# DIMENSIONALITY REDUCTION VIA REGRESSION ON HYPERSPECTRAL INFRARED SOUNDING DATA

*Valero Laparra, Jesús Malo, and Gustavo Camps-Valls* *

Image Processing Laboratory (IPL), Universitat de València, Spain. http://isp.uv.es

**ABSTRACT**

This paper introduces a new method for dimensionality reduction via regression (DRR). The method generalizes Principal Component Analysis (PCA) in such a way that reduces the variance of the PCA scores. In order to do so, DRR relies on a deflationary process in which a non-linear regression reduces the redundancy between the PC scores. Unlike other nonlinear dimensionality reduction methods, DRR is easy to apply, it has out-of-sample extension, it is invertible, and the learned transformation is volume-preserving. These properties make the method useful for a wide range of applications, especially in very high dimensional data in general, and for hyperspectral image processing in particular. We illustrate the performance of the algorithm in reducing the dimensionality of IASI hyperspectral image sounding data. We compare DRR with related and invertible methods such as linear PCA and Principal Polynomial Analysis (PPA) in terms of reconstruction error, and expressive power of the extracted features to estimate atmospheric variables.

***Index Terms***— Manifold learning, nonlinear dimensionality reduction, principal component analysis, Principal Polynomial Analysis, hyperspectral sounder, MetOp, IASI

## 1. INTRODUCTION

In recent years, a plethora of nonlinear dimensionality reduction methods has been presented trying to deal with manifolds that cannot be described with linear methods, see [1] for a comprehensive review. Approaches to the problem range from local methods [2–6], kernel-based and spectral decompositions [7–9], neural networks [10–12], and projection pursuit approaches [13, 14]. Despite the advantages of nonlinear methods, the fact is that classical principal component analysis (PCA) [15] is still the most widely used dimensionality reduction technique in real applications. The main reasons are that PCA: 1) is easy to apply, 2) involves solving a fast and convex problem, 3) its performance can be simply evaluated because of its invertibility, 4) the components can be interpreted, and 5) has a straightforward out-of-sample extension. In this paper, we present a nonlinear extension of PCA that shares the above mentioned appealing properties of PCA.

The above properties are not always present in the new dimensionality reduction algorithms due to either their complex formulations, introduction of a number of non-intuitive free parameters to be tuned, high computational cost, non-invertibility of the achieved transformation, strong assumptions about the manifold characteristics (e.g. Gaussianity), and the difficulty to obtain out-of-sample predictions. More plausibly, though, the limited adoption of nonlinear methods in daily practice has to do with the lack of feature and model interpretability. Actually, interpretation of the model is tightly related to invertibility of the learned transform: invertibility allows to both characterize the transformed domain, and to evaluate the quality of the transform. On the one hand, inverting the data back to the input domain is an important feature because one can understand the input physical units therein, while analyzing the results in the transformed domain is typically more complicated (if at all possible). On the other hand, regarding evaluation, invertible transforms like PCA allow simple assessment of the reconstruction errors. We should stress here that invertibility is scarcely achieved in the manifold learning literature. For instance, spectral methods do not generally yield intuitive mappings between the original and the intrinsic curvilinear coordinates of the low dimensional manifold.

In this paper, we introduce the dimensionality reduction based on regression (DRR) technique, which is a *nonlinear generalization of PCA* and still shares its important properties. DRR is computationally inexpensive, and it is robust since it reduces to solving a series of convex problems. DRR actually implements a volume-preserving and invertible map. Moreover, applying the learned transform to new samples is also straightforward, as in PCA.

The paper is organized as follows. Section 2 introduces the main characteristics of the algorithm. We focus on a very high-dimensional problem to estimate atmospheric state vectors from IASI hyperspectral sounding data with reduced dimensionality, which is described and motivated in Section 3. Section 4 compares DRR with PCA [15] and with recent nonlinear generalizations (e.g. Principal Polynomial Analysis, PPA [16, 17]) that yield better results than Non-Linear PCA based on neural networks [10, 12]. Comparisons are made both in terms of reconstruction error and of expressive power of the extracted features. We end the paper with some concluding remarks in Section 5.

## 2. DIMENSIONALITY REDUCTION VIA REGRESSION

PCA removes the second order dependencies between the components, i.e. the PCA scores are decorrelated [15]. Equivalently, PCA can be casted as the linear transformation that ensures minimum reconstruction error when a fixed number of dimensions are neglected. However, for general non-Gaussian sources, and in particular for hyperspectral signals, the PCA scores still display significant statistical relations [18][ch. 1]. The scheme presented in this work tries to remove the information that still remains between the PCA components.

### 2.1. DRR learning scheme

We propose a deflationary scheme in which each PCA component is nonlinearly predicted from the higher-variance components. For each coefficient, we estimate a given score from the rest, and then compute its residual. Only the non-predictable information (the

*This work was partially funded by the Spanish Ministry of Economy and Competitiveness, under the LIFE-VISION project TIN2012-38102-C03-01.

residual error) is retained in each case. The first score to be estimated is the last component, the one with less variance, and we continue the procedure until the first component is reached. Once we removed the estimate from one score, we will not use this score in the following iterations. Schematically, the so-called DRR on $n$-dimensional signals can be depicted as:

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \\ \alpha_n \end{pmatrix} \xrightarrow{\mathbf{R}_1} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{n-1} \\ y_n \end{pmatrix} \xrightarrow{\mathbf{R}_2} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{n-2} \\ y_{n-1} \\ y_n \end{pmatrix} \cdots \xrightarrow{\mathbf{R}_{n-1}} \begin{pmatrix} \alpha_1 \\ y_2 \\ \vdots \\ y_{n-2} \\ y_{n-1} \\ y_n \end{pmatrix},$$

where $\alpha_i$ are the Principal Component (PC) scores, i.e. the projected data $\alpha = Vx$, being $V$ a matrix containing the eigenvectors in rows and $x$ the original centered data. In the $i$-th step, we compute the residual of the approximation,

$$y_i = \alpha_i - \hat{\alpha}_i = \alpha_i - f_i(\alpha_{i-1}, \alpha_{i-2}, ..., \alpha_1),$$

where the residual, $y_i$, is the $i$-th dimension of the DRR domain.

Note that using linear regressions in $f_i(\cdot)$ would lead to zeros in $\hat{\alpha}_i$ (the components are decorrelated) and hence $y_i = \alpha_i$, i.e. linear-DRR reduces to PCA. Therefore, DRR generalizes PCA whenever using nonlinear functions in $f_i(\cdot)$. The flexibility of these functions with regard to the linear case will reduce the variance of the residuals, and hence the reconstruction error in dimensionality reduction.

### 2.2. Selecting the family of approximation functions

In practice, the functions $f_i(\cdot) = \hat{\alpha_i}$ reduce to the training of a nonlinear regression. In our experiments, we used the kernel ridge regression (KRR) [19] to implement the predictions $f_i(\cdot)$, although any alternative regression method could be applied. KRR can be quite convenient in this scheme because it implements flexible nonlinear regression functions, and reduces to solving a unique convex matrix inversion problem. KRR combines a good performance in prediction, offers a moderate training complexity, obtains an efficient consuming time for prediction, and also offers the possibility to generate multi-output nonlinear regression and, due to its tight relation with Gaussian Processes, confidence intervals for the predictions in a natural way. Finally, KRR has been widely used in many real problems, including remote sensing applications involving high dimensional data, but even in such cases a previous feature extraction was mandatory to attain significant results [20–22].

### 2.3. Inversion and out-of-samples extension

Given the DRR transformed vector, $(\alpha_1, y_2, y_3, \ldots, y_n)^\top$, and knowing the series of models $f_i(\cdot)$, the inverse is straightforward since it reduces to undo the sequence described above. At some point of the inversion sequence, we use the previous *known* PCs to predict the considered one using the *known* function, $f_i(\cdot)$, and then we use the *known* residual, $y_i$, to correct the prediction:

$$\alpha_i = \hat{\alpha}_i + y_i = f_i(\alpha_{i-1}, \alpha_{i-2}, ..., \alpha_1) + y_i$$

Note that forward and inverse DRR transforms can be applied to new data (out-of-samples extension) since there is no restriction in KRR prediction functions $f_i(\cdot)$.

### 2.4. Relation to previous methods and computational cost

Removal of the predictions from each dimension in (what can be considered) a sequential procedure, may sound similar to recently proposed methods based on drawing a *sequence* of Principal Curves: Sequential Principal Curves Analysis (SPCA) [23] and Principal Polynomial Analysis (PPA) [16, 17]. In those methods, each curve, either non-parametrical (SPCA) or analytical (PPA), accounts for one curvilinear dimension of the data. For instance, in PPA, the $i$-th curve is used to predict the $(n-i)$-dimensional subspace orthogonal to $\alpha_i$: $(\hat{x}_{i+1}, \hat{x}_{i+2}, \hat{x}_{i+3}, ..., \hat{x}_n) = f_i(\alpha_i)$, and hence, the same behavior (curve) is applied along the previously considered dimensions, $1, 2, \ldots, (i-1)$. This is a limitation of PPA that may restrict the success of the model in cases where the manifold displays more complex structure.

On the contrary, DRR is not about drawing a *sequence* of Principal Curves, but on using nonlinear regressions to remove the dependence between PCA components. The consideration of multiple components in the DRR regressions, $\hat{\alpha}_i = f_i(\alpha_{i-1}, \alpha_{i-2}, ..., \alpha_1)$, implies that the interaction at different locations of the space may be quite different and hence richer structures may be described with DRR.

Moreover, even though each Kernel regression in DRR is more expensive than fitting polynomials in PPA, note that DRR allows trivial parallel implementations. The sequence depicted above is just a convenient way to think on the transform, but the prediction of each component is done from a subset of the original PC scores. As a result, all the predictions $f_i(\cdot)$ can be done at the same time after the initial PCA, which is not possible in PPA.

### 2.5. DRR is a volume preserving transform

A nonlinear transform preserves the volume of the input space if the determinant of its Jacobian is one for all $x$. The nature of DRR ensures that its Jacobian fulfils this property. DDR can be seen as a sequential algorithm in which only one dimension is addressed at a time. In each step of this sequence, remaining relations among the principal components are reduced by subtracting the prediction of each PCA component obtained from the components of larger variance (taken from the previous DRR stage). Hence, the (global) Jacobian of DRR is the product of the Jacobians of the elementary transforms in this sequence. The $i$-th elementary transform leaves all components but the $i$-th dimension unaltered. Therefore, the Jacobian for this transform is the identity matrix except for the $i$-th row, where below the diagonal it contains the derivatives of the $i$-th regression function with regard to each component in the previous stage. Whatever these derivatives are (whatever regression function is used), the determinant of such a simple matrix (identity with a single non-zero row below the diagonal) is always one. Therefore, the determinant of the global Jacobian, including the PCA rotation, is guaranteed to be one.

## 3. EXPERIMENTAL SETUP

In this work we will analyze the benefits of using DRR methods for the estimation of atmospheric parameters from hyperspectral infrared sounding data with a reduced dimensionality.

### 3.1. Infrared sounders and the high-dimensional problem

Temperature and water vapor are atmospheric parameters of high importance for weather forecast and atmospheric chemistry studies [24, 25]. Observations from spaceborne high spectral resolution

infrared sounding instruments can be used to calculate the profiles of such atmospheric parameters with unprecedented accuracy and vertical resolution [26]. The Infrared Atmospheric Sounding Interferometer (IASI) [27,28], in which we will focus in this work, is a Fourier-transform instrument onboard the MetOp-A satellite. IASI spectra consist of 8461 spectral channels, between 3.62 and 15.5 $\mu$m, with a spectral resolution of 0.5 cm$^{-1}$ after apodization. Its spatial resolution is 25 km at nadir with an Instantaneous Field of View (IFOV) size of 12 km at a satellite altitude of 819 km. The huge input data dimensionality typically requires simple and computationally efficient data processing techniques.

One of the retrieval techniques available in the MetOp-IASI L2 PPF is a computationally inexpensive method based on linear regression of the principal components of the measured brightness temperature spectra and the atmospheric state parameters. However, linear methods often fail to reproduce the nonlinear dependencies in the state vector. Recent works have successfully tackled the problem of atmospheric parameter retrieval by using alternative nonlinear regression methods, such as artificial neural networks [29, 30], and kernel ridge regression [20–22]. We aim to introduce DRR in such scheme as an alternative to PCA. In this application it is important that dimensionality reduction minimizes the reconstruction error and that the identified features are useful in the retrieval stage.

### 3.2. Data Collection

We used a collection of 23 datasets of input (IASI/AMSU/MHS) data and output atmospheric variables (e.g. temperature, moisture, surface pressure). Results were similar for all the datasets so, for the sake of simplicity, we show absolute results for the first dataset and relative results for all the datasets. In each dataset, the input data are 110-dimensional. In particular, each dimension corresponds to: secant of satellite zenith angle (dim 1), radiance in 14 AMSU channels (channel 7 excluded, dims [2-14]), radiance in 5 MHS channels (dims [15 20]), 30 leading IASI band 1 PC scores (dims [21 50]), 30 leading IASI band 2 PC scores (dims [51 80]), and 30 leading IASI band 3 PC scores (dims [81 110]). The output data is 277 dimensional, in particular each dimension corresponds to: Ta, Wa, Ts, Sp (hPa), T profile (K) at 91 model levels, W profile (W) dew point temperature at 91 model levels, O profile (K) at 91 model levels, and a quality indicator computed as the mean absolute error of Ta, Ts, Wa and T[79], T[86], T[90], W[86], and W[90].

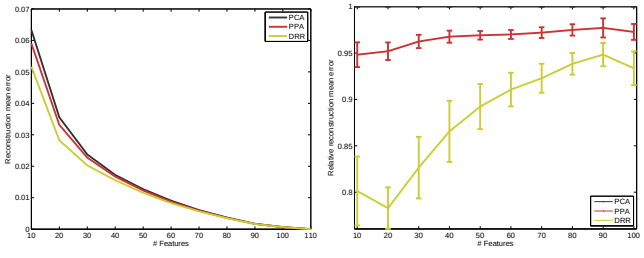


**Fig. 1**. **Reconstruction error.** Left: Absolute reconstruction error for different number of retained features obtained when using different DR methods on the first (just one) dataset. Right: Relative error (percentage) with regard to the error in PCA, mean and standard deviation have been obtained over the 23 (all) datasets.

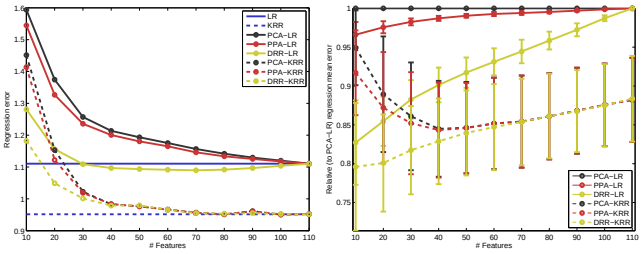


**Fig. 2**. Retrieval performance. Accuracy of the parameter retrieval (MAE) with regard to the number of retained features. Results are given for different feature extraction (PCA,DRR) and regression (LR,KRR) methods. Left: Absolute MAE for the first dataset. The performance has to be compared to the baseline behavior obtained using LR or KRR on the original data (full dimension in the input representation) represented by the blue lines. Right: Relative (to the PCA-LR MAE in each dimension) results. Results have been computed for the 23 dataset.

## 4. EXPERIMENTAL RESULTS

We evaluate the performance of DRR in terms of both the reconstruction error and the expressive power of the features to perform prediction of atmospheric profiles and physical variables.

### 4.1. Reconstruction error

In this experiment, we study the representation power of a small number of features extracted by DRR. The 110 input features are processed with PCA [15], PPA [16, 17] and the presented DRR method. Comparison of DRR with PPA is sensible due to the formal similarity between these nonlinear generalizations of PCA, and because PPA is better than previously reported techniques such as NLPCA [10, 12] or SPCA [23]. On the one hand, PPA overperforms NLPCA in reconstruction error [17], and, on the other hand, PPA is computationally feasible in high dimensional scenarios, as opposed to SPCA. Here the quality of the transformation is evaluated solely with the error in the input space computed from the original signals, $X$, and the obtained from the $r$ most relevant coefficients retained, $X_r$, i.e. $\mathcal{E} = \|X - X_r\|_F^2$. Figure 1 illustrates the effect of reconstructing the input data when using PCA, PPA and DRR for different numbers of components. Results in absolute and relative terms show that DRR obtains less reconstruction error than PCA for an arbitrary number of features.

### 4.2. Retrieval accuracy

Figure 2 illustrates the effect of using the features either from PCA, PPA or DRR for the retrieval of physical parameters. We used both linear regression and KRR in the features-to-parameters estimation. We plot the mean absolute error (MAE) for all the variables. These plots show the effect of using different (linear and non-linear) configurations for dimensionality reduction and retrieval. Using DRR features to estimate the state vectors has clear benefits. For instance, in the linear regressor framework (LR, solid lines), using just the 25% of the DRR features obtains the same accuracy as PCA when using all the components. Moreover, the benefits of using non-linear methods are clearer when combining them: when using DRR and the nonlinear regression, just 14% of the features are necessary to achieve the same performance as PCA combined with LR.

## 5. CONCLUSIONS

We introduced a novel method for dimensionality reduction via the application of a nonlinear regression to approximate each projection onto the principal directions from a subset of the other PC scores. The method is shown to generalize PCA and to achieve more data compression (smaller MSE for a fixed number of retained components) and better features for prediction (less approximation error in regression) than competitive methods like PCA and PPA. Besides, unlike other nonlinear dimensionality reduction methods, DRR is easy to apply, it has out-of-sample extension, it is invertible, and the learned transformation is volume-preserving. We focused on the challenging problem of atmospheric parameter retrieval from hyperspectral infrared sounding data. Extension of DRR to cope with multiset/output regression, as well as impact of the data dimensionality and noise sources, will be explored in the future. It is also planned to extend this study using as input IASI infrared radiances alone.

## 6. ACKNOWLEDGMENTS

The authors wish to thank Tim Hultberg from the EUMETSAT in Darmstadt, Germany, for kindly providing the IASI datasets.

## 7. REFERENCES

[1] J. A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer, 2007.

[2] Joshua B. Tenenbaum, Vin Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.

[3] S. T. Roweis, L. K. Saul, and G. E. Hinton. Global coordination of local linear models. In *Advances in Neural Information Processing Systems 14*, pages 889–896. MIT Press, 2002.

[4] J. J. Verbeek, N. Vlassis, and B. Krose. Coordinating principal component analyzers. In *In Proc. International Conference on Artificial Neural Networks*, pages 914–919. Springer, 2002.

[5] Y. W. Teh and S. Roweis. Automatic alignment of local representations. In *NIPS 15*, pages 841–848. MIT Press, 2003.

[6] Matthew Brand. Charting a manifold. In *NIPS 15*, pages 961–968. MIT Press, 2003.

[7] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, December 2000.

[8] B. Schölkopf, A. J. Smola, and K-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comp.*, 10(5):1299–1319, 1998.

[9] K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. In *Proc. IEEE CVPR*, pages 988–995, 2004.

[10] M. A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991.

[11] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, July 2006.

[12] M. Scholz, M. Fraunholz, and J. Selbig. *Nonlinear principal component analysis: neural networks models and applications*, chapter 2, pages 44–67. Springer, 2007.

[13] P. Huber. Projection pursuit. *Annals of Statistics*, 13(2):435–475, 1985.

[14] V. Laparra, G. Camps-Valls, and J. Malo. Iterative gaussianization: from ICA to random rotations. *IEEE Trans. Neur. Net.*, 22, 2011.

[15] I.T. Jolliffe. *Principal component analysis*. Springer, 2002.

[16] V. Laparra, S. Jiménez, G. Camps-Valls, and J. Malo. Nonlinearities and adaptation of color vision from sequential principal curves analysis. *Neural Comp.*, 24(10):2751–88, 2012.

[17] V. Laparra, S. Jiménez, D. Tuia, G. Camps-Valls, and J. Malo. Principal polynomial analysis. *Submitted to Int. J. Neur. Syst. (Invited Paper)*, 2014.

[18] G. Camps-Valls, D. Tuia, L. Gómez, S. Jiménez, and J. Malo. *Remote Sensing Image Processing. Synth. Lect. on Image, Video and Multimedia Proc.*, chapter The Statistics of Remote Sensing Images. Morgan & Claypool Publ., 2011.

[19] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[20] G. Camps-Valls, L. Guanter, J. Muñoz, L. Gómez, and X. Calbet. Nonlinear retrieval of atmospheric profiles from MetOp-IASI and MTG-IRS data. In *Proc. Im. Sig. Proc. Rem. Sens. XVI*, volume 7830, page 78300Z. SPIE, 2010.

[21] G. Camps-Valls, V. Laparra, J. Muñoz, L. Gómez, and X. Calbet. Kernel-based retrieval of atmospheric profiles from IASI data. In *IEEE Proc. IGARSS 11*, pages 2813–2816, Jul 2011.

[22] G. Camps-Valls, J. Muñoz and, L. Gómez, L. Guanter, and X. Calbet. Nonlinear statistical retrieval of atmospheric profiles from MetOp-IASI and MTG-IRS infrared sounding data. *IEEE Trans. Geosci. Rem. Sens.*, 50(5):1759–1769, 2012.

[23] V. Laparra, S. Jiménez, G. Camps-Valls, and J. Malo. Nonlinearities and adaptation of color vision from sequential principal curves analysis. *Neural Comp.*, 24(10):2751–88, 2012.

[24] K. N. Liou. *An Introduction to Atmospheric Radiation*. Academic Press, Hampton, USA, second edition, 2002.

[25] F. Hilton, N. C. Atkinson, S. J. English, and J. R. Eyre. Assimilation of IASI at the Met Office and assessment of its impact through observing system experiments. *Q. J. R. Meteorol. Soc.*, 135:495–505, 2009.

[26] H. L. Huang, W. L. Smith, and H. M. Woolf. Vertical resolution and accuracy of atmospheric infrared sounding spectrometers. *J. Appl. Meteor.*, 31:265–274, 1992.

[27] G. Chalon, F. Cayla, and D. Diebel. IASI: an advanced sounder for operational meteorology. In *Proceedings of the 52nd Congress of IAF*, Toulouse, France, 2001.

[28] Chalon G. Siméoni D., Singer C. Infrared atmospheric sounding interferometer. *Acta Astronautica*, 40:113–118, 1997.

[29] F Aires. A regularized neural net approach for retrieval of atmospheric and surface temperatures with the IASI instrument. *Journal of Applied Meteorology*, 41:144–159, 2002.

[30] W.J. Blackwell. A neural-network technique for the retrieval of atmospheric temperature and moisture profiles from high spectral resolution sounding data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(11), 2005.