INTERPOLATION OF SIGNALS WITH MISSING DATA USING PCA

P. Oliveira

Instituto Superior Técnico and Instituto de Sistemas e Robótica
Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal
pjcro@isr.ist.utl.pt

ABSTRACT

A non-iterative methodology for the interpolation of sampled signals with missing data resorting to Principal Component Analysis is introduced. Based on unbiased estimators for the mean and covariance of signals, corrupted by zero-mean noise, the Principal Component Analysis is performed and the signal is interpolated given the optimal solution of a weighted least squares minimization problem. Upper and lower bounds for the mean square interpolation error are also provided in the interval of validity of the method. A preliminary performance assessment, with 1-D and 2-D signals, is included based on the results of a series of Monte Carlo experiments.

1. INTRODUCTION

The problem of interpolation of sampled signals with missing data is central in a series of engineering problems. Autonomous robotic surveying [7], underwater positioning, remote sensing, digital communications (subject to bursts of destructive interferences), and computer vision (when 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 long time in solving interpolation problems, see [1, 3, 8, 13] and the references therein for an in-depth repository of available techniques. Iterative methods such as Projection Onto Convex Sets (POCS) and the Expectation/Maximization (EM) algorithm [10] are the most commonly used. The iterative characteristics, the domain of application (low-frequency or bandlimited signals) and the low convergence rates of these methods preclude their use on a number of real time applications.

Motivated by a terrain based navigation problem for underwater autonomous robotic activities [7, 9], this paper proposes a new methodology that departs from the aforementioned approaches. It introduces a non-iterative methodology for the interpolation of sampled signals with missing data based on Principal Component Analysis (PCA). Unbiased estimators for the mean and covariance of signals corrupted by zero-mean noise allow the PCA computation. The signal interpolation is tackled resorting to the optimal solution of a weighted least squares minimization problem. Moreover, upper and lower bounds for the mean square interpolation error and the interval of validity of the proposed method are provided.

PCA has already been used in interpolation problems with sampled signals with incomplete data. In [11], PCA is applied to sparse data from segmented images (not directly on the complete signal, as in the present work). Also in [2], 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: section 2 introduces unbiased estimators for the mean and covariance of discrete time signals with missing data and the PCA computation. Section 3 describes an optimal solution for the interpolation of signals, corrupted by zero-mean noise. The interval of validity of the proposed methodology and lower and upper bounds for the interpolation error variance are deduced, exploiting the PCA properties. Results from a series of Monte Carlo experiments with 1-D and 2-D signals are summarized in section 4, to allow a preliminary performance assessment of the proposed method. Finally, some conclusions are drawn and future work is unveiled in section 5.

2. PCA FOR SIGNALS WITH MISSING DATA

PCA 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 [4] 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 squares 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 signal processing applications such as data compression, image and voice processing, data mining, exploratory data analysis, pattern recognition, and time series prediction [4].

*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.

2.1. Mean and Covariance Estimators

Consider a signal $\mathbf{x} \in l_2$, i.e. with finite energy, from a real-valued stochastic process corrupted by zero mean noise, and an indicator index $\mathbf{i}$, represented as column vectors of length N. The underlying process can result from a non-homogeneous spatial survey, due to physical or kinematic constrains, or associated with the reception of a signal in a communications channel corrupted by bursts of noise that destroy completely the information contained in some samples. The index $\mathbf{i}(j), j = 1, \ldots, N$ is set to 1 if the j^{th} component of signal $\mathbf{x}$ is available and zero otherwise. In the latter, the component $\mathbf{x}(j)$ is set to zero, without loss of generality. Auxiliary results on unbiased and efficient estimators for the mean and covariance of signals with missing data, will now be introduced.

Lemma 2.1 *Given a set of M signals $\mathbf{x}_i$, with associated indexes $\mathbf{i}_i$, the auxiliary vector of counters $\mathbf{c} = \sum_{i=1}^{M} \mathbf{i}_i$, and $\mathbf{C} = \sum_{i=1}^{M} \mathbf{i}_i \mathbf{i}_i^T$:*
i) the estimator for the j^{th} component of the ensemble mean

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

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

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

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

The proof resorts to basic statistical signal processing theory that can be found for instance in [5].

2.2. Principal Component Analysis

PCA can now be computed, 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

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

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. $\sum_n \lambda_n / \sum_N \lambda_N$.

Departing from the perfect interpolation setup [12], the matrix $\tilde{\mathbf{U}} = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_n]$ with dimensions $R^{N \times n}$ will be used as the approximate PCA associated transformation. 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 squared 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, even under missing data; 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 allowing the interpolation of 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 data on the 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.*

Noise with null mean is assumed to be corrupting the underlying signal, departing from the gaussian noise assumptions in [2, 10]. To solve the interpolation problem at hand, 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 $\mathbf{L}_i \in R^{N \times N}$ is a diagonal matrix, filled with the indicator index $\mathbf{i}_i$. The interpolation operation can be formulated as finding $\tilde{\mathbf{r}}_i$ such that minimizes the weighted l_2 norm of the error, i.e.

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

Using the approximated PCA projection $\mathbf{r}_i = \tilde{\mathbf{U}}(\mathbf{v}_i + \mathbf{m}_x)$, the minimization can now be written as

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

that has the solution

$$\tilde{\mathbf{v}}_i = (\tilde{\mathbf{U}}^T\mathbf{L}_i\mathbf{W}\mathbf{L}_i\tilde{\mathbf{U}})^{-1}\tilde{\mathbf{U}}^T\mathbf{L}_i\mathbf{W}^T(\mathbf{x}_i - \mathbf{L}_i\mathbf{m}_x), \tag{2}$$

where the relations $\mathbf{L}\mathbf{L}^T = \mathbf{L}$ and $\mathbf{L}^T = \mathbf{L}$ were used. Resorting to optimal stochastic minimization techniques [6] the knowledge of the stochastic process allows the optimal choice of $\mathbf{W} = \mathbf{R}_{xx}^{-1}$. According to 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 diagram 1.

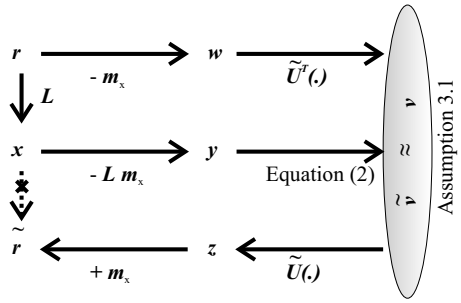

Fig. 1. Diagram describing the interpolation of sampling signals with missing data.

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 $n << N$. Moreover, this result can be interpreted as a generalization of the classical Yen interpolator [13] and of the minimax-optimal interpolators [3].

The minimization is well posed in the case where the expected number of samples available are greater than the selected number of principal components, also according with assumption 3.1, $N(1-\eta) > n$, where η is the percentage of missing samples in the signal. This leads to the validity interval for the proposed method

$$0 \leq \eta < \frac{N-n}{N}. \tag{3}$$

In the sequel, assume that n components are used from the total of N available. Lower and upper bounds on the variance of the interpolation error per sample σ^2 can be found given the PCA stochastic approximation properties, i.e,

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

leading in the missing data case to

$$\frac{\sum_{i=n+1}^{N} \lambda_i}{(N-1)(1-\eta)} \leq \sigma^2 \leq \frac{\sum_{i=1}^{N} \lambda_i}{(N-1)(1-\eta)}. \tag{4}$$

4. RESULTS FOR PERFORMANCE ASSESSMENT

In this section the results from a series of Monte Carlo experiments (20 for each parameter combination), using the estimators proposed in section 2 and the interpolation method introduced in section 3, applied to 1D and 2D signals, are presented. The bounds deduced for the interpolation under missing data are checked in the interval of validity of the method.

In Fig. 2, a non-bandlimited signal of length 128 samples considered with 18 samples missing, i.e. $\eta \sim 0.2$. After the selection of $M = 128 - N = 117$ mosaics, the ensemble mean and covariance are computed according with lemma

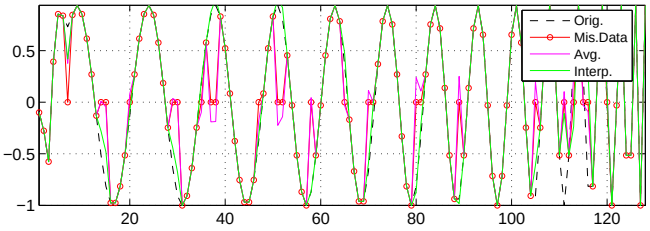

Fig. 2. 1D Signal interpolation under missing data (green) with $N = 7$, $n = 3$, and $\eta = 0.2$. In pink the results of local averaging (window of length N).

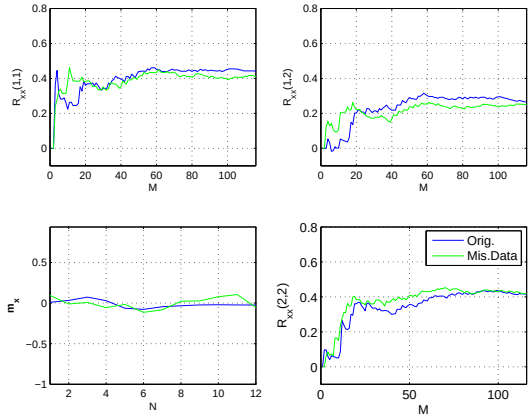

Fig. 3. Mean and covariance estimates (partial) from lemma 2.1.

2.1. The evolution of the mean and covariance estimates are depicted in Fig. 3. Given the PCA as computed in section 2, the missing data is interpolated based on the solution to the proposed l_2 minimization problem.

A performance study for a variation of the percentage of lost samples in the interval $\eta = [0.01, 0.73]$ is presented in Fig. 4, both for the proposed method and for a local averaging method, for $N = 11$. The bounds and the validity interval are observed and a graceful degradation is obtained. Clearly, the upper bound is very conservative.

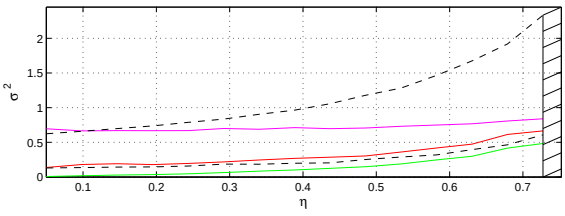

Fig. 4. Error variance with the proposed method for the interpolated samples (red), for all samples using PCA (green), and local averaging (pink). Upper and lower bounds (from 4) and validity barrier (from 3) in black, for $N = 11$ and $n = 3$.

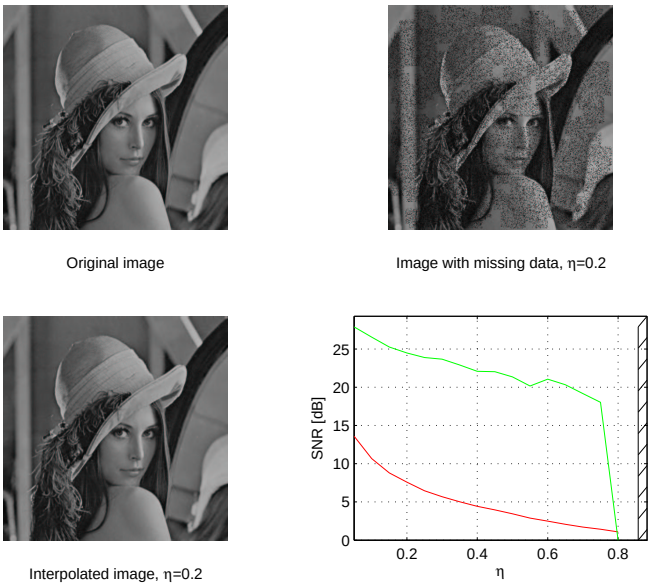

Fig. 5. 2D Signal interpolation under missing data, with $N = 7 * 7 = 49$, $n = 7$, and $\eta = 0.2$.

In Fig. 5, an 8 bit black and white image is considered. In the upper left and right panels of the figure, the original and an example of that image with missing data are depicted, respectively. In the lower left of the figure, the interpolated image obtained is presented. In the lower right, the evolution of the signal to noise ratio in the presence of missing data, over the interval $\eta = [0.05, 0.8]$, computed as

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

for the original image $\mathbf{I}$ and for the interpolated image $\tilde{\mathbf{I}}$, i.e. $SNR(\mathbf{I}, \tilde{\mathbf{I}})$, is presented, with a graceful degradation for $\eta \leq 0.7$. For larger values of η, Assumption 3.1 does not hold anymore, due to severe loss of information on the signal at hand. However, it is important to remark that an improvement in the excess of $15\ dB$ is achieved in most cases on the interval of validity of the method.

5. CONCLUSIONS AND FUTURE WORK

A new methodology to interpolate sampled signals with missing data is presented, supported on estimates from two efficient estimators for the mean and covariance of the underlying signals. Upper and lower bounds for the problem are presented and validated through a series of tests, with improved performance when compared with a local averaging method.

In the near future an in-depth benchmark with other methods, resorting to a collection of representative signals, will be performed and sensitivity studies on a series of parameters in the estimators, PCA, and in the interpolation method will be carried out. The extension of the proposed method to multidimensional signals is obvious but it should be elucidated the percentage of missing data acceptable to obtain interpolated signals relevant to the underlying problems. 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 [9].

6. REFERENCES

[1] J. Benedetto and P. Ferreira, *Modern Sampling Theory: Mathematics and Applications*, Birkhuser, 2000.

[2] V. Blanz, and T. Vetter, *Reconstructing the Complete 3D Shape of Faces from Partial Information*, it+ti Oldenburg, Verlag, 2002.

[3] H. Choi, and D. Munson, *Analyis and Design of Minimax-Optimal Interpolators*, IEEE Trans. on Signal Processing, vol.46, n.6, 1998.

[4] I. Jolliffe, *Principal Component Analysis*, Springer-Verlag, 2002.

[5] S. Kay, *Fundamentals of Statistical Signal Processing*, Prentice-Hall, 1993.

[6] T. Kaylath, A. Sayed, and B. Hassibi, *Linear Estimation*, Prentice-Hall, 2000.

[7] A. Pascoal, P. Oliveira, C. Silvestre, A. Bjerrum, A. Ishoy, J-P. Pignon, G. Ayela, and C. Petzelt, *MARIUS: An Autonomous Underwater Vehicle for Coastal Oceanography*, IEEE Robotics & Automation Magazine, 1997.

[8] F. Marvasti, *Nonuniform Sampling Theory and Practice Series: Information Technology: Transmission, Processing and Storage*, Springer-Verlag, 2001.

[9] P. Oliveira, *MMAE Terrain Reference Navigation for Underwater Vehicles using Eigen Analysis*, Proceedings of the 44th IEEE CDC/ECC 2005, Seville, Spain, December 2005.

[10] S. Roweis, *EM Algorithms for PCA and SPCA*, Proceedings of the Conference on Advances in Neural Information Processing Systems, 1997.

[11] H-Y. Shum, K. Ikeuchi, and R. Reddy, *Principal Component Analysis with Missing Data and Its Application to Polyhedral Object Modeling*, IEEE Trans. Pattern Analyis and Machine Intelligence, vol.17, n.9, 1995.

[12] M. Unser, *Sampling - 50 Years After Shannon*, Proceedings of the IEEE, vol.88, n.4, 2000.

[13] J. Yen, *On nonuniform sampling of bandlimited signals*, IRE Transactions on Circuit Theory, vol. CT-3, 1956.