Comment

E. F. Lock, A. B. Nobel, and J. S. Marron

The article by Crainiceanu et al. addresses an important and relatively undeveloped area of statistical research: the analysis of populations in which the data objects are matrices. In particular, they focus on collections of matrices that have the same row and column dimensions. Such datasets are increasingly prevalent in a number of scientific fields. Examples range from the analysis of facial data in image analysis, EEG data in neuroscience, fMRI data in medical imaging, and browsing data in the study of Internet traffic (see Table 1).

The method proposed by the authors, population value decomposition (PVD), is a useful way to simultaneously reduce the dimensionality of a collection of matrices. For matrices $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$, each of dimension $F \times T$, the PVD yields an approximation $\mathbf{Y}_i \approx \mathbf{P}\mathbf{V}_i\mathbf{D}$ for $i = 1, \ldots, n$. The low-dimensional representation $\mathbf{V}_i$ for each matrix is on a common set of coordinates determined by $\mathbf{P}$ and $\mathbf{D}$. This allows for the application of standard statistical approaches, such as regression and cluster analysis, to the lower-dimensional matrices $\mathbf{V}_i$, rather than the $\mathbf{Y}_i$'s. Furthermore, inspection of the population-wide left and right loading matrices $\mathbf{P}$ and $\mathbf{D}$ can aid in identifying the primary modes of variation among a population of matrices.

The authors present an interesting data analysis; however, we note that a model formally equivalent to PVD has been proposed in the computer science literature under the name *two-dimensional singular value decomposition* (2DSVD) (Ding and Ye 2005; Ye 2005). This literature also provides natural additional approaches for choosing the population-wide matrices $\mathbf{P}$ and $\mathbf{D}$. We discuss these approaches in Section 1.

We may regard the problem addressed in these articles by viewing the data as a three-way ($F \times n \times T$) array. In Section 2 we discuss and compare alternative approaches that treat the data structure as a three-way array. We show that two SVD-like decompositions for higher-order arrays, those of *Candecomp/Parafac* (Carroll and Chang 1970) and *Tucker* (Tucker 1966), are related to the PVD decomposition. In fact, both can be represented in the form $\mathbf{P}\mathbf{V}_i\mathbf{D}$, in which the $\mathbf{V}_i$ matrices have a particular structure.

In Section 3 we discuss some important issues and caveats related to the application of PVD and related methods. In Section 4 we compare PVD and other methods in an application to facial image data.

1. CHOICE OF P AND D

The PVD article suggests determining the entries of the individual matrices $\mathbf{V}_i$ via standard least squares regression, for a given choice of the population-wide matrices $\mathbf{P}$ and $\mathbf{D}$. However, the default method for choosing $\mathbf{P}$ and $\mathbf{D}$ is somewhat ad hoc, and a more principled approach would be desirable. We suggest formulating the estimation of $\mathbf{P}$, $\mathbf{D}$, and the $\mathbf{V}_i$ matrices together as a single least squares problem. That is, for given dimensions $A < F$ and $B < T$, find $\mathbf{P}: F \times A$, $\mathbf{D}: B \times T$, and $\mathbf{V}_i: A \times B$, $i = 1, \ldots, n$, to minimize the sum of squared residuals

$$\sum_{i=1}^{n} \|\mathbf{Y}_i - \mathbf{P}\mathbf{V}_i\mathbf{D}\|_F^2. \tag{1}$$

Here $\|\cdot\|_F$ defines the Frobenious norm; that is, $\|\mathbf{A}\|_F^2$ is just the sum of the squared entries of $\mathbf{A}$.

This approach to estimating the PVD model was previously explored by Ye (2005). Ye suggested an iterative least squares procedure that cycles among estimation of the matrices $\mathbf{P}$, $\mathbf{V}_1, \ldots, \mathbf{V}_n$, and $\mathbf{D}$ until convergence. Although this iterative procedure is not guaranteed to achieve the global minimum in criterion (1), Ye argued that the algorithm is insensitive to starting conditions and is generally successful at minimizing the sum of squared residuals.

An alternative approach for choosing $\mathbf{P}$ and $\mathbf{D}$, termed 2DSVD by Ding and Ye (2005), makes use of the aggregated row–row and column–column covariance matrices. In particular, $\mathbf{P}$ is determined by the first A singular vectors of the row-by-row covariance matrix $\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i\mathbf{Y}_i'$ and $\mathbf{D}$ determined by the first B singular vectors of the column-by-column covariance matrix $\frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i'\mathbf{Y}_i$. Equivalently, $\mathbf{P}$ can be computed as the first A left singular vectors of the aggregated matrix $[\mathbf{Y}_1\ \mathbf{Y}_2\ \cdots\ \mathbf{Y}_n]$, and $\mathbf{D}$ can be computed as the first B right singular vectors of $[\mathbf{Y}_1'\ \mathbf{Y}_2'\ \cdots\ \mathbf{Y}_n']'$. Although both computations give the same result, the latter may be more efficient if one of the dimensions is particularly large, and computing the covariance is impractical.

The justification for the 2DSVD algorithm is that the columns of $\mathbf{P}$ are chosen as the set of left-singular vectors that explain the most total variation across the columns of $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, and the rows of $\mathbf{D}$ are (independently) chosen as the set of right singular vectors that explain the most variation across the rows of $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. Because interactions between $\mathbf{P}$ and $\mathbf{D}$ are not accounted for, the resulting matrices do not necessarily minimize criterion (1), but they should come close to doing so. Indeed, 2DSVD could be used to determine the initial matrices $\mathbf{P}_0$ and $\mathbf{D}_0$ for an iterative least squares procedure, such as that described earlier.

The "default" method for estimating $\mathbf{P}$ and $\mathbf{D}$ proposed in the PVD article is essentially a two-stage SVD. The first few singular vectors of each matrix are found separately, then another SVD of the combined singular vectors determines the global left and right singular vectors $\mathbf{P}$ and $\mathbf{D}$. This method requires

Eric F. Lock is Doctoral Student (E-mail: *Eric.F.Lock@gmail.com*), Andrew B. Nobel is Professor (E-mail: *nobel@email.unc.edu*), and J. S. Marron is Professor (E-mail: *marron@email.unc.edu*), Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27514. This research was supported in part by National Institutes of Health grant R01 MH090936-01 and National Science Foundation grant DMS-0907177.

Journal of the American Statistical Association
September 2011, Vol. 106, No. 495, Applications and Case Studies
DOI: 10.1198/jasa.2011.ap11236

Table 1. Data examples where the objects are matrices of the same dimension

Objects	Value	Dimensions
Facial images	pixel intensity	horizontal × vertical
EEG recordings	electrical activity	frequency × time
fMRI scans	blood flow	voxel position × time
Browsing histories	visits from website *i* to website *j*	websites × websites

specifiying the number of singular vectors to take for each matrix, which may be somewhat arbitrary. We feel that the 2DSVD and least squares methods described earlier are more justified, and in most cases will be computationally simpler. However, in some datasets the aggregated matrix $[\mathbf{Y}_1 \ \cdots \ \mathbf{Y}_k]$ and/or one of $\mathbf{Y}'\mathbf{Y}$, $\mathbf{Y}\mathbf{Y}'$ are too large to store in memory. In these cases, the individual-level data compression used by the two-stage SVD may be necessary.

Although we have presented various alternative methods for selecting $\mathbf{P}$ and $\mathbf{D}$, we share the authors' opinion that the ideal choice of $\mathbf{P}$ and $\mathbf{D}$ may depend on the particular type of data, and there is no perfect method.

2. THREE–WAY METHODS

The authors apply PVD to data in which an EEG-based activity matrix of *Frequency* × *Time* is available for multiple subjects. Note that these data can be framed as a three-way array: *Frequency* × *Subject* × *Time*. Indeed, any dataset analyzed with PVD can be considered a three-way array of dimension $F \times n \times T$. In PVD, the second mode (subject, of dimension n) is treated differently than the other two modes. In this section we consider some other SVD-like decompositions for multiway arrays that treat all three modes similarly. These methods are also appropriate for multiway arrays with more than three modes. Thus they have potential for the analysis of fMRI data that is truly multidimensional: *Length* × *Width* × *Height* × *Time* × *Subject*.

There are two standard SVD-like extensions to multiway data: the Candecomp/Parafac and Tucker decompositions. Both have been studied in the analysis of tensors for several years, but are not widely known. The survey by Kolda and Bader (2009) is a well-written and accessible introduction to tensor notation, the aformentioned decompositions, and related software. We briefly discuss their relationship to PVD here, but refrain from using notation that may be unfamiliar.

2.1 The Candecomp/Parafac Decomposition

The Candecomp/Parafac (Carroll and Chang 1970) decomposition extends the notion of the SVD as a sum of rank-1 approximations. We can approximate an $F \times T$ matrix $\mathbf{Y}$ by combining the first r left singular vectors and corresponding right singular vectors; that is, the columns of $\mathbf{U}^{(1)}: F \times r$ are the first r left singular vectors of $\mathbf{Y}$, and the columns of $\mathbf{U}^{(2)}: T \times r$ are the first r right singular vectors of $\mathbf{Y}$, scaled appropriately, then

$$y_{ij} \approx \sum_{l=1}^{r} u_{il}^{(1)} u_{jl}^{(2)}$$

for $i = 1, \ldots, F$, $j = 1, \ldots, T$.

For a three-way array $\mathbf{Y}: F \times n \times T$, then, the Parafac decomposition yields matrices $\mathbf{U}^{(1)}: F \times r$, $\mathbf{U}^{(2)}: n \times r$, and $\mathbf{U}^{(3)}: T \times r$, so that

$$y_{ijk} \approx \sum_{l=1}^{r} u_{il}^{(1)} u_{jl}^{(2)} u_{kl}^{(3)}$$

for $i = 1, \ldots, F$, $j = 1, \ldots, n$, and $k = 1, \ldots, T$. The matrix $\mathbf{U}^{(i)}$ serves as a low-dimensional representation for variation in the ith mode.

The three-way Parafac decomposition also can be represented in the framework of the PVD model. If $\mathbf{P} := \mathbf{U}^{(1)}$, $\mathbf{D} := \mathbf{U}^{(3)}$, and $\mathbf{V}_j$ is a diagonal matrix whose entries are from the jth column of $\mathbf{U}^{(2)}$, $\mathbf{V}_j = \operatorname{diag}(\mathbf{U}_{\cdot j}^{(2)})$, then

$$\mathbf{Y}_{\cdot j \cdot} \approx \mathbf{P}\mathbf{V}_j\mathbf{D}$$

for $j = 1, \ldots, n$. Thus the three-way Parafac decomposition can be considered a PVD model in which the $\mathbf{V}_j$ matrices are diagonal.

2.2 The Tucker Decomposition

For a standard (two-mode) SVD, combining the ith left singular vector and the jth right singular vector does not improve an approximation when $i \neq j$. No such result holds for higher-order arrays. The Tucker decomposition (Tucker 1966), then, considers all combinations from a set of basis vectors in each mode. Thus a three-way Tucker decomposition consists of matrices $\mathbf{U}^{(1)}: F \times r_1$, $\mathbf{U}^{(2)}: n \times r_2$, and $\mathbf{U}^{(3)}: T \times r_3$ and a $r_1 \times r_2 \times r_3$ tensor $\boldsymbol{\Lambda}$, where

$$y_{ijk} \approx \sum_{l_1=1}^{r_1} \sum_{l_2=1}^{r_2} \sum_{l_3=1}^{r_3} \lambda_{l_1 l_2 l_3} u_{il_1}^{(1)} u_{jl_2}^{(2)} u_{kl_3}^{(3)}.$$

Here the ijkth entry of the tensor $\boldsymbol{\Lambda}$ weights the interactions among the ith column of $\mathbf{U}^{(1)}$, the jth column of $\mathbf{U}^{(2)}$, and the kth column of $\mathbf{U}^{(3)}$. The Parafac decomposition is a special case of the Tucker model where $r_1 = r_2 = r_3$ and $\lambda_{l_1 l_2 l_3} = 0$ unless $l_1 = l_2 = l_3$. Again, the matrix $\mathbf{U}^{(i)}$ serves as a low-dimensional representation for variation in the ith mode.

The three-way Tucker decomposition also can be given in the PVD framework, where the matrices $\mathbf{V}_j$ have a particular factorized form. If $\mathbf{P} := \mathbf{U}^{(1)}$ and $\mathbf{D} := \mathbf{U}^{(3)}$, then

$$\mathbf{Y}_{\cdot j \cdot} \approx \mathbf{P}\mathbf{V}_j\mathbf{D},$$

where $\mathbf{V}_j: r_1 \times r_3$ is

$$\mathbf{V}_{\cdot j \cdot} := \sum_{l_2=1}^{r_2} u_{jl_2}^{(2)} \boldsymbol{\Lambda}_{\cdot l_2 \cdot}.$$

Intuitively, here each $\mathbf{V}_j$ can be considered a weighted combination of basis matrices $\boldsymbol{\Lambda}_{\cdot 1 \cdot}, \ldots, \boldsymbol{\Lambda}_{\cdot r_2 \cdot}$, where the weights specific to the jth individual are given by the jth row of $\mathbf{U}^{(2)}$.

3. POTENTIAL ISSUES

There are important caveats in the application of PVD and related methods, such as those discussed in the previous section. We briefly discuss four common issues that must be considered before describing an application of PVD.

3.1 Registration

The PVD approach requires that the coordinates of the matrices $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ be aligned; that is, the (i,j) entry of the matrix $\mathbf{Y}_1$ must correspond to the (i,j) entry of the matrix $\mathbf{Y}_2$, and so on. This is a common issue in the analysis of image populations, where a slight shift or rotation of perspective can cause difficulties when integrating information across the images. Here registration methods that transform a collection of images to the same coordinate system can be useful. For an overview of image registration methods, see the survey by Zitova and Flusser (2003).

3.2 Scaling

Direct application of PVD also may be problematic if the matrices $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ differ in scale. For example, estimation of $\mathbf{P}$ and $\mathbf{D}$ can be unduly influenced by a select few matrices with a large amount of variability. A potential solution is to scale each matrix to have the same total variation, that is, sum of squares. This approach was suggested by, for example, Lock et al. (2011) as a preprocessing step for integrating across multiple data matrices (possibly of different dimension) available for the same set of objects.

To remove baseline differences between matrices, it is helpful to center the data by subtracting the overall mean from each matrix. To control total variability, one can then divide by the standard deviation of the matrix entries; that is, letting $\bar{y}_i$ be the mean and s_i be the standard deviation of the entries of $\mathbf{Y}_i$, define

$$\mathbf{Y}_i^{\text{scaled}} = \frac{\mathbf{Y}_i - \bar{y}_i}{s_i}.$$

The matrices $\mathbf{Y}_i^{\text{scaled}}$ then have the same total sum of squared entries.

We note, however, that the choice of a normalization procedure should depend on the type of data and goals of the analysis.

3.3 Dimensional Compatibility

Recall that for a rank $r+1$ SVD approximation, the first r right singular vectors, left singular vectors, and singular values remain the same. If the PVD model is estimated via minimizing the sum of squared residuals, then there is no such dimensional compatibility; that is, if either dimension A or B is changed, then all entries of the estimated matrices $\mathbf{P}$, $\mathbf{D}$, and each $\mathbf{V}_i$ may change. This is an important caveat when interpreting the columns and rows of $\mathbf{P}$ and $\mathbf{D}$. In many cases, changing A or B slightly might not lead to a dramatic change in the entries of $\mathbf{P}$, $\mathbf{D}$, and $\mathbf{V}_i$, but the stability of these estimates are worth considering. The Tucker and Parafac models described in Section 2 also are not necessarily compatible on different dimensions.

3.4 Choice of A and B

In light of the foregoing comments, the choices of A and B in the PVD approximation can be particularly important. In any case, the choices of A and B may be somewhat arbitary in practice, and a principled approach to choosing these dimensions is desired. We do not give a specific approach here, but note that certain ideas may be borrowed from related work. One potential criterion is a cross-validation–based estimate of the reconstruction error, similar to that used to determine the number of principal components by Wold (1978). Another potential approach is permutation testing, similar to the rank selection procedure described by Lock et al. (2011). Yet another potential approach may be motivated by random matrix theory (see Shabalin and Nobel 2010).

4. APPLICATION: FACIAL IMAGES

As an example, we apply PVD and related methods to the Database of Faces procured by AT&T Laboratories Cambridge. This is a publicly available database of $n = 400$ total gray-scale images for 40 individuals (10 per individual). Each image $\mathbf{Y}_i$, $i = 1, \dots, n$, is 92×112 in size. All subjects are in an upright, frontal position, but facial characteristics (e.g., smiling, not smiling; glasses, no glasses) vary in each image. We apply four factorization models to these data and compare the results. We apply the PVD model in which $\mathbf{P}$ and $\mathbf{D}$ are estimated by iteratively minimizing the sum of squared residuals, as in Section 1. We apply the Parafac and Tucker models, also estimated by least squares using the N-way MATLAB toolbox (Andersson and Rasmus 2000). We also try a SVD of the vectorized data; for each $\mathbf{Y}_i$, the rows are stacked to form a vector of length $92 \times 112 = 10{,}304$, and an SVD is applied to the resulting $10{,}304 \times 400$ matrix.

We compare these factorized approximations in terms of data compression for this example; that is, we consider the sum of squared residuals versus the total number of degrees of freedom (free parameters) needed for each model. For example, a PVD approximation with $A = B = 5$ requires $\mathbf{P}: 92 \times 5$, $\mathbf{D}: 5 \times 112$, and $\mathbf{V}_i: 5 \times 5$, $i = 1, \dots, 400$, or $92 \times 5 + 5 \times 112 + 5 \times 5 \times 400 = 11{,}020$ free parameters.

Figure 1(A) displays the sum of squared residuals for each model as the number of free parameters increases. In this analysis, for simplicity, we restrict $A = B$ for the PVD model and $r_1 = r_2 = r_3$ for the Tucker model. Relaxing these restrictions could give these methods additional power. The SVD of the vectorized data is by far the worst-performing method by this measure, whereas the other three methods are relatively comparable. This indicates that there are advantages to exploiting the two-dimensional nature of these images, rather than simply vectorizing them.

The resulting approximations for three of the facial images are shown in Figure 1(B). Here each method uses approximately 70,000 degrees of freedom, whereas the original data had $112 \times 92 \times 400 = 4{,}121{,}600$ total pixel values. The approximations resulting from an SVD of the vectorized data bear little resemblance to the original images. The other three methods are fairly comparable, although one could argue that the Parafac approximations give the best visual impression.

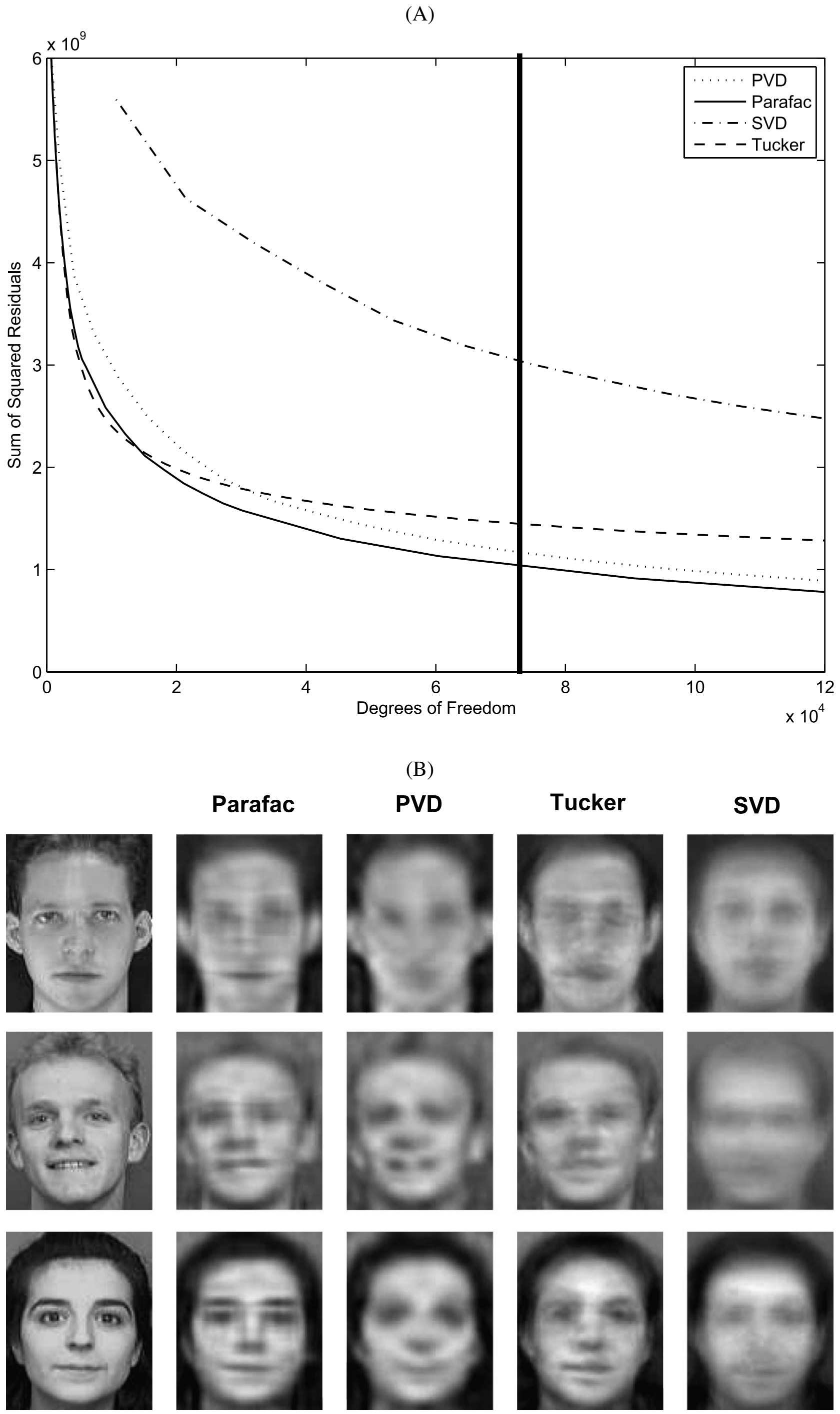

Figure 1. Application of PVD, Tucker, Parafac, and SVD factorizations to facial image data. (A) The sum of squared residuals versus the degrees of freedom used to fit the model for each method. (B) Three facial images (at left), and their reconstructions using the four methods. Each reconstruction uses similar degrees of freedom, close to the vertical line in (A). The Parafac approximation shown uses 72,480 ($r = 120$) degrees of freedom, PVD uses 70,252 ($A = B = 13$), Tucker uses 73,001 ($r_1 = r_2 = r_3 = 37$), and SVD uses 74,928 ($r = 7$).

All of the factorization methods compared here can be used to reduce the dimensionality and provide insight into the primary modes of variation among a collection of matrices. The relative success of these factorization methods will depend on the structure and dimensions of any given dataset. Here we have focused exclusively on data compression. There are other important considerations, such as which method provides the best interpretation for a given application.

ADDITIONAL REFERENCES

Andersson, C. A., and Rasmus, B. (2000), “The N-Way Toolbox for MATLAB,” *Chemometrics and Intelligent Laboratory Systems*, 52, 1–4. [800]

Carroll, J. D., and Chang, J. (1970), “Analysis of Individual Differences in Multidimensional Scaling via an N-Way Generalization of “Eckart–Young” Decomposition,” *Psychometrika*, 35, 283–391. [798,799]

Ding, C., and Ye, J. (2005), “Two-Dimensional Singular Value Decomposition (2DSVD) for 2D Maps and Images,” in *Proceedings of SIAM International*

Conference on Data Mining (SDM'05)*, Philadelphia, PA: SIAM, pp. 32–43. [798]

Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [799]

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2011), "Joint and Individual Variation Explained (JIVE) for the Integrated Analysis of Multiple Datatypes," available at *arXiv:1102.4110*. [800]

Shabalin, A. A., and Nobel, A. B. (2010), "Reconstruction of a Low-Rank Matrix in the Presence of Gaussian Noise," available at *arXiv:1007.4148*. [800]

Tucker, L. R. (1966), "Some Mathematical Notes on Three-Mode Factor Analysis," *Psychometrika*, 31, 279–311. [798,799]

Wold, S. (1978), "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models," *Technometrics*, 20, 397–405. [800]

Ye, J. (2005), "Generalized Low Rank Approximations of Matrices," *Machine Learning*, 61, 167–191. [798]

Zitova, B., and Flusser, J. (2003), "Image Registration Methods: A Survey," *Image and Vision Computing*, 21, 977–1000. [800]

Comment

Ying Nian Wu

The population value decomposition method proposed in this article 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 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 **P** and **D** for the whole population. In a population consisting of multiple clusters, it is possible that different clusters may have different **P** and **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 multistage principal component analysis (PCA). As such, it shares the limitations of PCA, such as the inability to capture the non-Gaussian and nonlinear properties in the data. Although the proposed method appears to be very sensible for SHHS data, it might 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 one another, 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) considered 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 represented by the following linear model:

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

where each B_k is a basis vector 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 (called the "$p > n$" 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 nonzero (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 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 the sets of nonzero $c_{m,k}$ can be different for different m. This is much more flexible than PCA. It is also related to the aforementioned clustering issue, where different clusters may lie in different low-dimensional subspaces.

Recently (Wu et al. 2010), we attempted to model such clusters. In our approach we first assume that the basis vectors are already learned or designed, and 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. We then model each cluster by

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

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

Ying Nian Wu is Professor, Department of Statistics, University of California, Los Angeles, CA 90095 (E-mail: *ywu@stat.ucla.edu*). I would like to acknowledge the support of NSF DMS 1007889.

Journal of the American Statistical Association
September 2011, Vol. 106, No. 495, Applications and Case Studies
DOI: 10.1198/jasa.2011.ap11332