weight $\mathbf{W} = \mathbf{R}_{xx}^{-1}$. Nevertheless, the covariance matrix is estimated from an incomplete data set, which may lead to numerical problems. Next, an approximated and more robust numerical solution is proposed. From (1), the covariance can be decomposed as $\mathbf{R}_{xx} = \mathbf{U}\Lambda\mathbf{U}^T$, where $\mathbf{\Lambda} \in \mathcal{R}^{N\times N}$ is the diagonal matrix, whose k^{th} diagonal element is λ_k. An approximation of the covariance matrix is obtained using the approximated PCA (Mertins 1999), i.e. $\tilde{\mathbf{R}}_{xx} = \tilde{\mathbf{U}}\tilde{\Lambda}\tilde{\mathbf{U}}^T$, where $\tilde{\mathbf{\Lambda}} \in \mathcal{R}^{n\times n}$ contains the eigenvalues corresponding to the first n principal components, thus

$$\tilde{\mathbf{R}}_{xx}^{-1} = \tilde{\mathbf{W}} = \tilde{\mathbf{U}}\tilde{\Lambda}^{-1}\tilde{\mathbf{U}}^T.$$

As the number of components increases, a more accurate result is obtained, with no extra complexity added.

A complexity analysis to the proposed methods previously introduced revealed that the underlying complexity is $O(\eta\, N^2\, M)$. This is a consequence of the application of (2) to the unknown samples, therefore it depends on the amount η of missing samples. It is important to remark that the matrix $\tilde{\mathbf{U}}^T\mathbf{L}_i\mathbf{W}\mathbf{L}_i\tilde{\mathbf{U}}$ to be inverted has dimension $n \times n$, presenting reduced computational complexity, given the choice of $n << N$. Interestingly enough, this result can be interpreted as a generalization of the classical Yen interpolator (Yen 1956) and the minimax-optimal interpolators (Choi and Munson 1998).

3.3 Lower and upper bounds

In order to evaluate the quality of the signal's interpolation, the variance of the interpolation error per signal sample, σ^2, is defined as,

$$\sigma^2 \equiv \frac{E[\|\tilde{\mathbf{r}} - \mathbf{r}\|]}{N-1}.$$

The next step consists of defining the lower and upper bounds for σ^2. In the sequel, assume that n components are used from the total of N available. Lower and upper bounds on σ^2 can be found given the PCA stochastic approximation properties, i.e,

$$\sum_{i=n+1}^{N} \lambda_i \leq E[\|\tilde{\mathbf{r}} - \mathbf{r}\|_2^2] \leq \sum_{i=1}^{N} \lambda_i,$$

where the lower bound represents the situation of best possible interpolation when limited to n principal components, and the upper bound represents the fact that the interpolated signal does not exceed the variance of the original signal.

When evaluating the above expression for the missing data case, it has to be considered the fact that no interpolation error exists on the known samples of the signal. As a result, the bounds must be adjusted to the ratio of missing samples in the signal, represented by $\eta \in [0, 1]$,

$$\eta \sum_{i=n+1}^{N} \lambda_i \leq E[\|\tilde{\mathbf{r}} - \mathbf{r}\|_2^2] \leq \eta \sum_{i=1}^{N} \lambda_i.$$

Scaling the bounds to a per sample basis, leads to the final result

$$\eta \frac{\sum_{i=n+1}^{N} \lambda_i}{N-1} \leq \sigma^2 \leq \eta \frac{\sum_{i=1}^{N} \lambda_i}{N-1}. \tag{3}$$

A simple validation on the bounds for certain values of η can be made. Consider the extreme case where $\eta = 0$, i.e. there is no missing samples. The bounds for this case are both null, which is correct because no interpolation error is present and consequently, a null value for σ^2 is obtained. Now consider the case $\eta \approx 1$, which means that almost no samples of the signal are available and the interpolation will correspond to the signal's variance as stated in the upper bound for high levels of η.

Assumption 3.1 can be interpreted as providing conditions when the interpolation is well posed or when the corresponding numerical tools can be applied. The number of samples available are required to be greater than to the selected number of principal components, i.e. $N(1-\eta) > n$. This leads to the following validity interval deduced from Assumption 3.1,

$$0 \leq \eta < \frac{N-n}{N}. \tag{4}$$

Interestingly enough, no limitation on the amount of missing data was found for the application of the method. Furthermore, the upper bound can be used to study the required performance for a transmission channel or to help on the design of the robotic survey mission, for a given terrain, in order to achieve a prescribed level of interpolation precision.

4 Extensions to the interpolation solutions

In this section, refinements to the solutions proposed in Sect. 3 and in Oliveira (2006) are presented. The motivation for the incorporation of these alternative methods is to include basic and frequently used tools associated to these type of problems. Hence, the extensions to the existing solutions consists of a different approach to the estimator of the covariance in presence of missing data, a regularization method to the least squares minimization problem and an expansion of the validity interval.

4.1 Mean substitution method

When dealing with the estimation of covariance based on missing data, several methods are available (Jolliffe 2002). An important group of such methods corresponds to the techniques

that add no extra computational complexity. A classical technique is the mean substitution method. The missing sample on the j^{th} component of the i^{th} variable is replaced by the corresponding component of the mean, i.e. the missing value $\mathbf{x}_i(j)$ is filled with the value $\mathbf{m}_x(j)$. Although originally the data set has missing samples, due to the mean substitution method, the estimator for the covariance from Proposition 1 is now applicable.

The choice on the mean substitution technique resorts to the fact that it leads to a positive semi-definite covariance matrix, in opposition to the estimator from Lemma 1. Also, better results are achieved for large values of missing data, when compared with common methods as Pairwise or Listwise (Jolliffe 2002). More sophisticated methods can be used, for example the Expectation Maximization (EM) algorithm (Roweis 1998). However, it leads to a more computationally consuming solution.

4.2 Dynamic principal components selection

Consider now the case when Assumption 3.1 is not verified, i.e. the available samples are less than or equal to the assigned number of principal components n. Under this situation, the solution to the minimization problem presented in Lemma 2 is ill-conditioned, resulting a violation on the validity interval given by (4). An alternative approach is suggested next, to be applied in those cases: the number of components used for the computation of the minimizing solution is set to the nearest integer below the current number of available samples. As a result of this procedure, the proposed reconstruction algorithm is extended to any amount of missing samples. Assumption 3.1. is always verified, given the adjustment on the principal components used relative to the existing information. Note that the lower and upper bounds in (3) remain valid throughout the whole interval $\eta \in [0, 1[$.

4.3 Tikhonov regularization

The solution to the interpolation problem introduced in Sect. 3 is obtained from a weighted least squares minimization problem. Thus, a regularization technique can be employed as a complement to the dynamic selection of the number of principal components, to ensure numerical robustness. A commonly used technique is the Tikhonov regularization, for which a well-known solution exists (Tikhonov et al. 1990).

With the purpose of ensuring a suitable reconstruction of the signal, it is desirable to have a smooth transition between the available and the recovered samples. To satisfy this requirement, a regularization term can be added to the reformulated minimization problem. The first order difference matrix $\mathbf{D} \in \mathcal{R}^{(N-1)\times N}$ (Tikhonov et al. 1990) and the auxiliary matrix $\overline{\mathbf{L}}_i \in \mathcal{R}^{N\times N}$, which is a diagonal matrix filled with the complementary values of the indicator index $\mathbf{l}_i$, are considered.

Lemma 3 *Considering the original signal $\mathbf{r}_i$, from which there is only available a signal with samples indexed by $\mathbf{L}_i$, the optimal interpolated and regularized signal $\tilde{\mathbf{r}}_i$, given the auxiliary matrices $\mathbf{D} \in \mathcal{R}^{(N-1)\times N}$ and $\overline{\mathbf{L}}_i \in \mathcal{R}^{N\times N}$, can be obtained solving the weighted least mean square problem with two terms expressed as*

$$\min_{\tilde{\mathbf{v}}_i \in \mathcal{R}^n} \|\mathbf{L}_i(\tilde{\mathbf{r}}_i - \mathbf{r}_i)\|^2_{2,\mathbf{W}} + \|\alpha \mathbf{D}(\mathbf{L}_i \mathbf{r}_i + \overline{\mathbf{L}}_i \tilde{\mathbf{r}}_i)\|^2_2,$$

with the solution that can be obtained resorting also to the PCA decomposition as

$$\tilde{\mathbf{v}}_i = (\tilde{\mathbf{U}}^T \mathbf{L}_i \mathbf{W} \mathbf{L}_i \tilde{\mathbf{U}} + \alpha^2 \mathbf{\Gamma}^T \mathbf{\Gamma})^{-1} (\tilde{\mathbf{U}}^T \mathbf{L}_i \mathbf{W} \mathbf{y}_i - \alpha^2 \mathbf{\Gamma}^T \mathbf{\Delta}), \tag{5}$$

where $\mathbf{\Gamma} \equiv \mathbf{D}\overline{\mathbf{L}}_i\tilde{\mathbf{U}}$ *is the regularization matrix,* $\mathbf{\Delta} \equiv \mathbf{D}(\mathbf{L}_i\mathbf{r}_i + \overline{\mathbf{L}}_i\mathbf{m}_x)$*, and* α *is a regularization parameter.*

Proof The minimization problem, including the specified regularization term, can be written as

$$\min_{\tilde{\mathbf{v}}_i \in \mathcal{R}^n} \|\mathbf{L}_i\tilde{\mathbf{U}}\tilde{\mathbf{v}}_i - \mathbf{y}_i\|^2_{2,\mathbf{W}} + \|\alpha\mathbf{D}(\mathbf{L}_i\mathbf{r}_i + \overline{\mathbf{L}}_i\tilde{\mathbf{r}}_i)\|^2_2.$$

This problem can be rewritten, using the relation $\tilde{\mathbf{r}}_i = \tilde{\mathbf{U}}\tilde{\mathbf{v}}_i + \mathbf{m}_x$, as

$$\min_{\tilde{\mathbf{v}}_i \in \mathcal{R}^n} \|\mathbf{L}_i\tilde{\mathbf{U}}\tilde{\mathbf{v}}_i - \mathbf{y}_i\|^2_{2,W} + \|\alpha\mathbf{D}\overline{\mathbf{L}}_i\tilde{\mathbf{U}}\tilde{\mathbf{v}}_i + \alpha\mathbf{D}(\mathbf{L}_i\mathbf{r}_i + \overline{\mathbf{L}}_i\mathbf{m}_x)\|^2_2.$$

Considering the definitions of $\mathbf{\Gamma}$ and $\mathbf{\Delta}$ above, this minimization problem results in the compact form

$$\min_{\tilde{\mathbf{v}}_i \in \mathcal{R}^n} \left\| \begin{pmatrix} \mathbf{L}_i\tilde{\mathbf{U}} \\ \alpha\mathbf{\Gamma} \end{pmatrix} \tilde{\mathbf{v}}_i - \begin{pmatrix} \mathbf{y}_i \\ -\alpha\mathbf{\Delta} \end{pmatrix} \right\|^2_{2,\mathbf{W}}.$$

This is also a weighted least mean square problem with the solution given by (5). □

The regularization parameter acts as a scaling factor involving the least square term and the regularization term of the minimization problem. For $\alpha = 0$, equation (5) reduces to the unregulated least squares solution presented in (2). Also, the Tikhonov regularization has a filtering feature which rejects principal components that are small (relative to α), while retaining components that are large. Thus, the value for α is set depending on the value of the smallest eigenvalues of the principal components that are desired to be disregarded.

A number of advantages associated to the application of a regularization technique can be delineated; (i) more adequate results are obtained, according to the choice of the regularization term, which privileges suitable solutions; (ii) the unregulated solution of equation (2) may result in an amplification of the corresponding interpolation error, in case of severe lack of samples, leading to an inaccurate result; (iii) there is no significant increment on the computational complexity, as the matrix $(\tilde{\mathbf{U}}^T\mathbf{L}_i\mathbf{W}\mathbf{L}_i\tilde{\mathbf{U}} + \alpha^2\mathbf{\Gamma}^T\mathbf{\Gamma})$ to be inverted, preserves the dimension of the unregulated solution.

5 Results for performance assessment

In this section, the results from the application of the proposed methodologies in Sect. 3 and the respective refinements from Sect. 4, to 1*D* audio and 2*D* image signals, are presented and discussed. Additionally, a performance assessment of other alternative methods is summarized to benchmark the proposed methods. It is important to remark that the generalization to multidimensional signals is immediate.

5.1 Alternative methods for interpolation of signals with missing data

Three alternative methods are considered to solve the interpolation problem in presence of missing data.

5.1.1 Papoulis–Gerchberg method

The first is the Papoulis–Gerchberg algorithm (P-G) (Gerchberg 1974; Papoulis 1975, 1973), that has been extensively used to solve the missing data problem in band-limited signals. It is an iterative method that recovers partially known signals under a smoothness constraint which implies the vanishing of a known subset of the samples of the Discrete Fourier Transform (DFT) of the data. This algorithm can only be applied to band-limited signals, in opposition to the interpolation of multidimensional signals using PCA, which does not rely on any assumption on the signal bandwidth. The implementation of one iteration of this algorithm involves: 1) the computation of the Fast Fourier Transform (FFT) of the signal with the unknown samples, 2) the application of a band-limiting filter, 3) the computation of the Inverse Fast Fourier Transform (IFFT), and finally 4) the replacement of the missing samples by the estimated data. The complexity of the algorithm resorts to $\mathcal{O}\{k \times S(3 + 2\log_2 S)\}$, where S is the total number of samples of the signal, k the number of iterations, and the use of optimized Discrete Fourier Transforms is assumed.

5.1.2 Averaging method

The second technique considered is an averaging technique, proposed as a naive procedure for recovering missing data. It consists of a non-iterative and computationally inexpensive method. The average of a local window, of length H, centered in the missing sample is computed and this value is assigned to the corresponding unknown sample. It is expected that the more complex methods previously proposed always outperforms this conceptually simple averaging technique.

5.1.3 Power factorization method

The third method considered is Power Factorization, as proposed in Hartley (2003), and represents the state of the art for modern reconstruction methods. This approach can be considered as a solution to low-rank approximation problems, in which a matrix of measured data must be approximated by a matrix of given low rank. The proposed method is iterative for approximating a data matrix, possibly with missing entries, with another matrix of small rank r, "provably convergent to a unique global optimum," as claimed informally by the authors. In this work, an implementation available from the authors of the benchmark survey (Buchanan and Fitzgibbon 2005), will be used in the 2D signal, for comparison purposes.

5.2 Results for 1D signals

The performance of the mean and covariance estimators in the presence of missing data previously proposed and the interpolation method introduced above, will be analyzed next when applied to an audio signal. The results obtained with the extensions discussed in Sect. 4 will also be included. Moreover, the bounds deduced for the interpolation under missing data are checked. Finally, the alternative algorithms are assessed. The results are achieved via a series of Monte Carlo experiments (20 for each parameter combination).

An audio signal (voice and musical instruments), as depicted in Fig. 2, was selected for the performance assessment study. The sampling frequency is $8192 Hz$ with a size of $S = 8000$ samples. The length of mosaics is selected as $N = 20$, with a total number of mosaics,

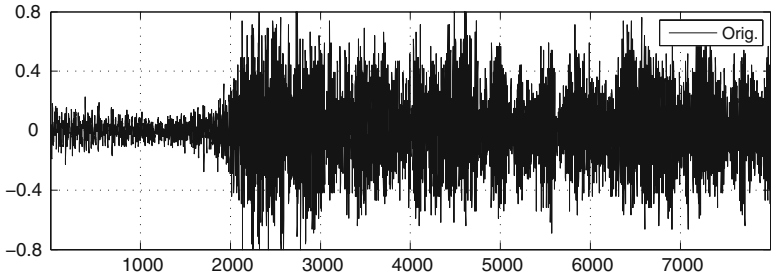

Fig. 2 Example of a 1D Signal, corresponding to the audio signal used for testing, with a length of 8000 samples

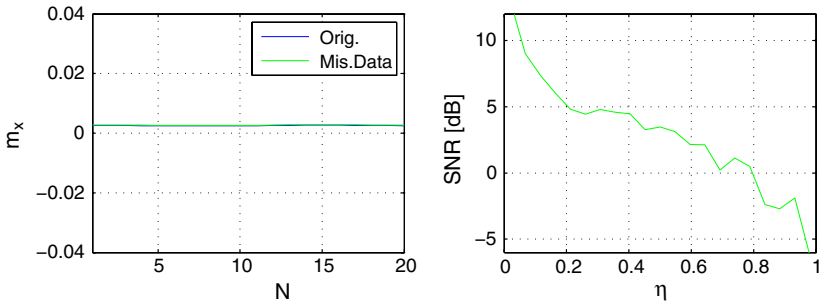

Fig. 3 On the left, mean estimate according to Lemma 1 for $\eta \sim 0.3$ (green). On the right, the $SNR(\mathbf{m}_x, \tilde{\mathbf{m}}_x)$ for $\eta \in [0.02, 0.98]$

$M = S - N + 1 = 7981$. The choice of an audio signal is just for illustrative purposes and it should be remarked that the proposed methodologies are applicable to any type of signal.

5.2.1 Assessment of the mean estimator

The first step in the estimation process is to compute the ensemble mean from the selected M mosaics. The results from applying the mean estimator proposed in Lemma 1, for a missing data $\eta \sim 0.3$, are shown in the left panel of Fig. 3.

Further analysis on the performance is possible when varying the ratio of lost samples over the interval $\eta \in [0.02, 0.98]$. The results are shown through the signal to noise ratio (SNR) computed as,

$$SNR(a, b) = 10\, log_{10}\left(\frac{\|a\|_2}{\|a - b\|_2}\right),$$

for the ensemble mean of the original signal $\mathbf{m}_x$ and for the estimated ensemble mean from the signal with missing samples $\tilde{\mathbf{m}}_x$, i.e. $SNR(\mathbf{m}_x, \tilde{\mathbf{m}}_x)$. The outcome of this experience is included also in Fig. 3, right panel.

5.2.2 Assessment of the covariance estimator

After the estimation of the mean, the estimation of the covariance matrix from the selected mosaics is computed. Results from Lemma 1 on the evolution of some covariance element's

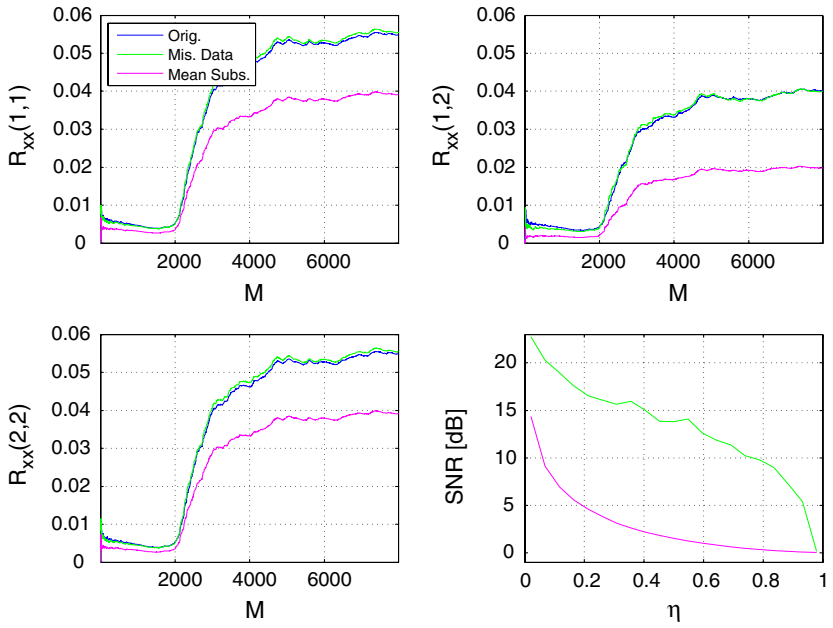

Fig. 4 Examples of the estimate of some covariance elements according to Lemma 1 (green) and mean substitution technique (pink), with $\eta \sim 0.3$. Additionally, $SNR(\mathbf{m}_x, \tilde{\mathbf{m}}_x)$ for each technique for $\eta \in [0.02, 0.98]$, in the lower left panel

estimation (in green) are depicted in Fig. 4. Additionally, in the interval $\eta \in [0.02, 0.98]$, it is computed the signal to noise ratio for the covariance matrix of the original signal and for the estimated covariance matrix from the signal with missing samples, i.e. $SNR(\mathbf{R}_{xx}, \tilde{\mathbf{R}}_{xx})$, with the results incorporated in the lower right panel of Fig. 4.

From the results obtained, it can be concluded that the estimators introduced in Lemma 1 are accurate since the obtained values for the mean and covariance are close to the computation of the covariance with the full signal, thus the estimators are shown to be efficient. Moreover, it always outperforms the mean substitution method, given the fact that the mean substitution method leads to a biased estimator. A smooth degradation along the increasing amount of missing data is verified for both techniques.

5.2.3 Applying the interpolation method

The results of the interpolation method to the audio signal (a l_2 signal), based on the mean and covariance estimators introduced in Lemma 1 can be found in Fig. 5, namely for the interval [2000, 2400], with a ratio of missing samples $\eta = 0.3$ (in red in Fig. 5). Note that the interpolated signal (in green) recovers accurately the missing information when compared to the original signal (in black).

A performance study for a variation on the ratio of lost samples in the interval $\eta \in [0.02, 0.98]$ is depicted in Fig. 6, for the basic method in Sect. 2 and for the possible combination of refinements. The dynamic number of principal components is always employed so that the method is not limited by the validity interval. The regularization parameter was set to $\alpha = 0.6\eta^2$, when applicable.

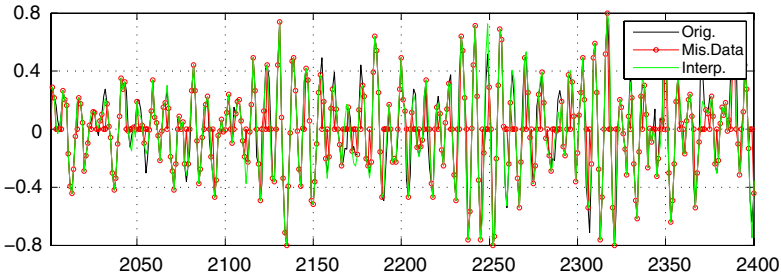

Fig. 5 1D signal interpolation under missing data (green) with $N = 20$, $n = 6$, and $\eta = 0.3$

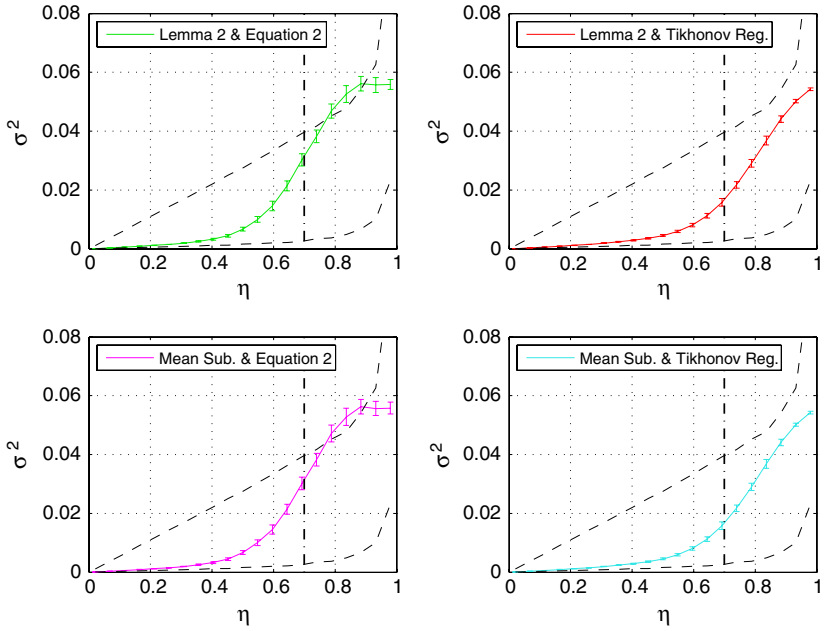

Fig. 6 Error variance for the proposed method and extensions, for $\eta \in [0.02, 0.98]$. Upper and lower bounds and validity point are depicted in black, for $N = 20$ and $n = 6$. The error bars show the $\pm$ one standard deviation across the 20 runs for each η. For the regularized solutions $\alpha = 0.6\,\eta^2$

Several conclusions can be drawn from Fig. 6, given the bounds computed from the estimator introduced in Lemma 1 and the validity interval obtained using (4). The bounds deduced for the interpolation under missing data are observed (in black) and clearly indicate the interval of possible results of the error's variance. The advantages on the use of the Tikhonov regularization are obvious: for values of η beyond the validity point, the regulated solution keeps the data integrity, making it possible to meet the proposed bounds, even when in presence of severe lack of data. Despite of the poor performance on the estimation of covariance for the mean substitution method, as in Fig. 4, the values of the error variance for the several options on the interpolation method are similar.

It is important to remark that the covariance is estimated based on the available samples. Consequently, a degradation on the bounds is verified as these are computed from the eigenvalues of the estimate of the covariance.

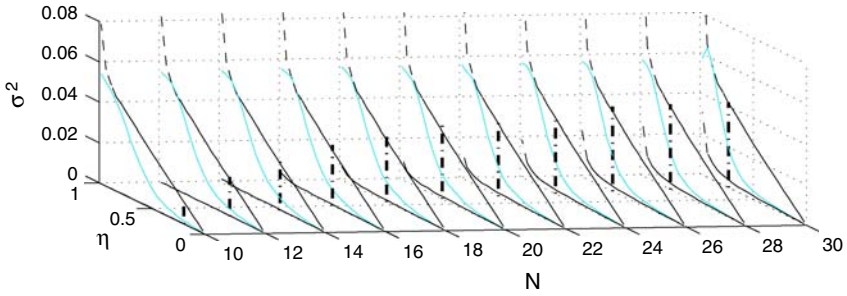

Fig. 7 Signal interpolation error variance with the proposed PCA based method for constant number of components ($n = 6$), for $\eta \in [0.02, 0.98]$ and mosaic length $N = 10, \ldots, 30$

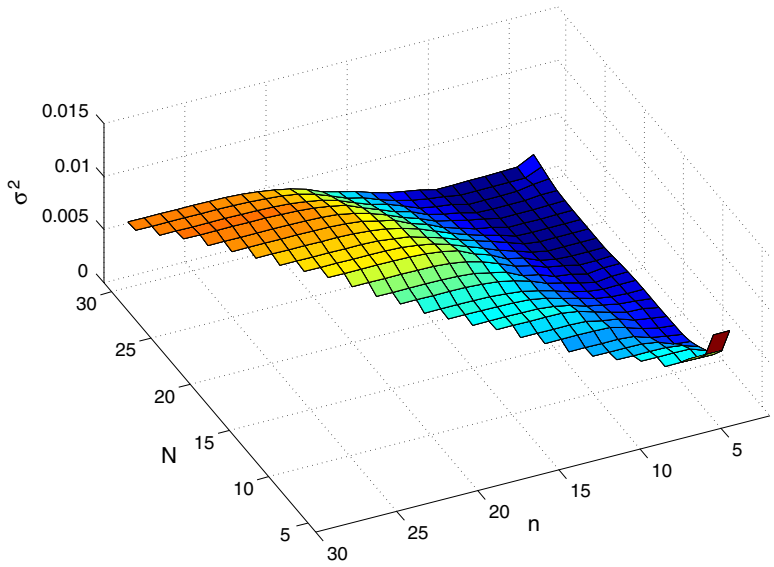

Fig. 8 Signal interpolation error variance with the proposed PCA interpolation method with constant ratio of missing samples $\eta \sim 0.3$, for a selected number of principal components $n = 4, \ldots, N$ and mosaic length $N = 5, \ldots, 30$

The performance of the proposed methodology depends on three parameters: the ratio of missing samples η, the mosaic length N, and the number of principal components n. To study the impact of these parameters on the overall interpolation performance, the results of a series of Monte Carlo experiments with the same signal are summarized next. In Fig. 7, the value for n is fixed, making it possible to show the influence of the mosaic length, and the ratio of missing samples on the interpolation results. On the other hand, in Fig. 8 the impact of the length of the mosaic and the selected number of components is depicted, for a fixed ratio of missing data.

Some conclusions can be drawn from the results obtained: (i) a smooth performance degradation was found in all the large combinations of parameters considered; (ii) in general the bounds are tight and are verified for every N; (iii) for constant missing data rates, the performance depends on the number of components used, where the selection of fewer components leads to an inaccurate interpolation, due to the disregard of significant components, while

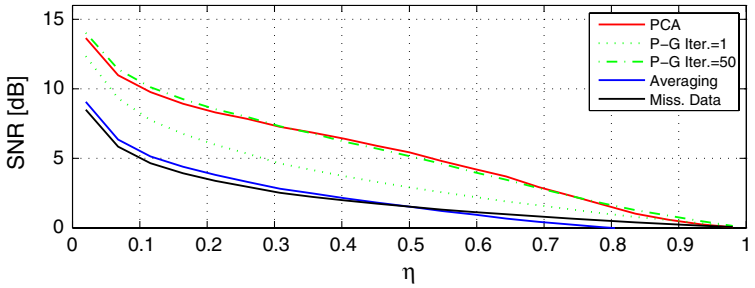

Fig. 9 1D Signal interpolation $SNR(\mathbf{r}, \tilde{\mathbf{r}})$ for $\eta \in [0.02, 0.98]$ with the PCA interpolation method (in red), the methods introduced in Sect. 5.1 and with no interpolation applied (in black)

selecting more components (thus including noisy components) increases the error variance, and (iv) longer mosaics are preferable as a better performance is achieved (with the number of components chosen kept constant), still it becomes more computationally demanding. The points (iii) and (iv) of this list corresponds to a performance compromise that should be tackled according to the type of application.

A comparison on the overall performance is depicted in Fig. 9 for the proposed method of interpolation and for the alternative methods introduced in Sect. 5.1. The goal is to assess the differences in the $SNR(\mathbf{r}, \tilde{\mathbf{r}})$ obtained for each algorithm. The P-G algorithm is an iterative method, which is compared to the non-iterative method proposed. Thus, for a fair assessment between methods, two runs differing in the number of iterations of the P-G algorithm are conducted. In order to implement the P-G algorithm it is necessary to set the bandwidth of the filter. The value is selected such that the main harmonics of the signal are present on the corresponding bandwidth thus leading to more advantageous results. For the signal in Fig. 2 the bandwidth was set to $B = 1300$, with results in green with line styles dashed and dot dashed, for the first and the 50th iteration results, respectively. Regarding the PCA interpolation method, all extensions are employed with the regularization parameter set to $\alpha = 0.6\eta^2$. For the averaging method the window size is $H = 7$ and is depicted in blue.

A graceful degradation on the $SNR(\mathbf{r}, \tilde{\mathbf{r}})$ for all methods was observed for increasing ratios of missing samples, as depicted in Fig. 9. As expected, the averaging method (in green) is always outperformed by the PCA interpolation and by the P-G methods, even when only one iteration of the P-G method (in pink) is considered. It is possible to conclude on the benefits of the proposed method, central to this paper, and the corresponding extensions where an improvement in excess of $10dB$ compared to the missing data signal can be achieved. Interestingly enough, equivalent results were obtained only after 50 iterations of the P-G algorithm.

5.3 Results for 2D signal

A second example of application is presented throughout this section. A classic 8 bit black and white image with a size of $512 * 512$ pixels, which is depicted on the left panel of Fig. 10, will be used. Also in the same figure, on the center panel, is an example of that image with missing data ($\eta = 0.3$) and the corresponding interpolated image on the right, where the methods detailed in Sect. 4 were used.

The results for the PCA interpolator and for the alternative methods are shown in Fig. 11. On the top row of Fig. 11, a series of images with several ratios of missing samples in the interval $\eta \in [0.2, \ldots, 0.95]$ are depicted. The second row illustrates the results obtained

Fig. 10 2D Signal interpolation with the PCA interpolation method, where $N = H * H = 7 * 7 = 49$, $n = 7$ and $\alpha = 0.006\,\eta^2$

Fig. 11 Results of 2D signal interpolation when alternative methods are applied, see 5.1 and the references therein for details

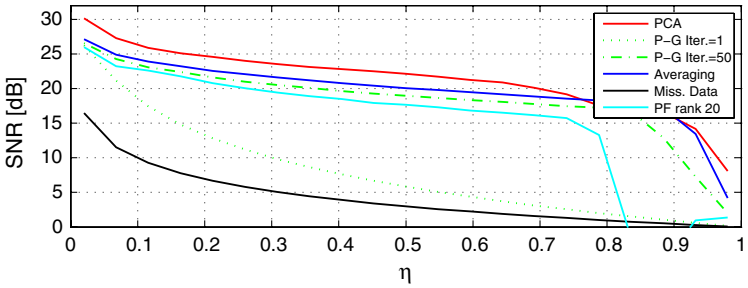

Fig. 12 2D Signal interpolation $SNR(\mathbf{r}, \tilde{\mathbf{r}})$ for $\eta \in [0.02, 0.98]$, with the PCA interpolation method (red), the methods introduced in 5.1, and for the original image with missing data (in black)

with the PCA interpolation method. Next, the results for the averaging method with square mosaics $N = H * H = 7 * 7 = 49$ are presented. In the fourth and fifth rows, results for the P-G algorithm with a filter bandwidth $B = 50$ for Iter = 1 and Iter = 50, respectively, are depicted. The last two rows are the reconstruction achieved with the Power Factorization method, for a rank 1 and a rank 50 decomposition, respectively, using the software available from the authors of Buchanan and Fitzgibbon (2005). From a careful inspection, it is immediate to conclude that all methods performed worst than the one central to this work. The advantage is even more interesting for high levels of missing data, i.e. $\eta > 0.5$.

To assess the performance of several classical and modern state of the art methods available, a comparison similar to the one presented for the 1D signal, i.e. resorting to the $SNR(\mathbf{r}, \tilde{\mathbf{r}})$, is carried out for the image in Fig. 10, with the results shown in Fig. 12. The parameters were kept as the ones established in Figs. 10 and 11. All methods presented a graceful degradation with the augment of missing data. For values above 0.75 the Power Factorization method fails (in cyan), next the P-G method fails, and finally, only for an excess of 0.9 of missing data, the averaging and the PCA based reconstruction method fail. In the case of the Power Factorization method, marginal performance increases can be obtained for higher rank approximation, however the performance collapses for smaller amounts of missing data.

The PCA reconstruction algorithm is sub-optimal, mainly due to the ambiguity on the choice of the mosaic dimension. To study the impact of this parameter on the overall performance of the proposed method, a study was carried out for a constant value of $\eta = 0.3$, for square mosaics $H \in \{3, 4, 5, 6, 7, 8, 9, 10, 11\}$. The results are presented in Fig. 13, with the number of components $n = 3 \ldots N$. Note that for a small percentage of components (20 % typical), the best performance is achieved, as can be seen in the plateau on the figure. As the computational requirements to obtain the PCA decomposition increases with the size of the mosaic, but very similar levels of performance are obtained, a moderate size for the mosaic size is advisable to be chosen (4 to 6). For larger percentages of components used, the performance of the the method degrades, due to the lack of information in the images to accurately compute all the components required.

From the results of the proposed method for the 2D signal in Fig. 12 it can be concluded that the PCA interpolation algorithm's performance is coherent and seems to be independent from the signal and its dimensionality. For the 2D signal, the PCA interpolator clearly stands as the method with the best results. For large values of η, due to severe lack of information, a pronounced degradation in the SNR is verified. The advantages of the proposed method are once again evident.

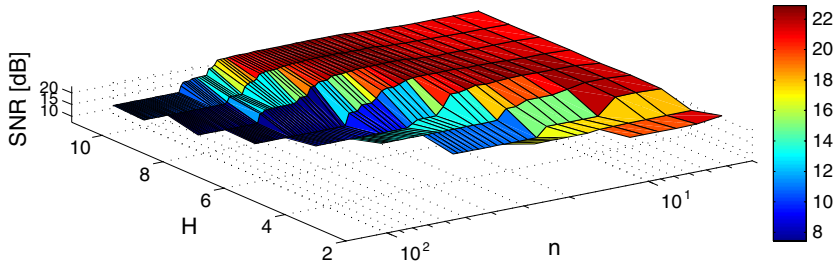

Fig. 13 Study on the proposed reconstruction method, based on the $SNR(\mathbf{r}, \tilde{\mathbf{r}})$, for a fixed missing data scenario of $\eta = 0.3$, and variable number mosaic size H and number of components $n < H * H$

6 Conclusions and future work

A new methodology to interpolate and regularize sampled signals with missing data is presented, supported on estimates from two efficient sub-optimal estimators for the mean and covariance of the underlying signals. Three refinements to the basic method in Oliveira (2006) are included with positive impact on the overall performance: (i) mean substitution, (ii) dynamic principal components selection and, (iii) Tikhonov regularization. These extensions naturally increased the numerical robustness of the interpolation method and removed the original limitations on the interval of validity, thus paving the way to the application of the present methods to a number of real problems in the interpolation of multidimensional signals. Tight upper and lower bounds were presented and validated through a series of tests, with improved performance when compared with local averaging, the Papoulis–Gerchberg, and the Power Factorization methods. No bandlimited nor Gaussian noise assumptions are required for the signals and noise present, respectively. Sensitivity studies on a series of parameters in the estimators revealed a graceful degradation on the interpolation performance. Ultimately, the application of the proposed methodology to data obtained in a series of surveying missions at sea, with unmanned underwater vehicles, is expected to be the key enabling tool to tackle terrain based navigation problems with feature based techniques (Oliveira 2007).

Acknowledgements Work supported by the Portuguese FCT POSI Programme under Framework QCA III and in the scope of project PDCT/MAR/55609/2004-RUMOS of the FCT.

References

Benedetto, J., & Ferreira, P. (2000). *Modern sampling theory: mathematics and applications,* Birkhäuser.

Blanz, V., & Vetter, T. (2002). *Reconstructing the complete 3D shape of faces from partial information*, it+ti Oldenburg, Verlag.

Buchanan, A., & Fitzgibbon, A. W. (2005). Damped newton algorithms for matrix factorization with missing data. *IEEE Computer Vision and Pattern Recognition Conference, 2*, 316–322.

Choi, H., & Munson, D. (1998). Analyis and design of minimax-optimal interpolators. *IEEE Transactions on Signal Processing, 46*(6), 1571–1579.

Gerchberg, R. (1974). Super-resolution through error energy reduction. *Optica Acta, 21*(9), 709–720.

Hartley, R. (2003). PowerFactorization: an approach to affine reconstruction with missing and uncertain data. In *Australia-Japan advanced workshop on computer vision*, Adelaide, Australia.

Jolliffe, I. (2002). *Principal component analysis*. Springer-Verlag.

Kailath, T., Sayed, A., & Hassibi, B. (2000). *Linear estimation*. Prentice Hall Information and System Sciences Series.

Kay, S. (1993). *Fundamentals of statistical signal processing – Vol. I Estimation theory*. Prentice-Hall Signal Processing Series.
Marvasti, F. (2001). *Nonuniform sampling theory and practice series: information technology: transmission, processing and storage*. Springer-Verlag.
Mertins, A. (1999). *Signal analysis: wavelets, filter banks, time-frequency transforms and applications*. John Wiley & Sons.
Oliveira, P. (2006). Interpolation of signals with missing data using PCA. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing—ICASSP06*, Toulouse, France.
Oliveira, P. (2007). MMAE terrain reference navigation for underwater vehicles using PCA. Accepted for publication on a special number of the *International Journal of Control, in Navigation, Guidance and Control of Uninhabited Underwater Vehicles, 80*(7), 1008–1017.
Papoulis, A. (1973). *A new method in image restoration*. JSTAC, Paper VI-3.
Papoulis, A. (1975). A new algorithm in spectral analysis and band-limited extrapolation. *IEEE Transactions on Circuits and Systems 19*, 735–742.
Pascoal, A., Oliveira, P., Silvestre, C., Bjerrum, A., Ishoy, A., Pignon, J.-P., et al. (1997). MARIUS: An autonomous underwater vehicle for coastal oceanography. *IEEE Robotics & Automation Magazine, 4*(4), 46–59.
Roweis, S. (1998). EM algorithms for PCA and SPCA. In *Advances in neural information processing systems*, Vol. 10, pp. 626–632.
Shum, H-Y., Ikeuchi, K., & Reddy, R. (1995). Principal component analysis with missing data and its application to polyhedral object modeling. *IEEE Transactions Pattern Analysis and Machine Intelligence, 17*(9), 854–867.
Tikhonov, A., Goncharsky, A., Stepanov, V., & YagolaKluwer, A. (1990). *Numerical Methods for the solution of Ill-posed problems* Academic Publishers.
Unser, M. (2000). Sampling–50 Years After Shannon. *Proceedings of the IEEE, 88*(4), 569–587.
Yen, J. (1956). On nonuniform sampling of bandlimited signals. *IRE Transactions on Circuit Theory, CT-3*.

Author Biographies

P. Oliveira completed the PhD in 2002 from the Instituto Superior Tecnico, Lisbon, Portugal. He is an Assistant Professor of the Department of Electrical Engineering and Computer Science of the Instituto Superior Técnico, Lisbon, Portugal and researcher in the Institute for Systems and Robotics, Lisbon, Portugal. The areas of scientific activity are Robotics and Autonomous Vehicles with special focus on the fields of Sensor Fusion, Navigation, Positioning, and Signal Processing. He participated in more than 10 Portuguese and European Research projects, in the last 15 years.

L. Gomes completed recently the MSc studies in Electrical Engineering and Computer Science from the Instituto Superior Tecnico, Lisbon, Portugal. The areas of scientific activity are Signal Processing for Robotics and Autonomous Vehicles applications.