Statistica Sinica **23** (2013), 963-988
doi:http://dx.doi.org/10.5705/ss.2011.245

METHODOLOGY AND THEORY FOR NONNEGATIVE-SCORE PRINCIPAL COMPONENT ANALYSIS

Peter Bajorski[1], Peter Hall[2] and Hyam Rubinstein[2]

[1]*Rochester Institute of Technology and* [2]*The University of Melbourne*

Abstract: We develop nonparametric methods, and theory, for analysing data on a random p-vector Y represented as a linear form in a p-vector X, say $Y = \mathbf{A}X$, where the components of X are nonnegative and uncorrelated. Problems of this nature are motivated by a wide range of applications in which physical considerations deny the possibility that X can have negative components. Our approach to inference is founded on a necessary and sufficient condition for the existence of unique, nonnegative-score principal components. The condition replaces an earlier, sufficient constraint given in the engineering literature, and is related to a notion of sparsity that arises frequently in nonnegative principal component analysis. We discuss theoretical aspects of our estimators of the transformation that produces nonnegative-score principal components, showing that the estimators have optimal properties.

Key words and phrases: Correlation, independent component analysis, nonparametric statistics, permutation, principal component analysis, rate of convergence, rotation.

1. Introduction and Summary

Principal component analysis, or PCA, aims to explain the variance-covariance structure of a set of variables through a relatively small number of linear combinations of those quantities. The many applications of PCA, for purposes such as classification and data interpretation, are legion, but in some instances PCA is not especially successful. One of the difficulties is that the principal components do not necessarily conform to specific constraints, such as nonnegativity, expected from a physical interpretation of the underlying factors. In this paper we develop properties of a version of PCA under the assumption that the principal components, or PCs, are nonnegative random variables.

Given observations of a random p-vector Y with possibly correlated components, nonnegative-score PCA (NSPCA) can be described as an attempt to find an unobserved p-vector X with nonnegative, uncorrelated components, such that Y can be represented as a linear form in X: $Y = \mathbf{A}X$, where $\mathbf{A}$ is a $p \times p$

deterministic (unknown) matrix. (Throughout this paper, "nonnegative" means "nonnegative with probability 1.") Without loss of generality, the components of X all have variance 1. The components of X are termed nonnegative-score canonical components.

In general, such components may not exist, but it might be possible to choose $\mathbf{A}$ such that the components of $X = \mathbf{A}^{-1}Y$ are uncorrelated, have unit variance, and are not far from being nonnegative. At another extreme, nonnegative-score canonical components might exist but not be uniquely defined.

Next we discuss several practical problems which motivate such an approach to analysis. In these examples, we consider observed realisations $y_1, \ldots, y_n$ of the vector Y and corresponding unobserved realisations $x_1, \ldots, x_n$ of the vector X. The vectors are p-dimensional.

Example 1 (Combinations of images.)**.** A database of n monochromatic images (e.g., Lee and Seung (1999); Hoyer (2004)) can be represented as a set of p-dimensional vectors $y_1, \ldots, y_n$, where p denotes the number of pixels in each image. The columns of $\mathbf{A}$ can be interpreted as prototype images containing important features present in all images from a given database. Each unobserved realisation x_i represents encoding of the observed image $y_i = \mathbf{A}x_i$, and in particular, the jth component $x_i^{(j)}$ describes the "amount" of the jth prototype image present in y_i. In this example, it is natural to suppose that the components of $\mathbf{A}$ are nonnegative (see Section 4.1 for numerical methods), although in general that assumption might not be valid. If a small number of nonnegative-score PCs (NSPCs) explain most of the total variability, then the database can often be helpfully described using only those prototype images.

Example 2 (Determining prototype diseases.)**.** A database of patients' medical records may contain information about a set of p symptoms. Each p-vector y_i describes the "intensity" of each of the p symptoms for the ith patient. Each column of $\mathbf{A}$ represents a configuration of symptoms that may characterise a disease (we call it a prototype disease). An unobserved x_i gives the intensity of a prototype disease in the ith patient. Uncorrelatedness of NSPCs means that the intensity of one prototype disease in a patient is not correlated with intensity of another prototype disease. Such lack of correlation is not likely to occur between conventional diseases, which is why the NSPCs are interpreted as prototype diseases with conventional diseases described in terms of combinations of prototype diseases. This approach can aid the understanding of relationships among diseases.

Example 3 (Hyperspectral image analysis.)**.** In a hyperspectral image, each pixel contains a representation of a spectrum (or spectral curve) of reflectance (or

radiance), digitised as a p-vector y_i. Each component of y_i represents reflectance in a narrow spectral band. Spatial information is often ignored in such (often low-resolution) images (e.g., Manolakis and Shaw (2002)), and the pixelated data can be interpreted as realisations $y_1, \ldots, y_n$ of a random vector Y. In the formula $Y = \mathbf{A}X$, each column of $\mathbf{A}$ might represent a spectral profile of a feature (or component) present in the image. The dimension p is often larger than 200, but there is always a high level of correlation among the reflectance values in different spectral bands. Often, classic PCA is used to reduce dimensionality, but the resulting columns of $\mathbf{A}$ cannot be interpreted as spectral profiles because of their negative values.

Further examples of linear mixing of the original sources are given by Oja and Plumbley (2004), (separation of mixed images), and Plumbley (2003) (separation of musical audio signals). The setting of compositional data analysis, where negatively correlated components of X are required, is related, but is usually handled by other means; see, for example, Aitchison (1982, 1983, 1986).

Thus, NSPCA and the associated concept of explaining variability in high dimensional data in terms of a much smaller set of variables, have potential to play an important role in a wide range of applications where nonnegativity is desirable. Our methodology allows (near) positivity to be imposed on the matrix $\mathbf{A}$ as well as on X; this is useful in many applications but is not usually possible in classic PCA.

There are two major contemporary approaches to analysis of data from the model $Y = \mathbf{A}X$, under the assumption of positive X. One is nonnegative independent component analysis (ICA); see Plumbley (2002, 2003), Oja and Plumbley (2004) and Plumbley and Oja (2004). ICA requires independence of components of X, whereas we assume only uncorrelatedness, a natural assumption when introducing a basis. Hence, our approach is both canonical and more general. The second approach is nonnegative matrix factorisation (NMF); see Donoho and Stodden (2003), Hoyer (2004) and Lee and Seung (1999). NMF offers a particularly useful set of numerical algorithms, but arguably without a clear underlying statistical model and usually with highly non-unique solutions. Some other approaches (see Han (2010), Sigg and Buhmann (2008) and Zass and Shashua (2006)), called nonnegative PCA, impose the nonnegativity constraint on the coefficients (loadings) of PCs. Those approaches do not require the components of X to be nonnegative and uncorrelated. In order to distinguish our approach from such nonnegative PCA, we call our approach NSPCA. Related work includes that of Ma et al. (2012), who construct two-dimensional nonnegative PCs. This approach is specific to imaging data since it uses the two-dimensional structure of images. Comparisons among the various approaches are made in Section 4.1.

Motivated by examples such as those above, and by a desire for a method that has greater statistical underpinning than NMF but without the relatively stringent assumptions of ICA, we make the following contributions in this paper.

(1) We show that, for a given distribution of a random vector Y, nonnegative-score canonical components exist and are unique, up to permutation of coordinates, if and only if there exists a vector of nonnegative-score canonical components satisfying a certain wedged-in condition. This restriction relaxes a more stringent assumption of Plumbley (2002) based on the notions of independence and "well groundedness." (See Section 2.2.)

(2) Given data $Y_1, \ldots, Y_n$ on the p-vector Y, we propose estimators of the $p \times p$ matrix $\mathbf{A}$ and of $X_i = \mathbf{A}^{-1} Y_i$. We then derive optimal convergence rates of general estimators, and we show that these rates are achieved by the suggested methods. Since $\mathbf{A}$ and X_i are determined only up to multiplication on the right and left, respectively, by a permutation matrix, the notion of convergence rate has a nonstandard form. (See Sections 2.3 and 3.)

2. Modelling and Estimating Nonnegative-Score Principal Components

2.1. Definition and elementary properties of nonnegative-score principal components

Assume that $Y = \mathbf{A}X$, where $\mathbf{A} = (a_{jk})$ is a $p \times p$ deterministic (unknown) matrix, $X = (X^{(1)}, \ldots, X^{(p)})^{\mathrm{T}}$ is an unobserved vector with nonnegative components, and the covariance matrix $\mathrm{cov}(X)$ of X is the identity matrix $\mathbf{I}$. We refer to the $X^{(j)}$'s as nonnegative-score canonical components. For $1 \le k \le p$, let $\lambda_k = \sum_j a_{jk}^2$ be the square of the length of the kth column of $\mathbf{A}$, and put $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Let $W = (W^{(1)}, \ldots, W^{(p)})^{\mathrm{T}} = \mathbf{\Lambda}^{1/2} X$, that is, $W^{(j)} = \lambda_j^{1/2} X^{(j)}$. Then $W^{(j)}$ is the nonnegative-score principal component (NSPC) corresponding to the nonnegative canonical component $X^{(j)}$. Clearly, $\mathrm{var}(W^{(j)}) = \lambda_j$. If the components of X are ordered such that $\lambda_1 \ge \ldots \ge \lambda_p$, then $W^{(j)}$ is termed the jth NSPC.

In this notation, $Y = \mathbf{A}\mathbf{\Lambda}^{-1/2} W = \mathbf{D}W$, where the columns of the matrix $\mathbf{D} = \mathbf{A}\mathbf{\Lambda}^{-1/2}$ have unit length. Of course, the NSPCs can be expressed as linear combinations of the observed variables $Y^{(j)}$ through the relationship $W = \mathbf{D}^{-1} Y$.

The following two results, direct analogues of their counterparts for standard PCA, are readily proved:

$$\sum_{j=1}^{p} \mathrm{var}\big(Y^{(j)}\big) = \sum_{j=1}^{p} \mathrm{var}\big(W^{(j)}\big)\,, \quad \mathrm{corr}\big(W^{(j)}, Y^{(k)}\big) = \frac{d_{jk}\, \lambda_j^{1/2}}{\{\mathrm{var}(Y^{(k)})\}^{1/2}}\,, \qquad (2.1)$$

the latter for $1 \le j, k \le p$, where d_{jk} denotes the (j,k)th component of $\mathbf{D}$. The first result in (2.1) asserts that the total variability in Y equals the total variability in W. Note that left multiplication of Y by an orthogonal matrix does not change the covariance structure of a vector of PCs (both classic and nonnegative), but it alters the directions of PCs as described by eigenvectors (for classic PCA) or by rows of the matrix $\mathbf{D}^{-1}$ (for NSPC).

Consider an illustrative example shedding some light on NSPCA and its relationship to classic PCA. Let Z be a random vector with its support in the nonnegative quadrant $\mathcal{Q}$, and such that $\mathrm{cov}(Z)$ is the identity matrix $\mathbf{I}$. Let $\mathbf{F}$ be a diagonal matrix of positive elements on the diagonal, and $\mathbf{P}$ an orthogonal matrix. Take $Y = \mathbf{PF}Z$. The classic PCs of Y are given by the vector $\mathbf{F}Z$, which is also a vector of NSPCs in this case (as one of possibly many solutions). If Z is wedged in $\mathcal{Q}$, then Z is a unique (up to a permutation) vector of nonnegative-score canonical components, and $\mathbf{F}Z$ is a unique vector of NSPCs. If the support of Z is far from the boundary of $\mathcal{Q}$, then the set of solutions for nonnegative-score canonical components may be quite large. Within that set of solutions, we may find a rotation $\mathbf{B}$ such that $\mathbf{B}^{\mathrm{T}}Z$ is still nonnegative, and $\mathbf{PFB}$ is also nonnegative, or close to being nonnegative. Note that $\mathbf{B}^{\mathrm{T}}Z$ would be a vector of nonnegative-score canonical components, and $W = \mathbf{FB}^{\mathrm{T}}Z$ would be a vector of NSPCs. This would allow a nonnegative representation $Y = \mathbf{D}W$, where $\mathbf{D} = \mathbf{PFBF}^{-1}$. An advantage of this representation is the nonnegativity of $\mathbf{D}$ (and W), but a disadvantage is that the first several components of W do not necessarily maximise variability as do classic PCs.

2.2. Uniqueness of the NSPCA solution

Here we state and discuss the wedged-in condition and show it to be a necessary and sufficient condition for uniqueness of the NSPCA solution.

Let $\mathcal{Q} = \{(x^{(1)}, \ldots, x^{(p)})^{\mathrm{T}} : x^{(j)} \ge 0 \text{ for } 1 \le j \le p\}$ denote the positive orthant of $\mathbb{R}^p$. Let $X = (X^{(1)}, \ldots, X^{(p)})^{\mathrm{T}}$ be a random p-vector with support $\mathcal{S} \subseteq \mathbb{R}^p$. We say that X is nonnegative (equivalently, nonnegative with probability 1) if $\mathcal{S} \subseteq \mathcal{Q}$.

Definition. *Pure Rotation.* A p-dimensional orthogonal transformation that is not a permutation of coordinates is called a pure rotation. The identity is included as a pure rotation.

Definition. *"Wedged in" condition.* We say that $\mathcal{S}$ is wedged in $\mathcal{Q}$ if $\mathcal{S} \subseteq \mathcal{Q}$ and any pure rotation (that is not the identity) of $\mathcal{S}$ takes at least one point of $\mathcal{S}$ outside $\mathcal{Q}$. We extend the definition of "wedged in" from $\mathcal{S}$ to (the distribution of) X by asserting that X is wedged in $\mathcal{Q}$ if and only if $\mathcal{S}$ is wedged in $\mathcal{Q}$.

The above definition of the wedged-in condition might sometimes be difficult to check in practice. In the following, we provide simple examples where the

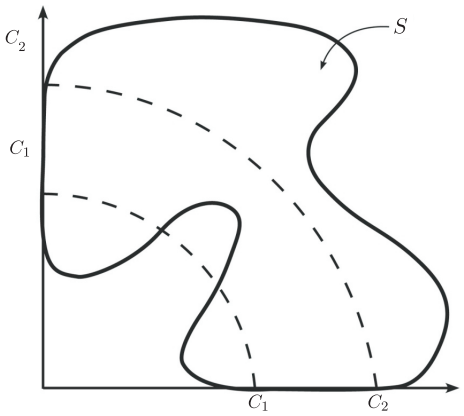

Figure 1. Example of configuration of the support, S, of the distribution of X, and the constants C_1 and C_2, when $p = 2$.

support $\mathcal{S}$ is wedged in $\mathcal{Q}$. Then we give an easier-to-check sufficient assumption (that we call well-groundedness) for the wedged-in condition to hold.

For example, if $\mathcal{S}$ is a subset of $\mathcal{Q}$ and the intersection of the closure of $\mathcal{S}$ with the positive half of any axis is nonempty, then $\mathcal{S}$ is wedged in $\mathcal{Q}$. However, this condition is not necessary for $\mathcal{S}$ to be wedged in. In particular, the set comprised of the "interior" of the hyperbola in $\mathcal{Q}$ with equation $y = 1/x$ is wedged in $\mathcal{Q}$, even though the hyperbola is not touching any of the two axes. Figure 1 gives, in the case $p = 2$, an example of a set $\mathcal{S}$ that is wedged in $\mathcal{Q}$ because any nontrivial rotation will necessarily move $\mathcal{S}$ outside the positive quadrant.

The term "wedged in" can be even better appreciated in three dimensions. Here one possibility for $\mathcal{S}$ to be wedged in $\mathcal{Q}$ is that the closure of $\mathcal{S}$ intersects with all three positive half axes. This means that $\mathcal{S}$ is "stuck" in the positive orthant $\mathcal{Q}$, in the sense that it cannot be rotated so that it stays within that orthant. On the other hand, a sphere of radius 1 with its center at the point $(1, 1, 1)$ is not wedged in $\mathcal{Q}$ because its rotation around the vector $(1, 1, 1)$ gives the same sphere, and hence it is not taken outside $\mathcal{Q}$. Note that this sphere touches each of the three plane boundaries of the positive orthant $\mathcal{Q}$ at only one point. Assume now that $\mathcal{S} \subseteq \mathcal{Q}$ touches (intersects) each of the three plane boundaries at a larger area such as a rectangle. In that case, $\mathcal{S}$ would be wedged in $\mathcal{Q}$. There are many other ways in which $\mathcal{S}$ could be wedged in $\mathcal{Q}$. A full characterisation of all those ways is beyond the scope of this paper though we give one sufficient condition for it in Proposition 1 at the end of this section.

The following theorem provides a necessary and sufficient condition for uniqueness of the NSPCA solution. The proof is deferred to Section 5.

Theorem 1. *Assume that for a given distribution of a random vector Y, a vector X of nonnegative-score canonical components exists. Then the vector X is unique, up to a permutation of coordinates, if and only if X is wedged in $\mathcal{Q}$.*

By way of comparison, classic canonical components in PCA are unique only up to a rotation. It can also be shown that the NSPCs are unique when all the λ_ks are distinct and the vector X is unique, up to a permutation of coordinates.

We now revisit the examples in Section 1 to discuss sparsity and the wedged-in condition. Recall that sparsity of a vector or matrix means that most of its elements are equal to zero. If, for some realisations x_i, all but one or two elements are equal to zero (or close to zero), then a wedged-in-$\mathcal{Q}$ distribution of X is suggested.

Example 1 (Combinations of images)**.** Lee and Seung (1999) advocate the assumption of sparsity of the unobserved x_i vectors, which is equivalent to having at least some of the observed images as linear combinations of a small number of prototype images. This type of sparsity helps create prototype images with localised features that correspond better to intuitive notions of parts of faces.

Example 2 (Determining prototype diseases)**.** In a medical database, sparsity of the unobserved x_i vectors means that a given patient usually does not suffer from many prototype diseases. Assuming that at least some patients do not have more than one or two prototype diseases is consistent with the wedged-in-$\mathcal{Q}$ condition.

Example 3 (Hyperspectral image analysis)**.** In this context, sparsity means that in at least some pixels, the number of features (or components) is quite small. This would be the case for uncomplicated scenes such as images of non-urban areas, or images with a relatively high resolution.

In order to formulate sufficient and easy-to-check assumptions for the wedged-in-$\mathcal{Q}$ condition, we provide the following definitions.

Definition. *"Well grounded" condition for a random variable.* A random variable $X^{(j)}$ (component of X) is well grounded if $P(0 \le X^{(j)} < \delta) > 0$ for each $\delta > 0$.

Definition. *"Well grounded" condition for a random vector.* A random vector $X = (X^{(1)}, \ldots, X^{(p)})$ is well grounded if there exists $\delta > 0$ such that, for each $\epsilon > 0$,

$$P\Big(X^{(j)} > \delta \text{ and, for all } k \neq j,\, 0 \le X^{(k)} < \epsilon\Big) > 0 \quad \text{for} \quad 1 \le j \le p\,.$$

The above definition may seem complex, but it is easy to check once the support of the distribution of X has been specified.

In the context of the components of X being mutually independent (as assumed by Plumbley (2002)), well-groundedness is a sufficient condition for the wedged-in condition. That is, if X is nonnegative, and if the components of X are mutually independent and well-grounded, then X is wedged in $\mathcal{Q}$.

The definition of well-groundedness for a random vector allows the following weaker characterisation of the wedged-in condition without the assumption of independence.

Proposition 1. *A sufficient condition that $X = (X^{(1)}, \ldots, X^{(p)})^{\mathrm{T}}$ be wedged in $\mathcal{Q}$ is that it be nonnegative and well grounded.*

2.3. Estimation of A and X from low-noise data

Here we provide a mathematical formulation of a class of problems of practical interest; see Section 1. Further, we suggest an algorithm for solving the problems.

Suppose we observe independent random vectors $Y_1, \ldots, Y_n$, each distributed as $Y = \mathbf{A}X$, where $X = (X^{(1)}, \ldots, X^{(p)})^{\mathrm{T}}$ is nonnegative, is wedged in $\mathcal{Q}$, and has identity covariance matrix, and where $\mathbf{A}$ is a nonsingular $p \times p$ deterministic (unknown) matrix. We refer to this as the low-noise case, arguing that small amounts of noise in observations of Y can be neglected without changing the method for inference. When noise is not negligible, neither $\mathbf{A}$ nor X is identifiable, even up to permutations.

We wish to estimate $\mathbf{A}$ and compute approximations to the random vectors $X_1, \ldots, X_n$ given by $Y_i = \mathbf{A}X_i$, for $1 \le i \le n$. We assume that $p < n$. Necessarily the approximations will be accurate only up to permutations of the coordinates of the X_i's, and likewise, $\mathbf{A}$ will be identifiable only up to multiplication on the right by a permutation matrix.

We suggest a four-step approach to defining estimators.

(i) Compute

$$\widehat{\mathbf{\Sigma}}_Y = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})\,(Y_i - \bar{Y})^{\mathrm{T}}\,, \tag{2.2}$$

the conventional estimator of $\mathbf{\Sigma}_Y = \operatorname{cov} Y$.

(ii) Given an orthogonal matrix $\mathbf{B}$, put

$$Z_i = \bigl(Z_i^{(1)}, \ldots, Z_i^{(p)}\bigr)^{\mathrm{T}} = Z_i(\mathbf{B}) = \mathbf{B}\,\widehat{\mathbf{\Sigma}}_Y^{-1/2}\,Y_i\,.$$

In the ICA literature, this is known as pre-whitening. Note that the observation vectors Z_i are such that their sample covariance matrix is the identity matrix. The orthogonal matrix $\mathbf{B}$ is needed for the purpose of rotation into the positive orthant $\mathcal{Q}$ in the next step.

(iii) Take

$$S^*(\mathbf{B}) = \max_{1\le j\le p} \max_{1\le i\le n} (-Z_i^{(j)}), \quad \widehat{\mathbf{B}} = \operatorname{argmin} S^*(\mathbf{B}), \tag{2.3}$$

where $\mathbf{B}$ is an orthogonal matrix. Define the negativity score

$$S(\mathbf{B}) = \max\{S^*(\mathbf{B}), 0\}. \tag{2.4}$$

In particular, $S(\mathbf{B}) = 0$ if and only if each $Z_i^{(j)} \ge 0$, and otherwise $S(\mathbf{B}) > 0$; $\widehat{\mathbf{B}}$ is a transformation for which the most negative value of $Z_i^{(j)} = Z_i^{(j)}(\mathbf{B})$ achieves its least negative value.

Since an orthogonal $\mathbf{B}$ can be represented as a composition of a permutation and a pure rotation, the minimisation to find $\widehat{\mathbf{B}}$ in (2.3) can be performed over pure rotations $\mathbf{P}$ parameterised as $\mathbf{P} = \exp(\mathbf{L})$ (Chevalley (1946)), where $\mathbf{L}$ is a skew-symmetric matrix ($\mathbf{L}^{\mathrm{T}} = -\mathbf{L}$). More specifically, we move along a geodesic $\mathbf{P}(t) = e^{t\mathbf{L}}\mathbf{P}_0$. A similar method is used by Plumbley (2003); see also Edelman, Arias, and Smith (1998, eq. (2.14)). Take $J(t) = S^*(e^{t\mathbf{L}}\mathbf{P}_0)$ as the negativity score in the direction of $\mathbf{L}$. This negativity score is not always differentiable, hence the following optimization procedure is used.

Step 1. Start with $\mathbf{P}_0$ as the identity matrix $\mathbf{I}$.

Step 2. Find $\mathbf{P}(t) = e^{tL}\mathbf{P}_0$ minimising $J(t)$ for $t \ge 0$ ($\mathbf{L}$ is defined below) through a line search.

Step 3. Use the optimum pure rotation $\mathbf{P}(t)$ found in Step 2 as a new value for $\mathbf{P}_0$.

Step 4. Repeat Steps 2 and 3 until convergence, or use another stopping rule.

The direction matrix $\mathbf{L}$ is chosen as described below, so as to ensure reduction of $J(t)$ in the vicinity of $t = 0+$. This is not necessarily the direction of steepest descent.

Let $\mathbf{V}$ be a $p \times n$ matrix consisting of vectors $(-\mathbf{P}_0\,\widehat{\mathbf{\Sigma}}_Y^{-1/2}\,Y_i)$ as columns, where $1 \le i \le n$. Then $J(t)$ is the maximum element of the matrix $e^{t\mathbf{L}}\,\mathbf{V}$. First assume that the maximum element of $\mathbf{V}$ is attained at exactly one entry (i_0, j_0) in that matrix. Then $J(t)$ is differentiable at zero, with $(dJ/dt)|_{t=0} = H_{i_0 j_0}$, where $H_{i_0 j_0}$ is the (i_0, j_0) entry in the matrix $\mathbf{H} = \mathbf{L}\mathbf{V}$. In this case, we can use the steepest descent direction $\mathbf{L} = \mathbf{Q} - \mathbf{Q}^{\mathrm{T}}$, where $\mathbf{Q}$ is a $p \times p$ matrix of zeros except for the i_0th column equalling the j_0th column of $\mathbf{V}$.

Let us now assume that the maximum element of $\mathbf{V}$ is attained at exactly k matrix elements in k different rows. Take $\mathbf{V}^*$ to be a $p \times k$ matrix of columns from $\mathbf{V}$ in which those maximum elements are located, and consider the case where $k = p$. For simplicity of notation we assume that the maximum elements of $\mathbf{V}^*$

are on the diagonal (otherwise use a permutation of coordinates). Partition $\mathbf{V}^*$ as

$$\mathbf{V}^* = \begin{bmatrix} a & b^{\mathrm{T}} \\ c & K \end{bmatrix},$$

where a is the maximum element of $\mathbf{V}^*$, b and c are $(p-1)$-dimensional vectors, and $\mathbf{L}$ is a $(p-1)\times(p-1)$ matrix.

With this notation it can be shown that the direction $\mathbf{L}$ that reduces all points of maximum at the same rate is given by

$$\mathbf{L} = \begin{bmatrix} 0 & -x^{\mathrm{T}} \\ x & \mathbf{M} \end{bmatrix},$$

where $\mathbf{M}$ is a $(p-1)\times(p-1)$ matrix with entries $M_{ij} = g_j\, K_{ij} - g_i\, K_{ji}$, vector $g = (g_1, \ldots, g_{(p-1)})$ is calculated as $c^{\mathrm{T}}\, \mathbf{E}^{-1}$, $\mathbf{E} = \mathbf{F} + \mathrm{diag}(b)$ is a $(p-1)\times(p-1)$ matrix, $\mathbf{F}$ is a $(p-1)\times(p-1)$ matrix with all rows being the same (equal to c^{T}) and $\mathrm{diag}(b)$ is a diagonal matrix with the vector b on the diagonal. The vector x is calculated as $\mathbf{E}^{-1} f$, where f is a $(p-1)$-dimensional vector with components $f_i = \sum_{1\le j\le p-1} M_{ij}\, K_{ji}$.

When $k < p$ we use the same partitioning of $\mathbf{V}^*$ and $\mathbf{L}$, but $\mathbf{L}$ is now a $(p-1)\times(k-1)$ matrix and b is a $(k-1)$-dimensional vector. Here x needs to be partitioned into $x = (x_{(1)}, x_{(2)})$, where $x_{(1)}$ is a $(k-1)$-dimensional sub-vector, and $x_{(2)}$ is $(p-k)$-dimensional. The vector c is partitioned in a similar way as $c = (c_{(1)}, c_{(2)})$. The matrix $\mathbf{M}$ is now defined as follows: $M_{ij} = g_j\, K_{ij} - g_i\, K_{ji}$ for $1 \le i, j \le k-1$; $M_{ij} = -g_i\, K_{ji}$ for $1 \le i \le k-1$, $k \le j \le p-1$; $M_{ij} = g_j K_{ij}$ for $k \le i \le p-1$, $1 \le j \le k-1$; and $M_{ij} = 0$ for $k \le i, j \le p-1$.

A new matrix $\mathbf{E}_{(1)}$ is defined similarly to $\mathbf{E}$ with c replaced by $c_{(1)}$, and $\mathbf{E}_{(2)}$ is a $(k-1)\times(p-k)$ matrix with all rows being the same (equal to $c_{(2)}^{\mathrm{T}}$). The vector f^* is a $(k-1)$-dimensional vector defined in the same way as f. Finally, the components of x are calculated as:

$$x_{(1)} = \mathbf{E}_{(1)}^{-1}(f - \mathbf{E}_{(2)} x_{(2)}), \quad x_{(2)} = c_{(2)}^{\mathrm{T}} - c_{(1)}^{\mathrm{T}} \mathbf{E}_{(1)}^{-1} \mathbf{E}_{(2)}.$$

In our experience, this minimisation process typically converges to a negativity score (defined in (2.4)) equal to zero, where the minimisation can be terminated (since a nonnegative solution has been found). Alternatively, the process stops when for one of the i values, say i_0, $(-Z_{i_0}^{(j)}) = S^*(\mathbf{B}) > 0$ for all $j = 1, \ldots, p$. In that case, we know that we have achieved a local minimum, which is typically also a global minimum for the type of data considered here. In order to see why, denote by a the value $S^*(\mathbf{B}) > 0$ at which the minimisation stopped. That is, $(-Z_{i_0})$ is a vector with all coordinates equal to a. Consider a rotation $\mathbf{P}$ of $(-Z_{i_0})$ that would be needed to further reduce all of its coordinates. Let the vector $Q = \mathbf{P}(-Z_{i_0})$ be such that all of its coordinates $Q^{(j)}$,

$j = 1, \ldots, p$, are smaller than a. If all coordinates $Q^{(j)}$ are also larger than $(-a)$, then $\|Q\| < \|Z_{i_0}\|$, which contradicts the assumption of $\mathbf{P}$ being a rotation (which preserves the vector norm). This means that for at least one j, say j_0, $Q^{(j_0)} \leq (-a)$. Hence, the rotation $\mathbf{P}$ results in the j_0 coordinate of $(-Z_{i_0})$ being transformed from $a > 0$ into a number not larger than $(-a)$. This proves that the original value $S^*(\mathbf{B})$ cannot be reduced by a local "small" rotation (in fact, the value always goes up after a small rotation), so this value is a local minimum. This still leaves the possibility of reaching a global minimum by the "larger" rotation $\mathbf{P}$ discussed above. However, for the type of data considered in this paper, that does not happen in practice. To see why, note that a is typically a small positive value and most points $(-Z_i)$ are in the negative orthant (or Z_i are in the positive orthant $\mathcal{Q}$). Consider a vector R with all coordinates equal to $(-r)$ (that is, positioned centrally in the negative orthant), with r significantly larger than a (as are typical coordinates of most points $(-Z_i)$). Under the rotation $\mathbf{P}$ discussed above, the vector R would be transformed into a vector with at least one coordinate at least as large as r (since R and $(-Z_{i_0})$ are collinear, but going in opposite directions). This will typically move the points $(-Z_i)$ significantly outside the negative orthant, resulting in an increase in the negativity score. Of course, one can find an artificial scenario with points $(-Z_i)$, where the rotation may indeed reduce the negativity score, but we have not encountered this in practice. If this were to occur, one could use a random search with a randomly generated L matrix defining the rotation.

It should be noted that the above described minimisation process can get stuck at some points, probably due to some other types of local minima or due to flatness of the optimized function. In those cases, a random rotation is used to continue the search, or an entirely different starting point is used.

Further work can be done to improve the minimisation process by modifications of the above method, or by some other methods. This would not impact any of the theoretical results in our paper, which hold for any solution to the minimisation (2.3).

The value of $\widehat{\mathbf{B}}$, defined in (2.3), may not be uniquely defined. For example, this is generally the case when $S(\widehat{\mathbf{B}}) = 0$, where there can be a continuum of transformations $\mathbf{B}$ such that none of the components $Z_i^{(j)}$ is negative. If a practitioner finds this non-uniqueness troubling, it can be overcome by taking $\widehat{\mathbf{B}}$ to be the transformation that maximises $\sum_i \sum_j Z_i^{(j)}(\mathbf{B})$ over orthogonal transformations $\mathbf{B}$ for which $S(\mathbf{B})$ achieves its maximum. In cases where $\widehat{\mathbf{B}}$ is not defined uniquely, the results in Section 3 apply to any version of it.

(iv) Take $\hat{\mathbf{A}} = \widehat{\mathbf{\Sigma}}_Y^{1/2}\widehat{\mathbf{B}}^{\mathrm{T}}$ to be our estimator of $\mathbf{A}$, and $\widehat{X}_i^{\mathrm{perm}} = \widehat{\mathbf{B}}\widehat{\mathbf{\Sigma}}_Y^{-1/2}Y_i$ to be our approximation to X_i. The superscript "perm" indicates that the components

of X_i that are approximated by $\widehat{X}_i^{\rm perm}$ may be permutations of those of X_i, the same permutation applying for each i.

To motivate this methodology, let $\mathbf{\Sigma}_Y = \mathbf{A}\mathbf{A}^{\rm T}$ denote the covariance matrix of Y. If $Z = \mathbf{B}\mathbf{\Sigma}_Y^{-1/2}Y$, where $\mathbf{B}$ is a $p \times p$ matrix, then, equivalently, $Z = \mathbf{B}\mathbf{\Sigma}_Y^{-1/2}\mathbf{A}X$, and so $\operatorname{cov} Z = \mathbf{B}\mathbf{B}^{\rm T}$. It now follows from Lemma 1 in Section 4.2 that, if $\mathbf{B}$ is chosen so that $\operatorname{cov} Z = \mathbf{I}$ and Z is nonnegative, then

$$\mathbf{U} \equiv \mathbf{B}\,\mathbf{\Sigma}_Y^{-1/2}\,\mathbf{A} \tag{2.5}$$

must be a permutation matrix. Moreover, $\mathbf{B}\mathbf{B}^{\rm T} = I$ and so $\mathbf{B}$ is orthogonal.

If we can find a $\mathbf{B} = \mathbf{B}_0$ such that $X^{\rm perm} \equiv \mathbf{B}_0\mathbf{\Sigma}_Y^{-1/2}Y$ is nonnegative and satisfies $\operatorname{cov} X^{\rm perm} = I$, then $\mathbf{B}_0$ is orthogonal and from (2.5),

$$X^{\rm perm} = \mathbf{B}_0\,\mathbf{\Sigma}_Y^{-1/2}\,\mathbf{\Sigma}_Y^{1/2}\,\mathbf{B}_0^{-1}\mathbf{U}\,X = \mathbf{U}\,X\,. \tag{2.6}$$

In particular, the p-vector $X^{\rm perm}$ is obtained by permuting the components of X. The procedure suggested in points (i)–(iv) above amounts to replacing $\mathbf{\Sigma}_Y$ by $\widehat{\mathbf{\Sigma}}_Y$, then choosing $\mathbf{B} = \widehat{\mathbf{B}}$ to make $\mathbf{B}\widehat{\mathbf{\Sigma}}_Y^{-1/2}Y_i$ "as positive as possible" in the sense of the score-function $S(\mathbf{B})$, and finally, taking $\hat{\mathbf{A}} = \widehat{\mathbf{\Sigma}}_Y^{1/2}\widehat{\mathbf{B}}^{\rm T} = \widehat{\mathbf{\Sigma}}_Y^{1/2}\widehat{\mathbf{B}}^{-1}$ (in place of $\mathbf{A}_1 = \mathbf{\Sigma}_Y^{1/2}\mathbf{B}^{-1}$) and $\widehat{X}_i^{\rm perm} = \widehat{\mathbf{B}}\,\widehat{\mathbf{\Sigma}}_Y^{-1/2}Y_i$ (in place of $X_i^{\rm perm} = \mathbf{B}\,\mathbf{\Sigma}_Y^{-1/2}Y_i$) to be our estimator of $\mathbf{A}$ and approximation to $X_i^{\rm perm}$, respectively.

3. Properties of Solution in Low-Noise Case

In this section, we give theory for the estimator $\hat{\mathbf{A}}$, and for the approximations $\widehat{X}_i^{\rm perm}$, suggested in Section 2.3. Throughout we keep the number of dimensions, p, fixed and take the sample size, n, to diverge. Our main result, Theorem 2, describes how accurately our method estimates the true matrix $\mathbf{A}$ and approximates the nonnegative-score canonical components given by X. The theorem can be paraphrased by stating that, if the density of X at points distant u from the boundary of $\mathcal{Q}$ is larger than a constant times u^{c-1} as $u \downarrow 0$, then, for an appropriate permutation matrix $\mathbf{U}$, the rate of convergence of $\hat{\mathbf{A}}\mathbf{U}$ to $\mathbf{A}$, and of $\widehat{X}_i^{\rm perm}$ to $\mathbf{U}X_i$, is $\max(n^{-1/c}, n^{-1/2})$ as $n \to \infty$. Moreover, this rate of convergence is optimal in a minimax sense; see Theorem 3.

One difficulty in formally phrasing such results is that of defining concisely the condition on the density stated in the previous paragraph. Writing $C_1 < C_2$ for positive constants, and $\Pi_1, \ldots, \Pi_p$ for the portions of $(p-1)$-dimensional planes that form the boundary of $\mathcal{Q}$, we require that the density be evaluated at a point that is not only distant u from one of $\Pi_1, \ldots, \Pi_p$, but is also not too close to the remaining $(p-1)$-dimensional planes and is distant between C_1 and C_2 from the origin. The result fails without the restriction that the point lies between C_1 and C_2.

One might consider, for instance, the case where $C_1 = 0$ and $\mathcal{S}$ intersects the boundary of $\mathcal{Q}$ only at the origin. There it is possible to construct examples where X is uniformly distributed on its support, in which case the density is either 0, or a positive constant value, at each point in $\mathbb{R}^p$; and the convergence rate is dictated completely by the behaviour of the support boundary in the neighbourhood of the origin. Our regularity conditions, given in the next paragraph, exclude such cases from specific treatment; this greatly simplifies our discussion. In particular, we ask that the support of the distribution of X have a significant presence in the region defined by $C_1 \le \|X\| \le C_2$ and $X \in \mathcal{Q}$, including parts of this region close to the coordinate planes. Figure 1 illustrates a configuration in the case $p = 2$.

Against this background, we define classes $\mathcal{A}$ of matrices $\mathbf{A}$ and $\mathcal{F}$ of distributions F. Let $0 < C_1 < C_2 < \infty$ and $C_3, C_4, c, \epsilon, \delta > 0$, such that $\delta < \frac{1}{2}\, p^{-1/2}\, C_1$. Write $\mathcal{F} = \mathcal{F}(C_1, \ldots, C_4, c, \epsilon, \delta)$ for the class of distributions F of X such that, for each $1 \le j \le p$, the density of the distribution is at least $C_3\, u^{c-1}$ at points $x \in \mathcal{Q}$ distant u perpendicularly from Π_j, provided that $C_1 \le \|x\| \le C_2$ and x is not closer than δ to all remaining Π_k's, where $k \ne j$. (Here, $\|x\|$ denotes the Euclidean length of the p-vector x.) We also assume that $E(\|X\|^{4+\epsilon}) \le C_4$, which ensures uniform integrability of fourth moments of distributions of X in $\mathcal{F}$; this condition could instead be imposed directly. Fourth moments are needed because we require root-n consistency of $\widehat{\mathbf{\Sigma}}_Y$.

Given $0 < C_5 < C_6 < \infty$, let $\mathcal{A} = \mathcal{A}(C_5, C_6)$ denote the class of $p \times p$ matrices $\mathbf{A}$ for which all the eigenvalues of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ lie in the interval $[C_5, C_6]$. If $\mathbf{M}$ is a $p \times p$ matrix, write $\|\mathbf{M}\|$ for any conventional norm of $\mathbf{M}$ for which $\|\mathbf{M}x\| \le \|\mathbf{M}\|\, \|x\|$ for a vector x, where $\|\mathbf{M}x\|$ and $\|x\|$ are the standard Euclidean norms. The square root of the sum of the squares of the components of $\mathbf{M}$, its Frobenius norm, is one possibility. Take $\delta_n = \max(n^{-1/c}, n^{-1/2})$.

Theorem 2. *There exists a sequence of permutation matrices* $\mathbf{U}_n$*, depending on the data, such that, for any measurable versions of* $\hat{\mathbf{A}}$ *and* $\widehat{X}_i^{\mathrm{perm}}$ *defined at* (iv) *of Section* 2.3*,*

$$\limsup_{n\to\infty} \sup_{\mathbf{A}\in\mathcal{A}} \sup_{F\in\mathcal{F}} P_F\big(\big\|\hat{\mathbf{A}}\,\mathbf{U}_n - \mathbf{A}\big\| > C\,\delta_n\big) \to 0\,, \tag{3.1}$$

$$\limsup_{n\to\infty} \sup_{\mathbf{A}\in\mathcal{A}} \sup_{F\in\mathcal{F}} P_F\big(\big\|\widehat{X}_i^{\mathrm{perm}} - \mathbf{U}_n X_i\big\| > C\,\delta_n\,\|X_i\|\big) \to 0 \tag{3.2}$$

as $C \to \infty$.

Results (3.1) and (3.2) should be compared with (2.5) and (2.6), respectively. They give rates of convergence of $\hat{\mathbf{A}}$ to $\mathbf{A}$, and of $\widehat{X}_i^{\mathrm{perm}}$ to X_i. However, since $\mathbf{A}$ and X_i are uniquely defined only up to multiplication by permutation matrices,

the "rate of convergence" has to be interpreted appropriately. Note that the same permutation, $\mathbf{U}_n$, is used in both (3.1) and (3.2).

The next theorem provides a lower bound of the same order as the upper bound given by Theorem 2. It applies to all choices of $\mathcal{F} = \mathcal{F}(C_1, \ldots, C_4, c, \epsilon, \delta)$ such that $0 < \delta < C_1(4p)^{-1/2}$, and $\mathcal{A} = \mathcal{A}(C_5, C_6)$ for which $0 < C_1 < C_2 < \infty$, $0 < C_5 < C_6 < \infty$ and $C_3, C_4, c, \epsilon > 0$, with C_3 chosen sufficiently small and C_4 sufficiently large as functions of the other constants. Let $\mathcal{U}_n$ denote the class of all random permutation matrices that are measurable functions of the data $Y_1, \ldots, Y_n$.

Theorem 3. *Let $\tilde{\mathbf{A}}$ denote any estimator of $\mathbf{A}$. Then there exists a constant $C > 0$ such that*

$$\limsup_{n\to\infty} \sup_{\mathbf{A}\in\mathcal{A}} \sup_{F\in\mathcal{F}} \inf_{\mathbf{U}\in\mathcal{U}_n} P_F(\|\tilde{\mathbf{A}}\,\mathbf{U} - \mathbf{A}\| > C\,\delta_n) > 0\,. \tag{3.3}$$

To interpret (3.3), note that the result holds trivially if $\mathbf{U}$ is a poor choice for the approximation "$\tilde{\mathbf{A}}\,\mathbf{U} \approx \mathbf{A}$." The theorem shows that, in a sense made precise by (3.3), no choice of $\mathbf{U}$ can render the approximation more accurate than $O(\delta_n)$.

In principle, Theorem 3 leaves open the possibility that, although $\mathbf{A}$ cannot be estimated at a faster rate than δ_n, individual vectors X_i can be estimated at a faster rate. To see that this cannot happen, suppose it can, and in fact that we can compute $\widetilde{X}_1, \ldots, \widetilde{X}_n$ such that, for a sequence $u_n \downarrow 0$, and a random permutation matrix $\mathbf{U}$,

$$\sup_{1\le i\le n} P_F\big(\big\|\widetilde{X}_i - \mathbf{U}X_i\big\| \le u_n\,\delta_n\,\|X_i\|\big) \to 0\,. \tag{3.4}$$

Suppose too that for some $C_1 > 0$, $P(\|X\| \le C_1) = 1$, and that the matrix $\mathbf{A}$ for which $E\|Y - \mathbf{A}X\| = 0$ is unique. Let us use a least-squares approach to estimate $\mathbf{A}$, choosing $\mathbf{B} = \widetilde{\mathbf{B}}$ to minimise $\sum_i \|\mathbf{B}\widetilde{X}_i - Y_i\|^2$. This proposal is motivated by the fact that, if $\widetilde{X}_i$ is close to $\mathbf{U}X_i$, as suggested by (3.4), then $\mathbf{B}\widetilde{X}_i - Y_i$ is close (on average) to $(\mathbf{BU} - \mathbf{A})X_i$, and therefore if $\mathbf{B}\widetilde{X}_i - Y_i$ is small then $\mathbf{BU}$ is close to $\mathbf{A}$. Indeed, for each $C_2 > 0$, $P(\|\widetilde{\mathbf{B}}\mathbf{U} - A\| \le C_2\delta_n) \to 1$, where $\mathbf{U}$ is as at (3.4). This contradicts (3.3). The proofs of Theorems 2 and 3 can be found in Hall, Bajorski, and Rubinstein (2009).

4. Numerical Results

4.1. Comparison to other methods

Our methodology of NSPCA is now compared to four related methods discussed in Section 1. Recall the assumption used in this paper that the observed

vector Y is a linear transformation of a nonnegative W having uncorrelated components, that is, $Y = \mathbf{D}W$, where W is the vector of NSPCs. Hence, our goal is to estimate $\mathbf{D}$ and approximate values of W rather than to try to find linear transformations of Y with maximum variability, which is the goal of some of the other approaches to nonnegative PCA. Based on notation of Section 2.1, the matrix $\mathbf{D}$ is estimated by $\hat{\mathbf{D}} = \hat{\mathbf{A}}\hat{\mathbf{\Lambda}}^{-\mathbf{1/2}}$, where $\hat{\mathbf{A}}$ is defined in Step (iv) of the algorithm presented in Section 2.3, and $\hat{\mathbf{\Lambda}}$ is the diagonal matrix with $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$ being the squares of the lengths of the columns of $\hat{\mathbf{A}}$. Similarly, $\hat{W} = \hat{\mathbf{\Lambda}}^{1/2}\hat{X}$, where $\hat{X}$ is the approximation of X defined in Step (iv) of the algorithm presented in Section 2.3.

The first method being compared here is nonnegative ICA (NICA), which is the most similar to NSPCA. It uses a different negativity score (based on the sums of squares) and more stringent assumptions of independence and "well groundedness" as discussed in Section 1.

The second method, expectation-maximisation NPCA (EMNPCA), is based on maximisation of variability of linear combinations with positive coefficients. These nonnegative PCs are constructed sequentially for a total of p PCs. Rather than rely on perfect orthogonality of the loadings vectors, a quadratic penalty term (see formula (20) in Sigg and Buhmann (2008)) is introduced, resulting in quasi-orthogonality of the loadings vectors. The obtained nonnegative PCs are typically correlated. Note that even orthogonality of the loadings vectors does not guarantee uncorrelatedness of nonnegative PCs in EMNPCA.

The third method we evaluate is nonnegative semi-disjoint PCA (NSDPCA), introduced in Zass and Shashua (2006). This method is also based on maximisation of variability of linear combinations with positive coefficients. However, it identifies a whole set of $k \leq p$ nonnegative PCs, rather than constructing them in a sequential fashion. With $k < p$, one could not estimate the matrix $\mathbf{D}$, which would make it more difficult to compare NSDPCA to the other methods. Hence, we used $k = p$. Using smaller k would result in nonnegative PCs more similar to those produced by EMNPCA. We have also made comparisons to classic PCA.

For each of the five methods, we have evaluated how close the estimated $\hat{\mathbf{D}}$ is to the true $\mathbf{D}$ and how close the approximation vectors $\hat{W}_i$ (scores of PCs) are to the simulated W_i's based on the standardized root mean squared errors of approximation as defined by

$$D_{\text{diff}} = \frac{\|\hat{\mathbf{D}} - \mathbf{D}\|}{\|\mathbf{D}\|}, \quad W_{\text{diff}} = \frac{\|\hat{\mathbf{W}} - \mathbf{W}\|}{\|\mathbf{W}\|},$$

where $\mathbf{W}$ [$\hat{\mathbf{W}}$] is an n by p matrix of W_i's [$\hat{W}_i$'s] as rows.

The evaluation was performed based on three cases of simulated data. In each case, we took $p = 10$ and $n = 1,000$. The components of X were generated

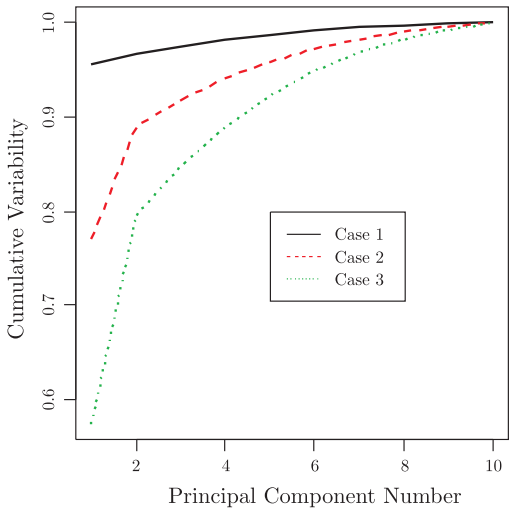

Figure 2. Cumulative variability of NSPCs in the three cases considered in Section 4.1.

Table 1. The approximation errors (as measured by D_{diff}) for the five methods and three cases considered in Section 4.1.

Case	NSPCA	NICA	EMNPCA	NSDPCA	PCA
1	0.2362	0.2360	0.3624	0.3598	0.7713
2	0.0547	0.0552	1.0538	0.4688	0.6716
3	0.0861	0.0859	0.9423	0.4755	0.6634

as independent from the same beta distribution with parameters 2 and 4, scaled to have variance 1. The matrices $\mathbf{D}$ were chosen as different in each case and such that the resulting percentages of variability explained by $W^{(j)}$'s (as described by the $\mathbf{\Lambda}$ matrix) were as shown in Figure 2. For example, variances of $W^{(1)}$ were 96, 77 and 57 percent of the total variability for Cases 1, 2 and 3, respectively.

The resulting estimation errors D_{diff} are shown in Table 1. We can see that NSPCA and NICA are performing very similarly to each other, with NSPCA sometimes slightly better and sometimes slightly worse than NICA. In Cases 2 and 3, both NSPCA and NICA are able to recover the matrix $\mathbf{D}$ quite precisely. The remaining three methods are significantly worse at recovering $\mathbf{D}$. This is not surprising since they are not optimized for the recovery of the nonnegative uncorrelated components. Instead, they try to find components with maximum variability. In Case 1, both NSPCA and NICA are not performing as well as in the other two cases. This might be due to the fact that most λ_i's are very close to each other (except for λ_1). Similar behaviour is well known in classic PCA.

Table 2. The estimation errors (as measured by D_{diff}) for the five methods and three cases considered in Section 4.1.

Case	NSPCA	NICA	EMNPCA	NSDPCA	PCA
1	0.581	0.589	2.342	2.307	8.740
2	0.343	0.330	13.437	9.029	5.879
3	0.400	0.407	11.605	7.626	5.305

The approximation errors W_{diff} are shown in Table 2. We can again see that NSPCA and NICA are performing very similarly to each other, with NSPCA sometimes slightly better and sometimes slightly worse than NICA. The remaining three methods are significantly worse at recovering $\mathbf{W}$. The very large approximation errors are due to the fact that recovered PCs are entirely different from the true NSPCs (since the three methods have assumptions very different from those of NSPCA).

It is also of interest to investigate the percent of variability explained by PCs produced by the five methods. The results are represented in Figures 3, 4 and 5, showing the cumulative percent of variability in Cases 1, 2 and 3, respectively. In each figure, the solid line represents the true cumulative variability of the underlying NSPCs as assumed in the simulations. The NICA produces results very similar to those of NSPCA (the lines would overlap) so, for clarity of presentation, the NICA results are not shown. In Case 1 (Figure 3), all methods produce very high cumulative variabilities that are close to the true values. Interestingly enough, the EMNPCA and NSDPCA lines overlap the true values almost perfectly, despite the fact that they estimate different PCs as demonstrated in previous considerations summarized in Tables 1 and 2.

In Case 2 (Figure 4), NSPCA recovers the approximately correct amounts of the cumulative variability, while the other methods either overestimate or underestimate that variability. As expected, EMNPCA (which sequentially maximises variability) explains more variability than NSPCA. If the main goal were to maximise variability, then EMNPCA would be a preferred method. One could also increase the variability explained by NSDPCA, but that would require a construction (and hence maximisation of variability) of a smaller number of PCs. In our approach, the main goal is to reconstruct the underlying signals $W^{(j)}$ rather than maximise the variability.

Case 3 (Figure 5) is again an example where NSPCA is best in estimating the true variability, while the other methods either overestimate or underestimate that variability.

The three simulation examples used in this section demonstrate good performance of NSPCA in recovering the underlying mixing matrix $\mathbf{D}$ and nonnegative uncorrelated signals in $W^{(j)}$ assumed by the model presented in Section

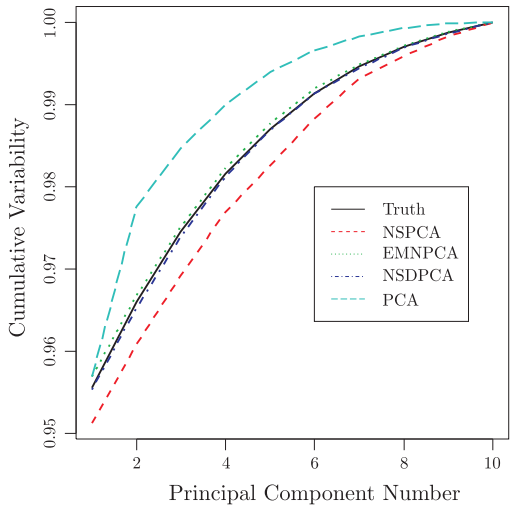

Figure 3. Cumulative percent of variability explained by PCs produced by various methods in Case 1 considered in Section 4.1.

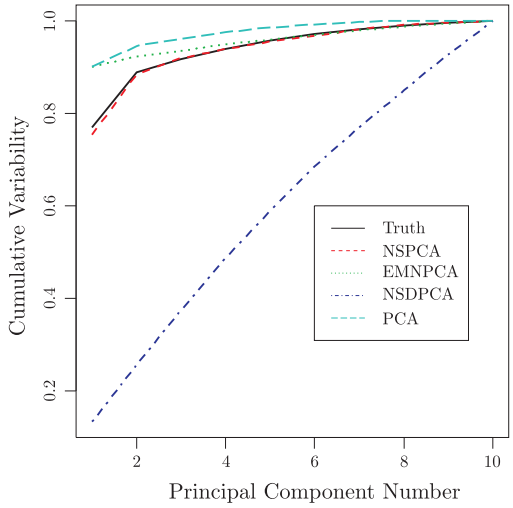

Figure 4. Cumulative percent of variability explained by PCs produced by various methods in Case 2 considered in Section 4.1.

2.1 (and used throughout the paper). The competing NICA methodology provides computationally similar results, but it uses more stringent assumptions of independence and "well groundedness" (see Sections 1 and 2.2). The remaining competing methodologies (EMNPCA and NSDPCA) discussed here are designed for different purposes (of variability maximisation), and hence, they do not per-

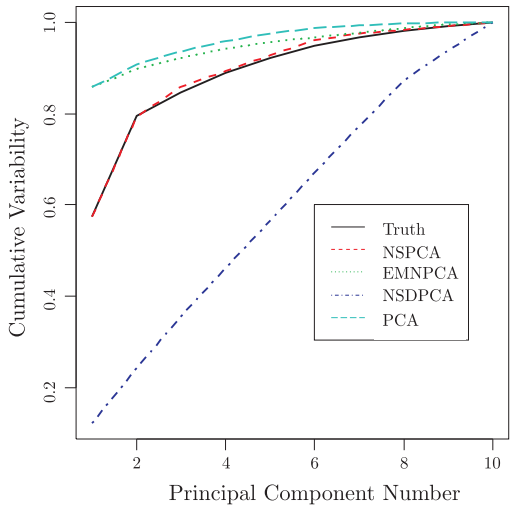

Figure 5. Cumulative percent of variability explained by PCs produced by various methods in Case 3 considered in Section 4.1.

form well in recovering the underlying mixing matrix $\mathbf{D}$ and signals in $W^{(j)}$'s.

4.2. Simulation study

Further simulations were performed to see how precisely NSPCA can recover the underlying model in final samples. The approach and notation here is consistent with that of Theorem 2. Recall that Theorem 2 shows the asymptotic behaviour of our estimators when the sample size approaches infinity. In simulations, we assumed that the components $X^{(1)}, \ldots, X^{(p)}$ were independent and followed the same power-function distribution, with distribution function $(x/\gamma)^c$ for $0 \le x \le \gamma$, where $c > 0$ and the scale parameter $\gamma = (2c^{-1}+1)^{1/2}(c+1)$ was chosen to obtain variance 1. In this case, the joint distribution of X belonged to the class $\mathcal{F}$ defined in Section 3.

Figure 6 shows simulated distributions of the random variables

$$G = \frac{\|\hat{\mathbf{A}}\,\mathbf{U} - \mathbf{A}\|}{\delta_n \|\mathbf{A}\|}, \quad H = \frac{\|\widehat{X}_i^{\text{perm}} - \mathbf{U}X_i\|}{\delta_n \|X_i\|},$$

where $\delta_n = \max(n^{-1/c}, n^{-1/2})$ is the rate of convergence given by Theorem 2. All simulations were repeated 4,000 times. Since the distribution of H does not depend on i, then all observations (for $1 \le i \le n$) from all repetitions were combined in the figures. Hence, the results for the distribution of H are, in effect, summaries of $4,000n$ observations.

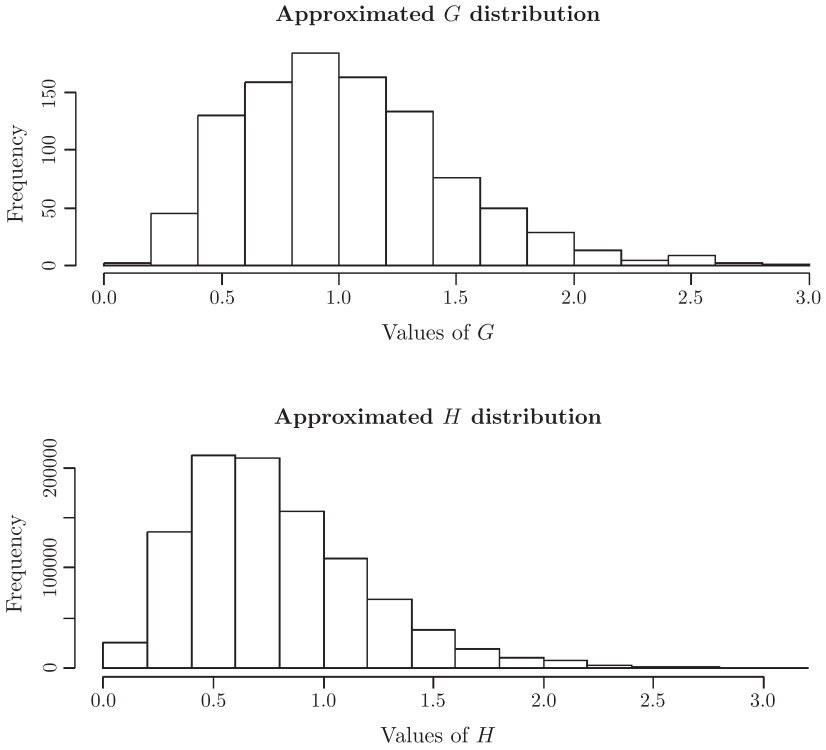

Figure 6. Histograms of the simulated G and H distributions for $(p, c, n) = (3, 1, 10^3)$, based on simulations discussed in Section 4.2.

Figure 6 shows approximations to the distributions of G and H in the case $(p, c, n) = (3, 1, 10^3)$. Similar plots for varying values of p, c, and n show almost identical shapes of distributions.

In the setting of Figure 6, the 95th percentile of the simulated G distribution is about 1.85, which is equivalent to 6% error. When n is reduced to 100, the error increases to 20%, and for $n = 10^4$, the error decreases to 2%. This reduction in the error of estimation with increasing n is also observed for other values of p and c, as expected.

We also investigated the changes in the estimation error as a function of c for p ranging from 2 to 5. Figure 7 shows the errors in estimation of $\mathbf{A}$ and X (with $p = 5$) as described by the 95th percentiles of the simulated distributions of $\|\hat{\mathbf{A}}\,\mathbf{U} - \mathbf{A}\|/\|\mathbf{A}\|$ and $\|\widehat{X}_i^{\mathrm{perm}} - \mathbf{U}X_i\|/\|X_i\|$, respectively. The error usually decreases when c changes from 1 to 2, and then increases when c increases to 3 and 4. Similar behaviour is also observed for $p = 2$, 3, and 4.

Note that when varying p, the fixed matrix $\mathbf{A}$ also changes, so a direct comparison of errors across different values of p could be misleading.

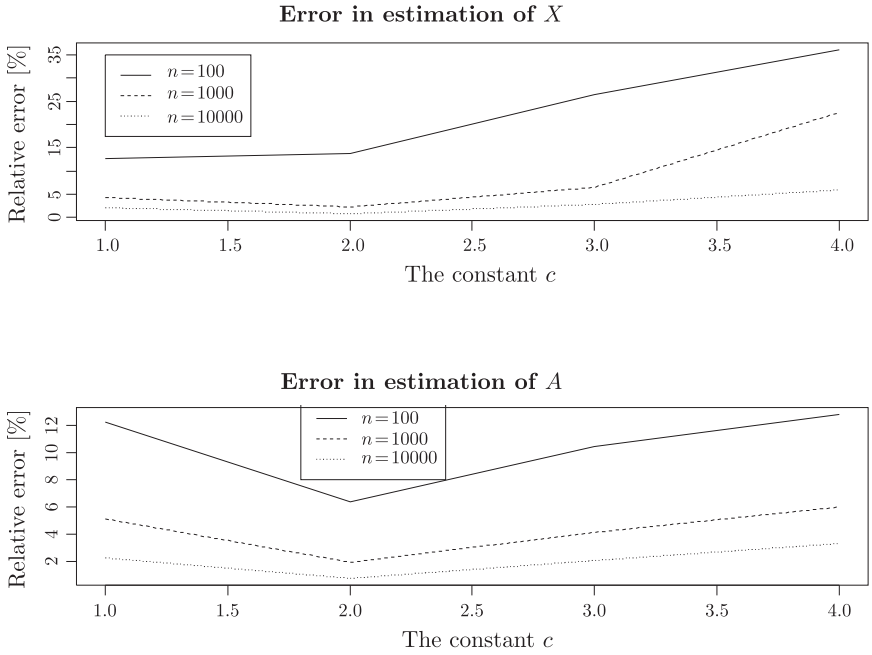

Figure 7. Errors in the estimation of A and X as defined by the 95th percentiles of the simulated distributions of $\|\hat{\mathbf{A}}\,\mathbf{U} - \mathbf{A}\|/\|\mathbf{A}\|$ and $\|\widehat{X}_i^{\text{perm}} - \mathbf{U}X_i\|/\|X_i\|$, respectively, based on simulations for $p = 5$ and sample sizes of $n = 10^2$, 10^3 and 10^4 as discussed in Section 4.2.

4.3. Example with hyperspectral imaging data

Here we used a hyperspectral image of an urban area in Rochester, NY, USA, near the Lake Ontario shoreline. The image was gathered using NASA's Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) technology. We selected the sub-image shown in Figure 8(b), consisting of n = 10,000 pixels, in order to reduce the computational burden. The same sub-image was used by Bajorski (2011a,b) and in the book by Bajorski (2012), where more information can be found. We used $p = 102$ spectral bands in the range from 400 to 1340nm in order to avoid the water absorption wavelengths. The numerical results reported here are generally close to those obtained using methods suggested by Plumbley (2002, 2003), Oja and Plumbley (2004) and Plumbley and Oja (2004), although as noted in earlier sections our assumptions and negativity score are different.

The NSPCs were estimated using the methodology described in this paper. The first columns of $\mathbf{D}$ (see section 2.1) representing profiles of the NSPCs (explaining 98.9 percent variability) are shown in Figure 9. The first NSPC can be interpreted as the infrared component of the image. The second and third NSPCs represent the visible spectrum (400 to 700 nm) with the second (third) NSPC

Figure 8. (a) AVIRIS image of Rochester, NY and Lake Ontario. (b) Actual 100×100 pixel image used in section 4.3.

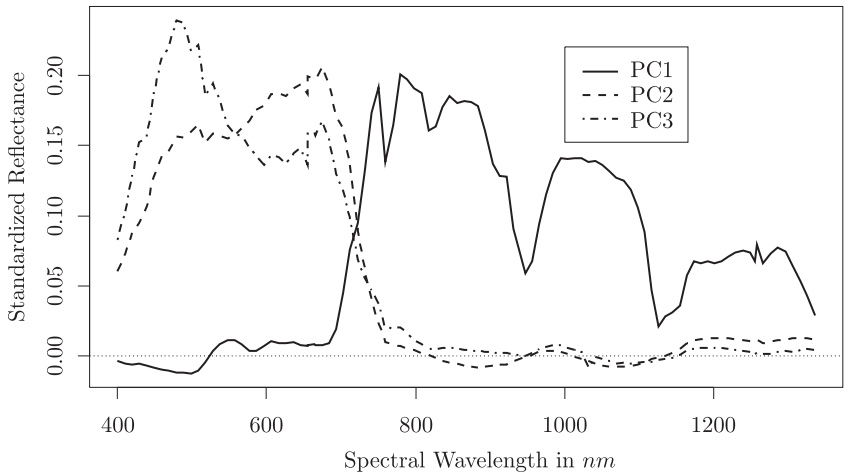

Figure 9. Profiles of the first three nonnegative PCs estimated from an AVIRIS hyperspectral image (see section 4.3).

representing more of the red (blue) wavelength. In this context, the classic PCs are more difficult to interpret because they represent contrasts between spectral bands, and their values are often negative.

4.4. Density data example

The density of photographic film was measured along a 21-step log exposure

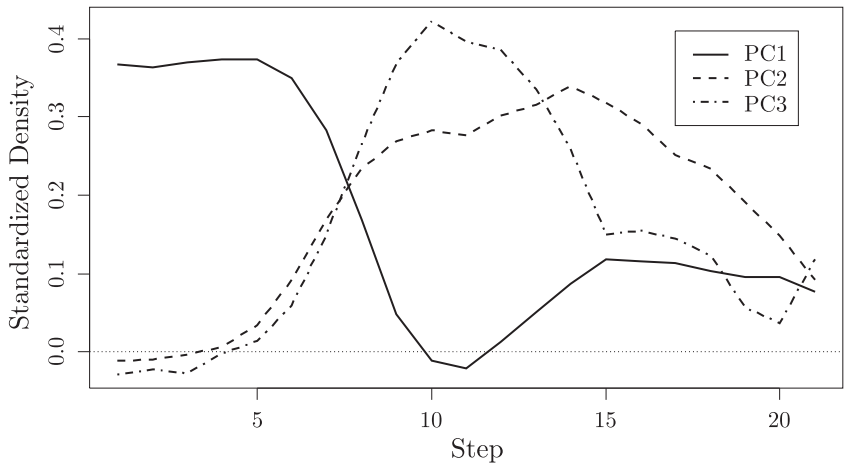

Figure 10. Profiles of the first three nonnegative PCs estimated from the density data (see section 4.4).

series. Each set of 21 response values is often plotted against an index and referred to as a Dloge curve (e.g., Farnell (1966); Hunt (1987, Sec. 14.14). The curve characteristically starts at a minimum density, increases with greater exposure, and then reaches some maximum density. The 32 observations in this dataset were generated from a 2^5 full factorial experiment, where five variables, each at two levels, were manipulated as design variables in the making of a photographic emulsion.

Although the actual dimension here is 21, the effective dimension is relatively low. Indeed, the data can be well approximated using only four to nine principal components, and so the nominal dimension of 21 is not intrinsic.

We again estimated the NSPCs using our methodology. The profiles (columns of $\mathbf{D}$) of the first three NSPCs (explaining 94.9 percent variability) are shown in Figure 10. The benefit of using the NSPCs is that they can be interpreted as additive effects, as opposed to contrasts usually available in classic PCA, explaining different aspects of the data. For example, the first NSPC can be interpreted as showing variability in the emulsion density, represented by the starting points of the Dloge curves. The second NSPC explains variability in the final stages of a Dloge curve; that is, the maximum density achieved in the process. The third NSPC is responsible for the variability in the middle range of the Dloge curves, where a steep incline is observed, interpreted by emulsion scientists as representing film contrast.

LaLonde and Bajorski (2006) used these NSPCs as response variables to model the log exposure data with the five design variables as predictors. They

also compared this method with two other approaches, multivariate linear regression with all 21 response variables and with principal components as predictors. The advantage of using NSPCs was ease of interpretation.

5. Appendices

5.1. Method for computing nonnegative A.

As indicated in Examples 1 and 2 in Section 1, we may sometimes require $\mathbf{A}$ to be a matrix of positive values. Since $\mathrm{cov}(Y) = \mathbf{A}\mathbf{A}^{\mathrm{T}}$, then the canonical representation $Y = \mathbf{A}X$, with positive $\mathbf{A}$, will only be possible when the elements of Y are positively correlated. In practice, small levels of negativity might be acceptable. However, initial estimates often give many negative elements in $\hat{\mathbf{A}}$, even for predominantly positively correlated data. Below we suggest a penalty-based method that minimises negativity of both $\hat{\mathbf{A}}$ and $\widehat{X}$.

Recall that, from a computational viewpoint, the matrix $\mathbf{B}$ is estimated by minimisation of a negativity score, say $S^*(\mathbf{B})$, for $\mathbf{Y}^*B^{\mathrm{T}}$, where $\mathbf{Y}^*$ is an $n \times p$ matrix of observations of $\widehat{\mathbf{\Sigma}}_Y^{-1/2}\, Y_i$. Since $\mathbf{A}$ is estimated by $\hat{\mathbf{A}} = \widehat{\mathbf{\Sigma}}_Y^{1/2}\widehat{\mathbf{B}}^{\mathrm{T}}$, where $\widehat{\mathbf{B}}$ is given at (2.3), we also want to minimise a negativity score, say $N(\mathbf{B})$, for $\widehat{\mathbf{\Sigma}}_Y^{1/2}\mathbf{B}^{\mathrm{T}}$. In order to balance both minimisations, we minimise $S^*(\mathbf{B}) + a\, N(\mathbf{B})$, where a is a chosen constant. This approach is especially useful when the solution to $Y = \mathbf{A}X$ is not unique and there is significant flexibility in choice of $\mathbf{B}$.

5.2. Wedged-in lemma

Lemma 1. *Assume that the vector X is wedged in $\mathcal{Q}$ and has identity covariance matrix. Let $\mathbf{H}$ be a square matrix such that $X_* = \mathbf{H}X$ is nonnegative and has identity covariance matrix. Then $\mathbf{H}$ is a permutation matrix.*

Proof. Since $I = \mathrm{cov}(X_*) = \mathbf{H}\mathbf{H}^{\mathrm{T}}$, $\mathbf{H}$ is an orthogonal transformation. Any orthogonal transformation can be expressed as a pure rotation followed by a permutation, that is, $\mathbf{H} = \mathbf{U}\mathbf{R}$ where $\mathbf{U}$ and $\mathbf{R}$ are permutation and pure rotation matrices, respectively. In particular, $\mathbf{R}X = \mathbf{U}^{\mathrm{T}}X_*$ is a permutation of X_*, and so is nonnegative. If $\mathbf{R}$ represents a pure rotation then $\mathbf{R}$ takes part of $\mathcal{S}$ outside $\mathcal{Q}$. This contradicts the fact that $\mathbf{R}X$ is nonnegative. Therefore $\mathbf{R}$ must be the identity, and so $\mathbf{H}$ must be a permutation matrix $\mathbf{U}$.

5.3. Proof of Theorem 1

Suppose we have two representations, say $Y = \mathbf{A}X$ and $Y = \mathbf{A}_*X_*$, of Y in terms of vectors X and X_* of nonnegative-score canonical components, and one of them, say X, is wedged in $\mathcal{Q}$. Then, $X_* = \mathbf{A}_*^{-1}\mathbf{A}X$. From Lemma 1, we deduce that $\mathbf{H} = \mathbf{A}_*^{-1}\mathbf{A}$ must be a permutation matrix $\mathbf{U}$. That is, $\mathbf{A}_* =$

$\mathbf{U}^{\mathrm{T}}\mathbf{A}$ and $X_* = \mathbf{U}X$, which proves the uniqueness of nonnegative-score canonical components up to the permutation of coordinates.

Conversely, suppose X is a unique random vector of nonnegative-score canonical components but is not wedged in $\mathcal{Q}$. It follows from the wedged-in condition that there exists a nondegenerate pure rotation matrix $\mathbf{R}$ such that $X_* = \mathbf{R}X$ is nonnegative. Then, $\mathrm{cov}(X_*) = \mathbf{R}\mathbf{R}^{\mathrm{T}} = I$. That is, X_* is also a vector of nonnegative-score canonical components, contradicting the uniqueness of X.

Acknowledgements

We are grateful to Professor Amnon Neeman for helpful conversations and to Professors Steven LaLonde and Joseph Voelkel for providing some of the data used in examples. We are also grateful to the journal editors and reviewers for helpful comments that guided us in improving this paper.

References

Aitchison, J. (1982). The statistical analysis of compositional data. (With discussion.) *J. Roy. Statist. Soc. Ser. B* **44**, 139-177.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika* **70**, 57-65.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman & Hall, London.

Bajorski, P. (2011a). Second Moment Linear Dimensionality as an Alternative to Virtual Dimensionality. *IEEE Trans. Geoscience and Remote Sensing.* **49**. 672-678.

Bajorski, P. (2011b). Statistical Inference in PCA for Hyperspectral Images. *IEEE J. Selected Topics in Signal Processing.* **5**, 438-445.

Bajorski, P. (2012). *Statistics for Imaging, Optics, and Photonics.* Wiley, New York.

Chevalley, C. (1946). *Theory of Lie Groups.* Princeton University Press, Princeton.

Donoho, D. and Stodden, V. (2003). When does Non-Negative Matrix Factorization give a correct Decomposition into Parts. *NIPS 2003 workshop on ICA: Sparse Representations in Signal Processing* Whistler, Canada.

Edelman, A., Arias, T. A. and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints, *SIAM J. Matrix Anal. Appl.* **20**, 303-353.

Farnell, G. C. (1966). The relationship between density and exposure. Chapter 4 in *The Theory of the Photographic Process*, (Edied by T. H. James), 3rd edition. Macmillan, New York.

Hall, P., Bajorski, P. and Rubinstein, H. (2009). Proofs for Theorems 2 and 3 of "Methodology and theory for nonnegative principal component analysis." Unpublished manuscript.

Han, X. (2010). Nonnegative Principal Component Analysis for Cancer Molecular Pattern Discovery. *IEEE/ACM Trans. Computational Biology and Bioinformatics* **7**, 537 - 549.

Hoyer, P. (2004). Non-negative matrix factorization with sparseness constraints. *J. Machine Learning Res.* **5**, 1457-1469.

Hunt, R. W. G. (1987). *The Reproduction of Color in Photography, Printing & Television.* Fountain Press, England.

LaLonde, S. M. and Bajorski, P. (2006). A comparison of three approaches to modeling a multivariate response in a designed experiment. In *Proceedings of the Joint Statistical Meetings, Section on Physical and Engineering Science*, 1740-1747.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788-791.

Ma, P., Yang, D., Ge, Y., Zhang, X. and Qu, Y. (2012). Face recognition using two-dimensional nonnegative principal component analysis. *J. Electron. Imaging* **21**, 033011.

Manolakis, D. and Shaw, G. (2002). Detection algorithms for hyperspectral imaging applications. *IEEE Sig. Proc. Mag.* **19**, 29-43.

Oja, E., and Plumbley, M. D. (2004). Blind separation of positive sources by globally convergent gradient search. *Neural Comput.* **16**, 1811-1825.

Plumbley, M. D. (2002). Conditions for nonnegative independent component analysis. *IEEE Sig. Proc. Lett.* **9**, 177-180.

Plumbley, M. D. (2003). Algorithms for nonnegative independent component analysis. *IEEE Trans. Neural Networks* **14**, 534-543.

Plumbley, M. D. and Oja, E. (2004). A "nonnegative PCA" algorithm for independent component analysis. *IEEE Trans. Neural Networks* **15**, 66-76.

Sigg, C. D. and Buhmann, J. M. (2008). Expectation-Maximization for Sparse and Non-Negative PCA. *Proceedings of the 25-th International Conference on Machine Learning*, Helsinki, Finland.

Zass, R. and Shashua, A. (2006). Nonnegative Sparse PCA. *Advances in Neural Information and Processing Systems.* MIT Press.

Graduate Statistics Department, Rochester Institute of Technology, 98 Lomb Memorial Drive, Rochester, NY 14623–5604, USA.

E-mail: Peter.Bajorski@rit.edu

Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia.

E-mail: halpstat@ms.unimelb.edu.au

Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia.

E-mail: H.Rubinstein@ms.unimelb.edu.au

(Received October 2011; accepted September 2012)