

Discussion

Ying Nian Wu
UCLA Department of Statistics

The Population Value Decomposition (PVD) proposed by this paper is an interesting advance in analyzing massive high-dimensional data. I am impressed by the simplicity of the model and the associated computational algorithm. Its application in the Sleep Heart Health Study (SHHS) demonstrates the usefulness of the proposed methodology.

The proposed computational algorithm is based on subject-specific singular value decompositions. Is it possible to find a more rigorous algorithm that minimizes some objective function?

The proposed model assumes the same \mathbf{P} and \mathbf{D} for the whole population. If the population consists of multiple clusters, it is possible that different clusters may have different \mathbf{P} and \mathbf{D} . Is it possible to extend the model and algorithm to address this issue?

As the authors point out, the proposed method can be considered a multi-stage Principal Component Analysis (PCA). As such, it shares the limitations of PCA, such as the inability to capture the non-Gaussian and non-linear properties in the data. While the proposed method appears very sensible for SHHS data, it may not be adequate for other types of image data, such as natural scene images.

As to dimension reduction, it is worthwhile to mention the work of Olshausen and Field (1996) on sparse coding that goes beyond PCA or factor analysis. For PCA, one finds a small number of orthogonal basis vectors that capture most of the variations in the data. In sparse coding, however, one finds a large dictionary of basis vectors, that are not necessarily orthogonal to each other, so that each observed signal can be represented by a small number of basis vectors selected from the dictionary, but different signals may be represented by different sets of selected basis vectors.

Specifically, Olshausen and Field (1996) consider the modeling of natural image patches (e.g., 12×12 images, so the signal is 144 dimensional vector). Let $\{\mathbf{I}_m, m = 1, \dots, M\}$ be the set of M image patches. They are represented by the following linear model:

$$\mathbf{I}_m = \sum_{k=1}^K c_{m,k} B_k + \epsilon_m, \quad (1)$$

where each B_k is a basis vector that is of the same dimensionality as \mathbf{I}_m , and $c_{m,k}$ is the coefficient. In the language of linear regression, \mathbf{I}_m is the response vector, and $(B_k, k = 1, \dots, K)$ are the regressors or predictors. It is often assumed that the number of regressors K is greater than the dimensionality of the response vector (which is called “ $p > n$ ” problem in regression). Meanwhile, it is also assumed that $(c_{m,k}, k = 1, \dots, K)$ is sparse in that for each \mathbf{I}_m , only a small number of $c_{m,k}$ are non-zero (or significantly different from 0). Given the dictionary of regressors $(B_k, k = 1, \dots, K)$, inferring $(c_{m,k}, k = 1, \dots, K)$ is a variable selection problem. But here the twist is that the regressors $(B_k, k = 1, \dots, K)$ are unknown, and they are to be learned from the training data $\{\mathbf{I}_m, m = 1, \dots, M\}$. Interestingly, by enforcing sparsity on $(c_{m,k}, k = 1, \dots, K)$, the $(B_k, k = 1, \dots, K)$ learned from natural image patches are localized, oriented and elongated wavelets. This provides a statistical justification for the use of wavelets in representing natural images.

The sparsity of $(c_{m,k}, k = 1, \dots, K)$ leads to dimension reduction of \mathbf{I}_m . However, unlike PCA, the dimension reduction in sparse coding is adaptive or subject-specific, because for different m , the sets of non-zero $c_{m,k}$ can be different. This is much more flexible than PCA. It is also related to

the clustering issue mentioned above, where different clusters may lie in different low-dimensional sub-spaces.

Recently, in Wu, Si, Gong, and Zhu (2010), we have attempted to model such clusters. First we assume that the basis vectors are already learned or designed, so there is a dictionary of localized, oriented and elongated wavelets $\{B_{x,s,\alpha}\}$, indexed or attributed by location x , scale s and orientation α . Each $B_{x,s,\alpha}$ is like a stroke for sketching the image. Then we model each cluster by

$$\mathbf{I}_m = \sum_{i=1}^n c_{m,i} B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}} + \epsilon_m, \quad (2)$$

where $(B_{x_i, s, \alpha_i}, i = 1, \dots, n)$ is the set of a small number n of basis vectors selected from the dictionary for representing the cluster. $(B_{x_i, s, \alpha_i}, i = 1, \dots, n)$ is like a template with n strokes. We allow small perturbations $(\Delta x_{m,i}, \Delta \alpha_{m,i}, i = 1, \dots, n)$ in locations and orientations, so that the template is deformable. Different clusters are represented by different templates $(B_{x_i, s, \alpha_i}, i = 1, \dots, n)$. We assume that the scale s is fixed.

We have done some preliminary experiments on finding such clusters. Figure 1 displays 4 templates obtained from 320 images (120×120 pixels) of animal faces by model-based clustering, where in each template $(B_{x_i, s, \alpha_i}, i = 1, \dots, n = 60)$, B_{x_i, s, α_i} is illustrated by a bar at the location x_i , scale s and orientation α_i .



Figure 1: Templates learned from images of animal faces by model-based clustering. Each template consists of a set of wavelet basis elements, each of which is illustrated by a bar. Number of training images = 320; Image height and width = 120×120 pixels; Number of clusters = 4; Number of selected wavelet elements = 60.

It is still unclear whether one could scale up the clustering experiments to learn thousands of templates or part-templates from image patches of natural scenes or various object categories. The templates of those clusters may become the “visual words” for representing images, and these visual words lead to sparser representations of natural images than wavelets.

I would also like to mention the recent work of Hinton and Salakhutdinov (2006) on dimension reduction based on the so-called “auto-encoder” network, which is a multi-layer neural network with a low-dimensional central layer for reconstructing the high-dimensional observed signal. The connection weights of this network are pre-trained by learning restricted Boltzmann machine layer by layer. This auto-encoder network appears to be able to capture some structures that elude PCA.

The above dimension reduction methods may not be applicable to the data that the authors deal with. I bring them up mainly to expand the discussion of existing tools for unsupervised learning. I would like to end my discussion by applauding what the authors have achieved in this interesting paper.

Acknowledgment

I would like to acknowledge the support of NSF DMS 1007889.

References

- [1] Hinton, G. E. and Salakhutdinov, R. R. (2006) Reducing the dimensionality of data with neural networks. *Science*, 313, 504-507.
- [2] Olshausen, B. A. and Field, D. J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381, 607-609.
- [3] Wu, Y. N., Si, Z., Gong, H., and Zhu, S. C. (2010) Learning active basis model for object detection and recognition. *International Journal of Computer Vision*, 90, 198-235.