Nonlinear Dimensionality Reduction for Regression

Minyoung Kim[1] and **Vladimir Pavlovic**[1]

[1]Department of Computer Science
Rutgers University
Piscataway, NJ 08854
{mikim,vladimir}@cs.rutgers.edu

The task of *dimensionality reduction for regression* (DRR) is to find a low dimensional representation $\mathbf{z} \in \mathbb{R}^q$ of the input covariates $\mathbf{x} \in \mathbb{R}^p$, with $q \ll p$, for regressing the output $\mathbf{y} \in \mathbb{R}^d$. DRR can be beneficial for visualization of high dimensional data, efficient regressor design with a reduced input dimension, but also when eliminating noise in data $\mathbf{x}$ through uncovering the essential information $\mathbf{z}$ for predicting $\mathbf{y}$. However, while dimensionality reduction methods are common in many machine learning tasks (discriminant analysis, graph embedding, metric learning, principal subspace methods) their use in regression settings has not been widespread.

The crucial notion related to DRR is the *sufficiency in dimension reduction* (SDR, [1, 2, 3]), which states that one has to find the linear subspace bases $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_q]$ with $\mathbf{b}_l \in \mathbb{R}^p$, (in the nonlinear case, $\mathbf{B} = \{\mathbf{b}_1(\cdot), \ldots, \mathbf{b}_q(\cdot)\}$, where $\mathbf{b}_l(\cdot)$ is a nonlinear basis function) such that $\mathbf{y} \perp \mathbf{x} \mid \mathbf{B}^\top \mathbf{x}$. As this condition implies that the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ equals to that of $\mathbf{y}$ given $\mathbf{z} = \mathbf{B}^\top \mathbf{x}$, the dimension reduction entails no loss of information for the purpose of regression. The minimal subspace with this property is called the *central subspace*[1].

A number of methods originating in the statistics community have tackled the task of recovering the central space. The kernel dimension reduction (KDR) [2] and the manifold KDR [4] directly reduces the task of imposing conditional independence to the optimization problem that minimizes the conditional covariance operator in RKHS (reproducing kernel Hilbert space). However, the manifold KDR introduces a tight coupling between the central space and the separately learned input manifold, which restricts its applicability to transductive settings. Moreover, both methods introduce non-convex objectives, potentially suffering from existence of local minima. An alternative inverse regression (IR) approach [3, 5] exploits the fact that the inverse regression $\mathbb{E}[\mathbf{x}|\mathbf{y}]$ can lie on the subspace spanned by $\mathbf{B}$, leading to the possibility of estimating $\mathbf{B}$ from the slice-driven covariance estimates of the IR. While KSIR [5] overcomes the linearity of SIR [3] its performance may still suffer from the need for target $\mathbf{y}$ slicing, which can be unreliable for high-dimensional targets.

In this work we propose the Covariance Operator based Inverse Regression (COIR), a novel nonlinear method for DRR that jointly exploits the kernel Gram matrices of both input and output. COIR estimates the variance of the inverse regression under the IR framework and, at the same time, avoids the slicing by the effective use of covariance operators in RKHS. As a result the central subspace effectively arises from the solution of the eigenequation

$$\frac{1}{n}\mathbf{K}_\mathbf{y}(\mathbf{K}_\mathbf{y} + n\epsilon \mathbf{I}_n)^{-1}\mathbf{K}_\mathbf{x}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha},$$

as

$$\mathbf{z}(x) = n\mathbf{k}(x)^T {\mathbf{K}_\mathbf{x}}^{-1}\boldsymbol{\alpha}.$$

While this approach generalizes that of KSIR (a special case of COIR) it also allows a closed-form solution to the nonlinear central subspace estimation problem. We demonstrate the benefits of the proposed method on two regression problems involving high-dimensional and noisy data.

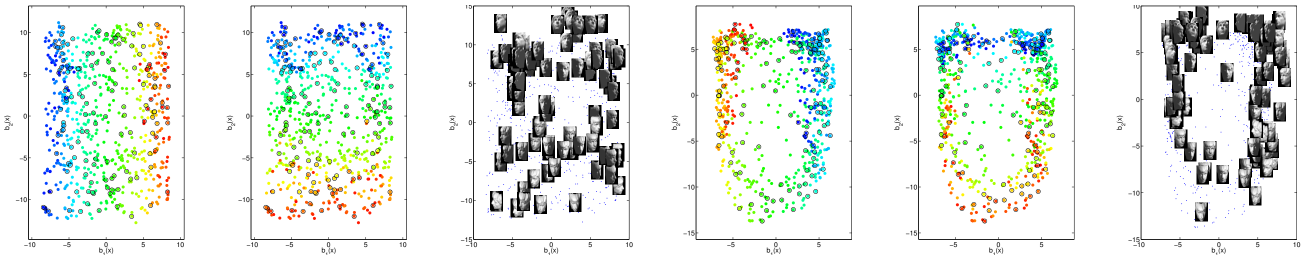
A. Estimation of Head Pose

From the face dataset (http://isomap.stanford.edu/datasets.html), we consider the task of predicting the 2D pose (horizontal and vertical rotation angles) from (64×64) image. We show 2D central subspaces estimated by COIR and KSIR in Fig. 1. COIR lays out the data almost linearly along the output values with good generalization. In KSIR, however, the data points are often mixed significantly (e.g., the red points intermingled with the green points).

B. Hand-written Digit Image Denoising

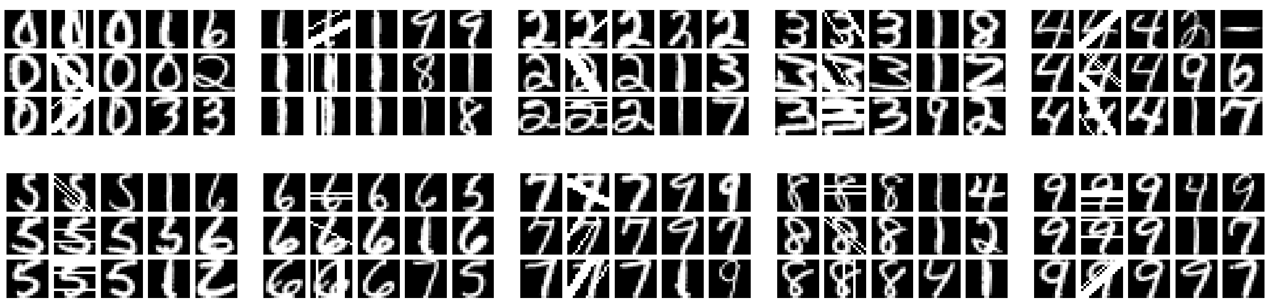
We devise an image denoising experiment with the USPS hand-written digit images. By adding highly non-iid noise, random scratch lines with varying thickness and orientation, on the normalized (16×16) digit images, we consider the task of predicting the original unscratched image (output $\mathbf{y}$) from the scratched version (input $\mathbf{x}$). The test RMS errors using NN and Gaussian Process regressors learned with the dimension-reduced input data, and selected denoised test images by the NN regression are depicted in Fig. 2. We see that COIR is robust to noise with improved prediction accuracy compared to the regression based on the image input itself.

[1]Although the *subspace* is usually meant for a linear case, however, we abuse it referring to both linear and nonlinear cases.



(a) COIR; L/R (b) COIR; U/D (c) COIR; Face (d) KSIR; L/R (e) KSIR; U/D (f) KSIR; Face

Figure 1: 2D central subspaces of face images. The points are colored by the true Left/Right pose angles in (a)/(d), Up/Down in (b)/(e), and (c)/(f) shows the face images superimposed. The test points are depicted in black circles.



Input Space	COIR	KSIR	Image **x**	Input Space	COIR	KSIR	Image **x**
NN Regression	8.5334	11.4909	9.3605	GP Regression	8.1454	10.7259	9.1036

Figure 2: Denoising USPS scratched digit images. **Top**: Each 5-tuple is composed of, from left to right, (1^{st}) the noise-free test image, (2^{nd}) randomly scratched image, (3^{rd}) denoised by NN on COIR, (4^{th}) NN on KSIR, and (5^{th}) NN on the scratched image $\mathbf{x}$ itself. **Bottom**: Test (RMS) errors using NN and Gaussian Process regressors.

Topic: Learning Algorithms, Pattern Recognition
Preference: Oral/Poster

References

[1] R. D. Cook. *Regression graphics*. Wiley Inter-Science, 1998.

[2] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 2004.

[3] K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 1991.

[4] J. Nilsson, F. Sha, and M. I. Jordan. Regression on manifolds using kernel dimension reduction, 2007. ICML.

[5] H. M. Wu. Kernel sliced inverse regression with applications on classification, 2006. ICSA Applied Statistics Symposium.