In: Proc. EUSIPCO-2013, XXI European Signal Processing Conference, Marrakech, Morocco, Sept. 9-13, 2013.

LOW-RANK SIGNAL APPROXIMATIONS WITH REDUCED ERROR DISPERSION

Vicente Zarzoso, Marianna Meo and Olivier Meste

Laboratoire I3S, Université Nice Sophia Antipolis, CNRS
Les Algorithmes, Euclide-B, 2000 route des Lucioles, BP 121, 06903 Sophia Antipolis, France
{zarzoso, meo, meste}@i3s.unice.fr

ABSTRACT

Low-rank representations of multivariate signals are useful in a wide variety of applications such as data compression, feature extraction and noise filtering. Several matrix decomposition techniques like principal component analysis and independent component analysis have been proposed so far for reduced-rank signal representation. However, these methods have no effect on the error dispersion across observations, which may lead to poor representation of some input variables. To render a more uniform description of the observed data, this work puts forth a novel technique for reduced error dispersion (RED) based on a p-norm minimization criterion, with $p > 1$. The RED criterion is minimized by an iterative algorithm alternating between a gradient descent update and a least squares (LS) step via singular value decomposition. Links with existing weighted LS approaches are also established. A simulation study demonstrates the satisfactory convergence of the proposed algorithm and its ability to approximate the observed data with improved reconstruction error uniformity at a negligible impact on the average error.

Index Terms— Matrix approximations, principal component analysis (PCA), reduced error dispersion (RED), weighted least squares (WLS).

1. INTRODUCTION

Approximating a matrix by another with lower rank is a fundamental problem in matrix algebra and signal processing, finding application in a variety of fields such as data compression, feature extraction and noise removal. A key aspect of most real-world phenomena is their multivariate nature, and modern measurement systems are indeed able to acquire an increasingly high number of data variables. Typical examples are biomedical signal acquisition schemes such as the ECG and the EEG that record the physiological phenomena of interest (electrical activity from the heart and the brain, respectively) using simultaneous multiple leads. Focusing on a single variable or lead separately while neglecting the others and their relationships often leads to wrong model interpretation and poor signal reconstruction. This calls for the development of multivariate decomposition techniques that, through suitable constraints, are able to render a faithful representation of the input signal with reduced dimensionality while preserving its most meaningful features. In ECG data compression, for instance, one may require not only the average reconstruction error to be small, but also the reconstruction quality to be evenly distributed across leads.

Principal component analysis (PCA) is arguably the most popular low-rank signal approximation technique [1]. It decomposes the data matrix as a sum of rank-1 components retaining most of the input variance and minimizing the average reconstruction error in the least square (LS) sense. Despite its usefulness in data analysis, PCA has no control on signal dispersion among input variables, so that reconstruction accuracy can be hampered by these scattering effects. This drawback is shared by recent variants. Robustness to outliers can be improved through alternative error measures such as the 1-norm and correntropy [2, 3]. Also based on the 1-norm and related constraints, sparse PCA [4] tries to simplify the interpretation of principal components by reducing the number of non-zero loadings. Another line of work has kept exploring the benefits of the classical 2-norm; for instance, minimizing the uniform error, defined as the maximum error 2-norm over the data samples, leads to cost-effective visual tracking algorithms with remarkable performance [5]. Yet, as for classical PCA, none of these variants is designed to guarantee a reconstruction error more evenly spread over the input variables. Similar limitations are encountered by independent component analysis (ICA) [6, 7], since independence constraints do not take into account the reconstruction error distribution across inputs.

To bridge this gap, the present investigation puts forward a novel criterion for achieving reduced error dispersion (RED) of low-rank matrix approximations. By means of a p-norm cost with $p > 1$, this criterion aims not only at minimizing the mean reconstruction error as in PCA, but also its dispersion among input variables. This cost function can be efficiently optimized by a gradient-descent iteration followed by orthog-

This work is partly funded by the French National Research Agency under contract ANR-2010-JCJC-0303-01 "PERSIST". M. Meo is funded by a doctoral grant from the French Ministry of Higher Education and Research, and also partly supported by the DreamIT Foundation in partnership with the University of Nice Sophia Antipolis.

onal projection on the space of low-rank matrices via singular value decomposition (SVD). Interestingly, the RED approach can be interpreted as a special instance of the weighted PCA technique of [8, 9, 10] with non-constant weights, which are effectively handled by the proposed optimization algorithm. After presenting the new method, its performance is assessed by numerical experiments, showing its good convergence properties and its ability to fit the input matrix with improved reconstruction error uniformity.

2. LOW-RANK MATRIX APPROXIMATIONS

2.1. Problem formulation

Let us consider a multivariate signal represented by L variables observed over N samples, denoted x_{ij}, $i = 1, 2, \ldots, L$, $j = 1, 2, \ldots, N$, where typically $N > L$. For simplicity in the mathematical derivations that follow, the observed data are assumed to be real valued, although extensions to the complex case are possible with minor modifications. The observed samples are often arranged in the form of a data matrix $\mathbf{X} \in \mathbb{R}^{L\times N}$ with entries $[\mathbf{X}]_{ij} \stackrel{\text{def}}{=} x_{ij}$. Low-rank representations aim at fitting the data matrix by the rank-R bilinear model

$$\hat{\mathbf{X}} = \mathbf{H}\mathbf{S}^T \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{L\times R}$ and $\mathbf{S} \in \mathbb{R}^{N\times R}$, with $R < L$; symbol $(\cdot)^T$ denotes the transpose operator. Clearly, this model is invariant to post-multiplication of its factors by an non-singular matrix $\mathbf{A} \in \mathbb{R}^{R\times R}$ and its inverse transpose, since $(\mathbf{HA})(\mathbf{SA}^{-T})^T = \mathbf{HS}^T$, so that additional constraints are necessary to reduce this ambiguity. For instance, PCA assumes model factors $\mathbf{H}$ and $\mathbf{S}$ with orthogonal columns, whereas ICA imposes statistical independence between the R random variables whose realizations are defined along the columns of $\mathbf{S}$.

If the interest lies in the reconstruction accuracy, model (1) is estimated by minimizing a given function of the residuals subject to the assumed constraints. Such a function is often related to the mean square error (MSE) in the estimation of the input variables, defined as

$$\varepsilon_i \stackrel{\text{def}}{=} \frac{1}{N}\sum_{j=1}^{N}(x_{ij} - \hat{x}_{ij})^2 \qquad i = 1, 2, \ldots, L. \tag{2}$$

The reconstruction errors for all input variables can be stored in a vector $\boldsymbol{\varepsilon} \stackrel{\text{def}}{=} [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L]^T$.

2.2. Principal component analysis

The low-rank estimate provided by PCA consists in the minimization of the residual MSE averaged over the input variables:

$$\Psi_{\text{PCA}} = \overline{\varepsilon} = \frac{1}{L}\sum_{i=1}^{L}\varepsilon_i = \frac{1}{L}\|\boldsymbol{\varepsilon}\|_1 \tag{3}$$

where the second equality is due to the positivity of ε_i, implying $\varepsilon_i = |\varepsilon_i|$, $i = 1, 2, \ldots, L$, Combining eqns. (1)–(3), the minimization of Ψ_{PCA} can be formulated as the ordinary least squares (LS) fitting problem:

$$\arg\min_{\mathbf{H},\mathbf{S}} \|\mathbf{X} - \mathbf{H}\mathbf{S}^T\|_{\text{Fro}}^2 \tag{4}$$

where $\|\cdot\|_{\text{Fro}}$ represents the Frobenius norm. The minimizers of this function with respect to $\mathbf{H}$ and $\mathbf{S}$ while keeping the other factor constant are easily deduced as:

$$\mathbf{H}_{\text{opt}} = \mathbf{X}\mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1} \qquad \mathbf{S}_{\text{opt}} = \mathbf{X}^T\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}. \tag{5}$$

Iterating between these two equations leads to an alternating LS (ALS) algorithm for the computation of the optimal factors of the PCA model. However, subject to the orthogonality constraints on the model factors, the above equations admit well-known closed-form solutions for $\mathbf{H}$ and $\mathbf{S}$ in terms of the R dominant eigenvectors of matrices $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$, respectively. An alternative equivalent solution exists in terms of the SVD of the data matrix, by retaining the singular vectors associated with the R largest singular values of $\mathbf{X}$ [1].

2.3. Reduced error dispersion

Since PCA merely focuses on the average reconstruction MSE, it lacks control over the actual distribution of the error across different variables. As a result, some input variables may be better approximated than others, which may be undesirable in some applications. This observation is further supported by the use of the 1-norm in eqn. (3), presenting certain tendency to make the error distribution sparse. To surmount this shortcoming, one could think of including in the PCA cost (3) a penalty term based on the error standard deviation, defined as

$$\sigma_\varepsilon = \sqrt{\frac{1}{L}\sum_{i=1}^{L}\varepsilon_i^2 - \overline{\varepsilon}^2} \tag{6}$$

giving rise to $\Psi_{\text{RED}}^{(2)} \stackrel{\text{def}}{=} (\overline{\varepsilon}^2 + \sigma_\varepsilon^2) = \frac{1}{L}\sum_{i=1}^{L}\varepsilon_i^2 = \|\boldsymbol{\varepsilon}\|_2^2/L$. Clearly, this combined cost takes into account both the average error and the error dispersion. The natural extension of this idea leads to the following criterion based on the p-norm of the fitting error:

$$\Psi_{\text{RED}}^{(p)} \stackrel{\text{def}}{=} \frac{1}{L}\|\boldsymbol{\varepsilon}\|_p^p = \frac{1}{L}\sum_{i=1}^{L}\varepsilon_i^p \tag{7}$$

with $p > 1$. We refer to this function as *reduced error dispersion (RED)* criterion. Note that if $p = 1$, cost (7) remains valid, but reduces to standard PCA (3) and loses explicit control on error disperson. For convenience, the shorthand notation Ψ will be used for $\Psi_{\text{RED}}^{(p)}$ in the sequel, where the value of p will generally be clear from the context.

2.4. Alternating least squares solution

If $\mathbf{X}$ is full rank, the stationary points of the RED criterion (7) with respect to $\mathbf{H}$ and $\mathbf{S}$ are given by:

$$\mathbf{H}_{\text{opt}} = \mathbf{X}\mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1} \qquad \mathbf{S}_{\text{opt}} = \mathbf{X}^T\mathbf{E}_p\mathbf{H}(\mathbf{H}^T\mathbf{E}_p\mathbf{H})^{-1} \tag{8}$$

where

$$\mathbf{E}_p \stackrel{\text{def}}{=} \text{Diag}([\varepsilon_1^{p-1}, \varepsilon_2^{p-1}, \ldots, \varepsilon_L^{p-1}]). \tag{9}$$

Since $\mathbf{E}_p$ depends on $\mathbf{H}$ and $\mathbf{S}$ through eqns. (1)–(2), the above expression for $\mathbf{S}_{\text{opt}}$ actually defines a set of $(L \times R)$ polynomial equations of degree $(2p-1)$ in the entries of $\mathbf{S}$. Except for $p = 1$ (PCA), these equations are coupled and nonlinear, which prevents the explicit expression of factor $\mathbf{S}$ as a function of $\mathbf{H}$ and $\mathbf{X}$ only. As a result, the ALS algorithm defined by (8) shows poor convergence for $p > 1$, as confirmed by our preliminary experiments, and we are left to consider alternative solutions.

2.5. Iterative optimization: gradient descent

The iterative procedure proposed in this work consists of performing unconstrained gradient descent of the cost (7) with respect to the approximation $\hat{\mathbf{X}}$. This is followed by an additional step enforcing the required rank-R model structure, as will be detailed in the following section. Differentiating (7) with respect to $\hat{x}_{ij}$ results in the gradient matrix:

$$\nabla\Psi(\hat{\mathbf{X}}) = -\frac{2p}{LN}\mathbf{E}_p(\mathbf{X} - \hat{\mathbf{X}})$$

where matrix $\mathbf{E}_p$ is defined in eqn. (9). The gradient-descent update is then given by

$$\hat{\mathbf{X}}_k^+ = \hat{\mathbf{X}}_k - \mu\nabla\Psi(\hat{\mathbf{X}}_k) \tag{10}$$

for a suitable step-size parameter μ. To choose an appropriate step-size value, we assume that criterion (7) can be perfectly cancelled and the current estimate $\hat{\mathbf{X}}_k$ is close to the optimal solution. The first-order Taylor approximation of Ψ around $\hat{\mathbf{X}}_k$ reads:

$$\Psi(\hat{\mathbf{X}}_k^+) \approx \Psi(\hat{\mathbf{X}}_k) + \text{trace}\Big(\nabla\Psi(\hat{\mathbf{X}}_k)^T(\hat{\mathbf{X}}_k^+ - \hat{\mathbf{X}}_k)\Big).$$

Plugging in update (10), we have

$$\Psi(\hat{\mathbf{X}}_k^+) \approx \Psi(\hat{\mathbf{X}}_k) - \mu\|\nabla\Psi(\hat{\mathbf{X}}_k)\|_{\text{Fro}}^2$$

which, by nulling the left-hand side of the expression, leads to

$$\mu = \Psi(\hat{\mathbf{X}}_k)/\|\nabla\Psi(\hat{\mathbf{X}}_k)\|_{\text{Fro}}^2.$$

This yields the gradient-descent update with adaptive step size:

$$\hat{\mathbf{X}}_k^+ = \hat{\mathbf{X}}_k - \frac{\Psi(\hat{\mathbf{X}}_k)\nabla\Psi(\hat{\mathbf{X}}_k)}{\|\nabla\Psi(\hat{\mathbf{X}}_k)\|_{\text{Fro}}^2}. \tag{11}$$

2.6. Enforcing the low-rank structure

The above gradient iteration is not consistent with the rank-R structure assumed for the fitted model $\hat{\mathbf{X}}$ in (1). To enforce this structure, update $\hat{\mathbf{X}}_k^+$ in (11) is projected on the space of rank-R matrices with dimensions $(L \times N)$. This can be achieved by solving the LS problem (4) using $\hat{\mathbf{X}}_k^+$ instead of the original data matrix $\mathbf{X}$. As a result, the final update, denoted $\hat{\mathbf{X}}_{k+1}$, is obtained by truncating the R dominant terms of the SVD of $\hat{\mathbf{X}}_k^+$ [11].

To recap, an iteration of the proposed algorithm consists of gradient-descent iteration (11) followed by the truncated SVD to ensure the rank-R structure. This two-step iteration is repeated until convergence. As a stopping criterion, we check if the relative distance between successive iterates or between their cost values lie below a certain threshold η, i.e.,

$$\frac{\|\hat{\mathbf{X}}_{k+1} - \hat{\mathbf{X}}_k\|_{\text{Fro}}}{\|\hat{\mathbf{X}}_k\|_{\text{Fro}}} < \eta \quad \text{or} \quad \frac{|\Psi(\hat{\mathbf{X}}_{k+1}) - \Psi(\hat{\mathbf{X}}_k)|}{\Psi(\hat{\mathbf{X}}_k)} < \eta. \tag{12}$$

After convergence, the columns of factor $\mathbf{H}$ in model (1) can be obtained as the R dominant left singular vectors of $\hat{\mathbf{X}}_{k+1}^+$ scaled by the corresponding singular values, whereas $\mathbf{S}$ is made up of the associated right singular vectors.

2.7. Links with weighted LS solutions

Weighted LS (WLS) problems aim at finding

$$\arg\min_{\mathbf{H},\mathbf{S}} \|\mathbf{W} \odot (\mathbf{X} - \mathbf{H}\mathbf{S}^T)\|_{\text{Fro}}^2 \tag{13}$$

where $\mathbf{W} \in \mathbb{R}^{L\times N}$ is a matrix of positive weights and $\odot$ stands for the Hadamard (element-wise) product [8]. Developing eqn. (13) and comparing it with eqns. (1), (2) and (7), the RED criterion can be considered as a particular WLS objective with weights

$$w_{ij}^2 = \varepsilon_i^{p-1} \tag{14}$$

for $i = 1, 2, \ldots, L$, and $j = 1, 2, \ldots, N$. According to this connection, an ALS procedure — referred to as criss-cross multiple regressions in [8] — could be employed for minimizing the RED cost, and indeed it particularizes to (8). However, whereas the weights of the original WLS criterion are set to constant values before starting the iterations, RED weights (14) depend on the fitting error for the current values of the model factors [through eqns. (1)–(2)], and thus vary with successive iterations, making the ALS iteration suboptimal. This may help explain the bad performance of the ALS algorithm observed at the end of Sec. 2.4. On the other hand, the two-step algorithm put forward in Secs. 2.5–2.6 bears strong resemblance with the majorizing function approach of [9, 10] for WLS optimization, but is actually derived from a gradient-descent iteration and handles the non-constant weights in a natural fashion. This algorithm has yielded satisfactory convergence in all our experiments, which are reported next.

3. EXPERIMENTAL RESULTS

3.1. Experimental set-up

Some numerical experiments are carried out to assess the performance of the RED criterion of Sec. 2.3 optimized by the iterative algorithm of Secs. 2.5–2.6. Results for different values of $p > 1$ are compared with those for $p = 1$, corresponding to classical PCA. More recent variants such as [2, 3, 4, 5] are not considered in this experimental study since, by design, these methods are not more likely to yield solutions with improved error dispersion than PCA, as discussed in the Introduction. At each Monte Carlo iteration, a full-rank data matrix $\mathbf{X}$ is generated as the product of two random matrices with dimensions $(L \times L)$ and $(L \times N)$ made up of zero-mean unit-variance Gaussian entries, with $N = 1000$. Arbitrary linear relationships between the rows of $\mathbf{X}$, linked to the cross-correlations of the underlying variables, are simulated in this way. The PCA solution is used as the initial point $\hat{\mathbf{X}}_0$ for the RED iterations. The threshold parameter is set to $\eta = 10^{-6}$ in stopping test (12). After convergence of the algorithm, several performance indices are computed. Besides the average error $\bar{\varepsilon}$ defining the PCA criterion (3), the standard deviation of the fitting error, σ_ε [eqn. (6)], is also computed to assess the degree of dispersion over input variables. As seen in Sec. 2.3, this index is related to the RED cost for $p = 2$. To quantify the degree of dispersion independently of the RED cost, we consider the error across inputs as a random distribution with discrete probabilities $\varepsilon_i / \sum_{i=1}^{L} \varepsilon_i$, and compute its Kullback-Leibler divergence from the discrete uniform distribution:

$$D_{\mathrm{KL}} = -\frac{1}{L} \sum_{i=1}^{L} \log \left(\frac{\epsilon_i}{\bar{\epsilon}} \right).$$

This index cancels out if and only if all reconstruction errors are identical, and is strictly positive otherwise. All performance indices are averaged over 100 independent Monte Carlo runs.

3.2. Influence of observation dimensions

The first set of experiments evaluates the influence of the signal subspace dimension L on RED's performance for several values of p, assuming rank-1 approximations ($R = 1$). The results are graphically summarized in Fig. 1. The error statistics worsen with increasing dimensions, as a model with fixed degrees of freedom is fitted to larger signal subspaces. However, error dispersion improves as p increases, especially for low data dimensions L, where reductions of up to 10 dB in error standard deviation can be observed (top plot of Fig. 1). The trends in error uniformity as described by D_{KL} are analogous (bottom plot). More remarkably, these favorable effects are obtained at a negligible impact on the average error (top plot). Although not shown here due to space limitations, the number of iterations for convergence increases slightly with p and L, but the algorithm required at most 250 iterations over the range of values considered for these parameters in this experiment.

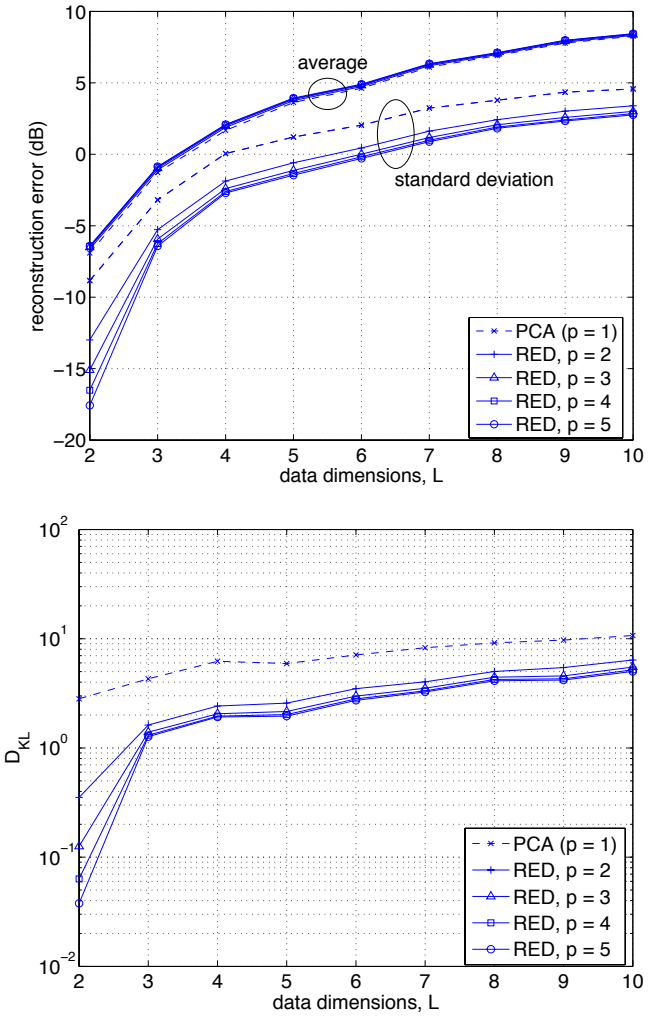

Fig. 1. RED criterion performance against observation dimensions, with $N = 1000$ samples and approximation rank $R = 1$. (Top) Reconstruction error statistics. (Bottom) Kullback-Leibler divergence from uniform error distribution.

3.3. Influence of model rank

The experiment is repeated by keeping $L = 5$ dimensions and fitting a model with increasing rank R from 1 to 4. The results, plotted in Fig. 2, confirm the reduced error dispersion obtained by RED with $p > 1$ as compared with PCA, while leaving the average error essentially unaffected (top plot). This time the error dispersion in terms of standard deviation improves with R, as the increasing degrees of freedom allow a more accurate model fitting in a signal subspace with constant dimensions. Nevertheless, only the RED criterion with sufficient p is able to enhance error uniformity as mea-

sured by the cost-independent index D_{KL} (bottom plot). The number of iterations (again not shown here for lack of space) increase with p but decrease slightly with R. For all values of p and R examined in this simulation, the algorithm required no more than 200 iterations for convergence.

4. CONCLUSIONS

The present work has put forward a novel technique for computing lower-rank representations of multivariate signals. By means of a p-norm criterion, the proposed method yields a low-rank approximation of the input data matrix minimizing not only the average MSE but also its dispersion among variables. Such features allow a more uniform representation of input observations, as well as the compensation of irregularity effects in the computed approximation. Although the criterion is linked to WLS optimization, special adjustments are required to take into account the varying nature of the weights implicitly used in the RED cost. The numerical analysis confirms the good convergence of the proposed iterative algorithm for RED criterion minimization. On synthetic data, the low-rank RED fitting is indeed characterized by a higher degree of error uniformity among input variables than PCA and is able to maintain the average error at practically the same level as the classical decomposition technique. Further work is required to confirm the compression efficacy of the RED approximation on real data and the possibility to exploit its capabilities in actual applications requiring data dimensionality reduction and feature extraction. Extensions to multi-way array (tensor) decompositions are also worth exploring.

5. REFERENCES

[1] I. T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics. Springer, New York, 2nd ed., 2002.

[2] N. Kwak, "Principal component analysis based on L1-norm maximization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1672–1680, Sept. 2008.

[3] R. He, W.-S. Hu, B.-G.and Zheng, and X.-W. Kong, "Robust principal component analysis based on maximum correntropy criterion," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1485–1494, June 2011.

[4] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[5] J. Ho, K.-C. Lee, M.-Y. Yang, and D. Kriegman, "Visual tracking using learned linear subspaces," in *Proc. CVPR-2004, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington DC, USA, June 27–July 2, 2004, vol. I, pp. 782–789.

[6] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994, Special Issue on Higher-Order Statistics.

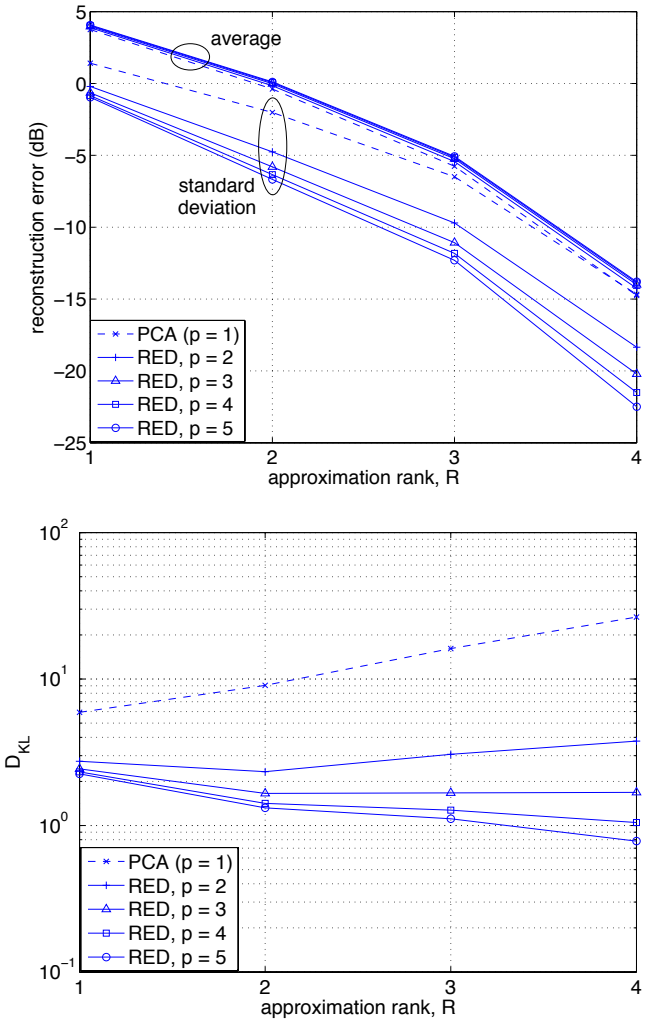

Fig. 2. RED criterion performance against approximation rank R, with $N = 1000$ samples and $L = 5$ input variables. (Top) Reconstruction error statistics. (Bottom) Kullback-Leibler divergence from uniform error distribution.

[7] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, Academic Press, Oxford, UK, 2010.

[8] K. R. Gabriel and S. Zamir, "Lower rank approximation of matrices by least squares with any choice of weights," *Technometrics*, vol. 21, no. 4, pp. 489–498, Nov. 1979.

[9] W. J. Heiser, "Convergent computation by iterative majorization: theory and applications in multidimensional data analysis," in *Recent Advances in Descriptive Multivariate Analysis*, W. J. Krzanowski, Ed., pp. 157–189. Oxford University Press, Oxford, UK, 1995.

[10] H. A. L. Kiers, "Weighted least squares fitting using ordinary least squares algorithms," *Psychometrika*, vol. 62, no. 2, pp. 251–266, June 1997.

[11] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore, MD, 3rd ed., 1996.