Biostatistics (2009), **10**, 3, *pp.* 515–534
doi:10.1093/biostatistics/kxp008
Advance Access publication on April 17, 2009

A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis

DANIELA M. WITTEN*

Department of Statistics, Stanford University, Stanford, CA 94305, USA
dwitten@stanford.edu

ROBERT TIBSHIRANI

Department of Health Research & Policy and Department of Statistics, Stanford University, Stanford, CA 94305, USA

TREVOR HASTIE

Department of Statistics, Stanford University, Stanford, CA 94305, USA

SUMMARY

We present a penalized matrix decomposition (PMD), a new framework for computing a rank-K approximation for a matrix. We approximate the matrix $\mathbf{X}$ as $\hat{\mathbf{X}} = \sum_{k=1}^{K} d_k \mathbf{u}_k \mathbf{v}_k^T$, where d_k, $\mathbf{u}_k$, and $\mathbf{v}_k$ minimize the squared Frobenius norm of $\mathbf{X} - \hat{\mathbf{X}}$, subject to penalties on $\mathbf{u}_k$ and $\mathbf{v}_k$. This results in a regularized version of the singular value decomposition. Of particular interest is the use of L_1-penalties on $\mathbf{u}_k$ and $\mathbf{v}_k$, which yields a decomposition of $\mathbf{X}$ using sparse vectors. We show that when the PMD is applied using an L_1-penalty on $\mathbf{v}_k$ but not on $\mathbf{u}_k$, a method for sparse principal components results. In fact, this yields an efficient algorithm for the "SCoTLASS" proposal (Jolliffe *and others* 2003) for obtaining sparse principal components. This method is demonstrated on a publicly available gene expression data set. We also establish connections between the SCoTLASS method for sparse principal component analysis and the method of Zou *and others* (2006). In addition, we show that when the PMD is applied to a cross-products matrix, it results in a method for penalized canonical correlation analysis (CCA). We apply this penalized CCA method to simulated data and to a genomic data set consisting of gene expression and DNA copy number measurements on the same set of samples.

Keywords: Canonical correlation analysis; DNA copy number; Integrative genomic analysis; L_1; Matrix decomposition; Principal component analysis; Sparse principal component analysis; SVD.

*To whom correspondence should be addressed.

1. INTRODUCTION

Consider a matrix $\mathbf{X}$ with n rows and p columns. In this paper, we present a new method for computing a rank-K approximation for $\mathbf{X}$:

$$\hat{\mathbf{X}} = \sum_{k=1}^{K} d_k \mathbf{u}_k \mathbf{v}_k^T, \tag{1.1}$$

where $\mathbf{u}_k$ and $\mathbf{v}_k$ are unit vectors in $\mathbb{R}^n$ and $\mathbb{R}^p$, respectively, and d_k are nonnegative constants. We estimate $\mathbf{u}_k$ and $\mathbf{v}_k$ subject to a penalty on their elements; as a result, we call this the "penalized matrix decomposition" (PMD) of $\mathbf{X}$.

In this paper, we will show that this decomposition has many uses:

1. Applying PMD to a data matrix can yield interpretable factors that provide insight into the data.
2. Applying PMD to a data matrix with L_1-constraints on the columns but not the rows yields an efficient algorithm for the SCoTLASS method for finding sparse principal components. This is similar to a method of Shen and Huang (2008).
3. Applying PMD to a cross-product matrix yields a new method for penalized CCA.

The main area of application in this paper relates to (3) above. In recent years, it has become increasingly common for biologists to perform 2 different assays on the same set of samples. For instance, both gene expression and DNA copy number measurements often are available on a set of patient samples. In this situation, an integrative analysis of both the gene expression and the copy number data is desired. If $\mathbf{X}$ and $\mathbf{Y}$ are $n \times p$ and $n \times q$ matrices with standardized columns, then PMD applied to the matrix of cross-products $\mathbf{X}^T\mathbf{Y}$ results in an efficient method for performing penalized CCA. This method can be applied to gene expression and copy number data in order to identify sets of genes that are correlated with regions of copy number change. We will demonstrate the use of our penalized CCA method for this purpose on a publicly available breast cancer data set.

In Section 2, we present the PMD. We show that the PMD can be used to identify shared regions of gain and loss in simulated DNA copy number data. In Section 3, we use the PMD to arrive at an efficient algorithm for finding sparse principal components, and we use PMD to unify preexisting methods for sparse principal component analysis (PCA). In Section 4, we extend the PMD framework in order to develop a method for penalized CCA, and we demonstrate its use on a breast cancer data set consisting of both gene expression and DNA copy number measurements on the same set of patients. Section 5 contains the discussion.

2. THE PMD

2.1 *General form of PMD*

Let $\mathbf{X}$ denote an $n \times p$ matrix of data with rank $K \leqslant \min(n, p)$. Without loss of generality, assume that the overall mean of $\mathbf{X}$ is 0. The singular value decomposition (SVD) of the data can be written as follows:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}_n, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}_p, \quad d_1 \geqslant d_2 \geqslant \cdots \geqslant d_K > 0. \tag{2.1}$$

Let $\mathbf{u}_k$ denote column k of $\mathbf{U}$, let $\mathbf{v}_k$ denote column k of $\mathbf{V}$, and note that d_k denotes the kth diagonal element of the diagonal matrix $\mathbf{D}$. Then, it is a well-known fact (see e.g. Eckart and Young, 1936) that for any $r \leqslant K$,

$$\sum_{k=1}^{r} d_k \mathbf{u}_k \mathbf{v}_k^T = \arg\min_{\hat{\mathbf{X}} \in M(r)} \|\mathbf{X} - \hat{\mathbf{X}}\|_F^2, \tag{2.2}$$

where $M(r)$ is the set of rank-r $n \times p$ matrices and $\|\cdot\|_F^2$ indicates the squared Frobenius norm (the sum of squared elements of the matrix). In other words, the first r components of the SVD give the best rank-r approximation to a matrix, in the sense of the Frobenius norm.

In this paper, we develop generalizations of this decomposition by imposing additional constraints on the elements of $\mathbf{U}$ and $\mathbf{V}$. We start with a rank-1 approximation.

Consider the following optimization problem:

$$\text{minimize}_{d,\mathbf{u},\mathbf{v}} \frac{1}{2}\|\mathbf{X} - d\mathbf{u}\mathbf{v}^T\|_F^2 \quad \text{subject to } \|\mathbf{u}\|_2^2 = 1, \ \ \|\mathbf{v}\|_2^2 = 1, \ P_1(\mathbf{u}) \leqslant c_1, \ P_2(\mathbf{v}) \leqslant c_2, \ d \geqslant 0. \tag{2.3}$$

Here, P_1 and P_2 are convex penalty functions, which can take on a variety of forms. Useful examples are

- lasso: $P_1(\mathbf{u}) = \sum_{i=1}^n |u_i|$ and
- fused lasso: $P_1(\mathbf{u}) = \sum_{i=1}^n |u_i| + \lambda \sum_{i=2}^n |u_i - u_{i-1}|$, where $\lambda > 0$

Only certain ranges of c_1 and c_2 will lead to feasible solutions, as discussed in Section 2.3. (Note that throughout this paper, the notation $\|\mathbf{u}\|_p$ indicates the L_p-norm of the vector $\mathbf{u}$, i.e. $\left(\sum_i u_i^p\right)^{\frac{1}{p}}$.) We now derive a more convenient form for this criterion.

The following decomposition holds.

THEOREM 2.1 Let $\mathbf{U}$ and $\mathbf{V}$ be $n \times K$ and $p \times K$ orthogonal matrices and $\mathbf{D}$ a diagonal matrix with diagonal elements d_k. Then,

$$\frac{1}{2}\|\mathbf{X} - \mathbf{U}\mathbf{D}\mathbf{V}^T\|_F^2 = \frac{1}{2}\|\mathbf{X}\|_F^2 - \sum_{k=1}^K \mathbf{u}_k^T\mathbf{X}\mathbf{v}_k d_k + \frac{1}{2}\sum_{k=1}^K d_k^2. \tag{2.4}$$

The theorem's proof is given in the Appendix. Hence, using the case $K = 1$, we have that the values of $\mathbf{u}$ and $\mathbf{v}$ that solve (2.3) also solve the following problem:

$$\text{maximize}_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}\mathbf{v} \quad \text{subject to } \|\mathbf{u}\|_2^2 = 1, \ \ \|\mathbf{v}\|_2^2 = 1, \ P_1(\mathbf{u}) \leqslant c_1, \ P_2(\mathbf{v}) \leqslant c_2, \tag{2.5}$$

and the value of d solving (2.3) is $\mathbf{u}^T\mathbf{X}\mathbf{v}$. The objective function $\mathbf{u}^T\mathbf{X}\mathbf{v}$ in (2.5) is bilinear in $\mathbf{u}$ and $\mathbf{v}$: that is, with $\mathbf{u}$ fixed, it is linear in $\mathbf{v}$, and vice versa. In fact, with $\mathbf{v}$ fixed, the criterion in (2.5) takes the following form:

$$\text{maximize}_{\mathbf{u}} \mathbf{u}^T\mathbf{X}\mathbf{v} \quad \text{subject to } P_1(\mathbf{u}) \leqslant c_1, \ \ \|\mathbf{u}\|_2^2 = 1. \tag{2.6}$$

This criterion is not convex due to the L_2-equality penalty on $\mathbf{u}$.

We can finesse this as follows. We define the (rank-1) PMD by

$$\text{Rank-1 PMD:} \quad \text{maximize}_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}\mathbf{v} \quad \text{subject to } \|\mathbf{u}\|_2^2 \leqslant 1, \ \ \|\mathbf{v}\|_2^2 \leqslant 1, \ P_1(\mathbf{u}) \leqslant c_1, \ P_2(\mathbf{v}) \leqslant c_2. \tag{2.7}$$

With $\mathbf{v}$ fixed, this criterion takes the form

$$\text{maximize}_{\mathbf{u}} \mathbf{u}^T\mathbf{X}\mathbf{v} \quad \text{subject to } P_1(\mathbf{u}) \leqslant c_1, \ \ \|\mathbf{u}\|_2^2 \leqslant 1, \tag{2.8}$$

which is convex. This means that (2.7) is biconvex, and this suggests an iterative algorithm for optimizing it. Moreover, it turns out that the solution to (2.8) also satisfies $\|\mathbf{u}\|_2^2 = 1$, provided that c_1 is chosen so that (for fixed $\mathbf{v}$) the $\mathbf{u}$ that maximizes

$$\mathbf{u}^T\mathbf{X}\mathbf{v} \quad \text{subject to } P_1(\mathbf{u}) \leqslant c_1 \tag{2.9}$$

has L_2-norm greater than or equal to 1. This follows from the Karush–Kuhn–Tucker conditions in convex optimization (see, e.g. Boyd and Vandenberghe, 2004). Therefore, for c_1 chosen appropriately, the solution to (2.8) solves (2.6).

The following iterative algorithm is used to optimize the criterion for the rank-1 PMD.

Algorithm 1: Computation of single-factor PMD model

1. Initialize $\mathbf{v}$ to have L_2-norm 1.
2. Iterate until convergence:
   (a) $\mathbf{u} \leftarrow \arg\max_{\mathbf{u}} \mathbf{u}^T\mathbf{X}\mathbf{v}$ subject to $P_1(\mathbf{u}) \leqslant c_1$ and $\|\mathbf{u}\|_2^2 \leqslant 1$.
   (b) $\mathbf{v} \leftarrow \arg\max_{\mathbf{v}} \mathbf{u}^T\mathbf{X}\mathbf{v}$ subject to $P_2(\mathbf{v}) \leqslant c_2$ and $\|\mathbf{v}\|_2^2 \leqslant 1$.
3. $d \leftarrow \mathbf{u}^T\mathbf{X}\mathbf{v}$.

In Section 2.2, we present an algorithm for obtaining multiple-factor solutions for the PMD. When P_1 and P_2 both are L_1-penalties, maximizations in Steps 2(a) and 2(b) are simple. This is explained in Algorithm 3 in Section 2.3.

It can be seen that without the P_1- and P_2-constraints, the algorithm above leads to the usual rank-1 SVD. Starting with $\mathbf{v}^{(0)}$, one can show that at the end of iteration i,

$$\mathbf{v}^{(i)} = \frac{(\mathbf{X}^T\mathbf{X})^i\mathbf{v}^{(0)}}{\|(\mathbf{X}^T\mathbf{X})^i\mathbf{v}^{(0)}\|_2}. \tag{2.10}$$

This is the well-known "power method" for computing the largest eigenvector of $\mathbf{X}^T\mathbf{X}$, which is the leading singular vector of $\mathbf{X}$.

In practice, we suggest using the first right singular vector of $\mathbf{X}$ as the initial value $\mathbf{v}$. In general, Algorithm 1 does not necessarily converge to a global optimum for (2.7); however, our empirical studies indicate that the algorithm does converge to interpretable factors for appropriate choices of the penalty terms. Note that each iteration of Step 2 in Algorithm 1 results in a decrease in the objective in (2.7).

The PMD is similar to a method of Shen and Huang (2008) for identifying sparse principal components; we will elaborate on the relationship between the 2 methods in Section 3.

2.2 *PMD for multiple factors*

In order to obtain multiple factors of the PMD, we minimize the single-factor criterion (2.7) repeatedly, each time using as the $\mathbf{X}$ matrix the residuals obtained by subtracting from the data matrix the previous factors found. The algorithm is as follows.

Algorithm 2: Computation of K factors of PMD

1. Let $\mathbf{X}^1 \leftarrow \mathbf{X}$.
2. For $k \in 1, \ldots, K$:
   (a) Find $\mathbf{u}_k$, $\mathbf{v}_k$, and d_k by applying the single-factor PMD algorithm (Algorithm 1) to data $\mathbf{X}^k$.
   (b) $\mathbf{X}^{k+1} \leftarrow \mathbf{X}^k - d_k\mathbf{u}_k\mathbf{v}_k^T$.

Without the P_1- and P_2-penalty constraints, it can be shown that the K-factor PMD algorithm leads to the rank-K SVD of $\mathbf{X}$. In particular, the successive solutions are orthogonal. This can be seen since the solutions $\mathbf{u}_k$ and $\mathbf{v}_k$ are in the column and row spaces of $\mathbf{X}^k$, which has been orthogonalized with respect to $\mathbf{u}_j$, $\mathbf{v}_j$ for $j \in 1, \ldots, k-1$. With P_1 and/or P_2 present, the solutions are no longer in the column and row spaces, and so the orthogonality does not hold. In Section 3.2, we discuss an alternative multifactor approach, in the setting where PMD is specialized to sparse principal components.

2.3 *Forms of PMD of special interest*

We are most interested in 2 specific forms of the PMD, which we call the "PMD(L_1, L_1)" and "PMD(L_1, FL)" methods. The PMD(L_1, L_1) criterion is as follows:

$$\text{maximize}_{\mathbf{u},\mathbf{v}}\, \mathbf{u}^T\mathbf{X}\mathbf{v} \quad \text{subject to } \|\mathbf{u}\|_2^2 \leqslant 1,\ \ \|\mathbf{v}\|_2^2 \leqslant 1,\ \ \|\mathbf{u}\|_1 \leqslant c_1,\ \ \|\mathbf{v}\|_1 \leqslant c_2. \tag{2.11}$$

This method results in factors $\mathbf{u}$ and $\mathbf{v}$ that are sparse for c_1 and c_2 chosen appropriately. As shown in Figure 1, we restrict c_1 and c_2 to the ranges $1 \leqslant c_1 \leqslant \sqrt{n}$ and $1 \leqslant c_2 \leqslant \sqrt{p}$.

Let S denote the soft thresholding operator; that is, $S(a, c) = \text{sgn}(a)(|a| - c)_+$, where $c > 0$ is a constant and where x_+ is defined to equal x if $x > 0$ and 0 if $x \leqslant 0$. We have the following lemma.

LEMMA 2.2 Consider the optimization problem

$$\text{maximize}_{\mathbf{u}}\, \mathbf{u}^T\mathbf{a} \quad \text{subject to } \|\mathbf{u}\|_2^2 \leqslant 1,\ \ \|\mathbf{u}\|_1 \leqslant c. \tag{2.12}$$

The solution satisfies $\mathbf{u} = \frac{S(\mathbf{a},\Delta)}{\|S(\mathbf{a},\Delta)\|_2}$, with $\Delta = 0$ if this results in $\|\mathbf{u}\|_1 \leqslant c$; otherwise, Δ is chosen so that $\|\mathbf{u}\|_1 = c$.

The proof is given in the Appendix. We solve the PMD criterion in (2.11) using Algorithm 1, with Steps 2(a) and 2(b) adjusted as follows.

Algorithm 3: Computation of single-factor PMD(L_1, L_1) model

1. Initialize $\mathbf{v}$ to have L_2-norm 1.
2. Iterate until convergence:
   (a) $\mathbf{u} \leftarrow \frac{S(\mathbf{X}\mathbf{v},\Delta_1)}{\|S(\mathbf{X}\mathbf{v},\Delta_1)\|_2}$, where $\Delta_1 = 0$ if this results in $\|\mathbf{u}\|_1 \leqslant c_1$; otherwise, Δ_1 is chosen to be a positive constant such that $\|\mathbf{u}\|_1 = c_1$.

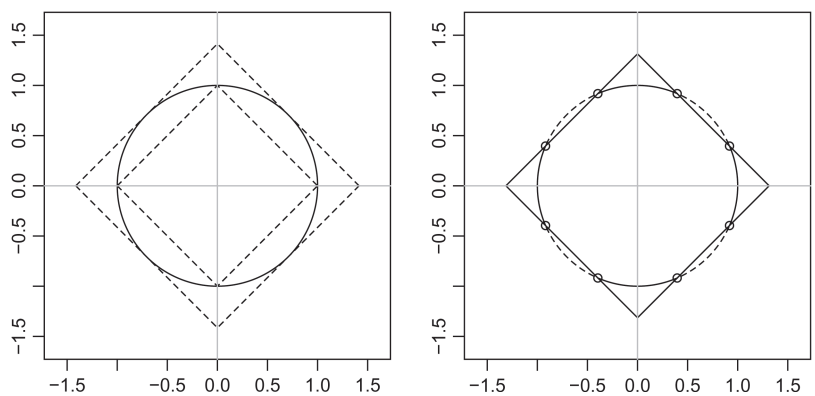

Fig. 1. A graphical representation of the L_1- and L_2-constraints on $\mathbf{u}$ in the PMD(L_1, L_1) criterion. The constraints are as follows: $\|\mathbf{u}\|_2^2 \leqslant 1$ and $\|\mathbf{u}\|_1 \leqslant c$. Here, $\mathbf{u}$ is two-dimensional, and the grey lines indicate the coordinate axes u_1 and u_2. Left: the L_2-constraint is the solid circle. For both the L_1- and L_2-constraints to be active, c must be between 1 and $\sqrt{2}$. The constraints $\|\mathbf{u}\|_1 = 1$ and $\|\mathbf{u}\|_1 = \sqrt{2}$ are shown using dashed lines. Right: The L_2- and L_1-constraints on $\mathbf{u}$ are shown for some c between 1 and $\sqrt{2}$. Small circles indicate the points where both the L_1- and the L_2-constraints are active. The solid arcs indicate the solutions that occur when $\Delta_1 = 0$ in Algorithm 3. The figure shows that in 2D, the points where both the L_1- and L_2-constraints are active do not have either u_1 or u_2 equal to 0.

(b) $\mathbf{v} \leftarrow \frac{S(\mathbf{X}^T\mathbf{u}, \Delta_2)}{\|S(\mathbf{X}^T\mathbf{u}, \Delta_2)\|_2}$, where $\Delta_2 = 0$ if this results in $\|\mathbf{v}\|_1 \leqslant c_2$; otherwise, Δ_2 is chosen to be a positive constant such that $\|\mathbf{v}\|_1 = c_2$.

3. $d \leftarrow \mathbf{u}^T\mathbf{X}\mathbf{v}$.

If one desires that $\mathbf{u}$ and $\mathbf{v}$ be equally sparse, one can simply fix a constant c and set $c_1 = c\sqrt{n}, c_2 = c\sqrt{p}$. For each update of $\mathbf{u}$ and $\mathbf{v}$, Δ_1 and Δ_2 are chosen by a binary search.

Figure 1 shows a graphical representation of the L_1- and L_2-constraints on $\mathbf{u}$ that are present in the PMD(L_1, L_1) criterion: namely, $\|\mathbf{u}\|_2^2 \leqslant 1$ and $\|\mathbf{u}\|_1 \leqslant c_1$. From the figure, it is clear that in two dimensions, the intersection of the L_1- and L_2-constraints results in both u_1 and u_2 nonzero. However, when $n = 2$, the dimension of $\mathbf{u}$, is at least 3, then the right panel of Figure 1 can be thought of as the hyperplane $\{u_i = 0, \forall i > 2\}$. In this case, the small circles indicate regions where both constraints are active and the solution is sparse (since $u_i = 0$ for $i > 2$).

The PMD(L_1, FL) criterion is as follows (where "FL" stands for the "fused lasso" penalty, proposed in Tibshirani *and others*, 2005):

$$\text{maximize}_{\mathbf{u},\mathbf{v}}\mathbf{u}^T\mathbf{X}\mathbf{v} \quad \text{subject to } \|\mathbf{u}\|_2^2 \leqslant 1, \ \ \|\mathbf{u}\|_1 \leqslant c_1, \ \|\mathbf{v}\|_2^2 \leqslant 1, \ \sum_j |v_j| + \lambda \sum_j |v_j - v_{j-1}| \leqslant c_2. \tag{2.13}$$

This method results in $\mathbf{u}$ sparse and $\mathbf{v}$ sparse and somewhat smooth (depending on the value of $\lambda \geqslant 0$). However, for simplicity, rather than solving (2.13), we solve a slightly different criterion which results from using the Lagrange form, rather than the bound form, of the constraints on $\mathbf{v}$:

$$\text{minimize}_{\mathbf{u},\mathbf{v}} -\mathbf{u}^T\mathbf{X}\mathbf{v} + \frac{1}{2}\mathbf{v}^T\mathbf{v} + \lambda_1 \sum_j |v_j| + \lambda_2 \sum_j |v_j - v_{j-1}| \quad \text{subject to } \|\mathbf{u}\|_2^2 \leqslant 1, \ \ \|\mathbf{u}\|_1 \leqslant c. \tag{2.14}$$

We can solve this by replacing Steps 2(a) and 2(b) in Algorithm 1 with the appropriate updates:

Algorithm 4: Computation of single-factor PMD(L_1, FL) model

1. Initialize $\mathbf{v}$ to have L_2-norm 1.
2. Iterate until convergence:
   (a) $\mathbf{u} \leftarrow \frac{S(\mathbf{X}\mathbf{v}, \Delta)}{\|S(\mathbf{X}\mathbf{v}, \Delta)\|_2}$, where $\Delta = 0$ if this results in $\|\mathbf{u}\|_1 \leqslant c$; otherwise, Δ is chosen to be a positive constant such that $\|\mathbf{u}\|_1 = c$.
   (b) $\mathbf{v} \leftarrow \arg\min_{\mathbf{v}} \left\{ \frac{1}{2}\|\mathbf{X}^T\mathbf{u} - \mathbf{v}\|^2 + \lambda_1 \sum_j |v_j| + \lambda_2 \sum_j |v_j - v_{j-1}| \right\}$.
3. $d \leftarrow \mathbf{u}^T\mathbf{X}\mathbf{v}$.

Step 2(b) can be performed using fast software implementing fused lasso regression, as described in Friedman *and others* (2007), Tibshirani and Wang (2008), and Hoefling (2009).

2.4 *PMD for missing data and choice of c_1 and c_2*

The algorithm for computing the PMD works even in the case of missing data. When some elements of the data matrix $\mathbf{X}$ are missing, those elements can simply be excluded from all computations. Let C denote the set of indices of nonmissing elements in $\mathbf{X}$. The criterion is as follows:

$$\text{maximize}_{\mathbf{u},\mathbf{v}} \sum_{(i,j)\in C} X_{ij}u_iv_j \quad \text{subject to } \|\mathbf{u}\|_2^2 \leqslant 1, \ \ \|\mathbf{v}\|_2^2 \leqslant 1, \ P_1(\mathbf{u}) \leqslant c_1, \ P_2(\mathbf{v}) \leqslant c_2. \tag{2.15}$$

The PMD can therefore be used as a method for missing data imputation. This is related to SVD-based data imputation methods proposed in the literature (see, e.g. Troyanskaya *and others*, 2001).

The possibility of computing the PMD in the presence of missing data leads to a simple and automated method for the selection of the constants c_1 and c_2 in the PMD criterion. We can treat c_1 and c_2 as tuning parameters and can take an approach similar to cross-validation in order to select their values. For simplicity, we demonstrate this method for the rank-1 case here.

Algorithm 5: Selection of tuning parameters for PMD

1. From the original data matrix $\mathbf{X}$, construct 10 data matrices $\mathbf{X}_1, \ldots, \mathbf{X}_{10}$, each of which is missing a nonoverlapping one-tenth of the elements of $\mathbf{X}$, sampled at random from the rows and columns.
2. For each candidate value of c_1 and c_2:
   (a) For $i \in 1, \ldots, 10$:
      i) Fit the PMD to $\mathbf{X}_i$ with tuning parameters c_1 and c_2 and calculate $\hat{\mathbf{X}}_i = d\mathbf{u}\mathbf{v}^T$, the resulting estimate of $\mathbf{X}_i$.
      ii) Record the mean squared error of the estimate $\hat{\mathbf{X}}_i$. This mean squared error is obtained by computing the mean of the squared differences between elements of $\mathbf{X}$ and the corresponding elements of $\hat{\mathbf{X}}_i$, where the mean is taken only over elements that are missing from $\mathbf{X}_i$.
   (b) Record the average mean squared error across $\mathbf{X}_1, \ldots, \mathbf{X}_{10}$ for tuning parameters c_1 and c_2.
3. The optimal values of c_1 and c_2 are those which correspond to the lowest mean squared error.

Note that in Step 1 of this method, we construct each $\mathbf{X}_i$ by randomly removing scattered elements of the matrix $\mathbf{X}$. That is, we are not removing entire rows of $\mathbf{X}$ or entire columns of $\mathbf{X}$, but rather individual elements of the data matrix. Similar approaches are taken in Wold (1978) and Owen and Perry (2009).

Though c_1 and c_2 can always be chosen as described above, for certain applications cross-validation may not be necessary. If the PMD is applied to a data set as a descriptive method, in order to obtain an intuitive understanding of the data, then one might simply fix c_1 and c_2 based on some other criterion. For instance, one could select small values of c_1 and c_2 in order to obtain factors that have a desirable level of sparsity.

2.5 *Relationship between PMD and other matrix decompositions*

In the statistical and machine learning literature, a number of matrix decompositions have been developed. We present some of these decompositions here, as they are related to the PMD. The best known of these decompositions is the SVD, which takes the form of (2.1). The SVD has a number of interesting properties, but the vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ of the SVD have (in general) no nonzero elements, and the elements may be positive or negative. These qualities result in vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ that are often not interpretable.

Lee and Seung (1999, 2001) developed the nonnegative matrix factorization (NNMF) in order to improve upon the interpretability of the SVD. The matrix $\mathbf{X}$ is approximated as

$$\mathbf{X} \approx \sum_{k=1}^{K} \mathbf{u}_k \mathbf{v}_k^T, \tag{2.16}$$

where the elements of $\mathbf{u}_k$ and $\mathbf{v}_k$ are constrained to be nonnegative. The factors $\mathbf{u}_k$ and $\mathbf{v}_k$ can be interpretable: the authors apply the NNMF to a database of faces and show that the resulting factors represent facial features. The SVD does not result in interpretable facial features.

Hoyer (2002, 2004) presents the nonnegative sparse coding (NNSC), an extension of the NNMF that results in nonnegative vectors $\mathbf{v}_k$ and $\mathbf{u}_k$, one or both of which may be sparse. Sparsity is achieved using an L_1-penalty. Since NNSC enforces a nonnegativity constraint, the resulting vectors can be quite different

from those obtained via PMD; moreover, the iterative algorithm for finding the NNSC vectors is not guaranteed to decrease the objective at each step.

Lazzeroni and Owen (2002) present the plaid model, which (in the simplest case) takes the form of (1.1). They seek d_k, $\mathbf{u}_k$, $\mathbf{v}_k$ that minimize

$$\sum_{i,j}\left(X_{ij}-\sum_{k=1}^{K} d_k u_{ik} v_{jk}\right)^2 \quad \text{subject to } u_{ik}\in\{0,1\},\ \ v_{jk}\in\{0,1\}. \tag{2.17}$$

Though the plaid model provides interpretable layers, it has the drawback that the criterion cannot be optimized exactly due to the nonconvex form of the constraints on $\mathbf{u}_k$ and $\mathbf{v}_k$.

2.6 *An example: PMD for DNA copy number data*

We now consider a simple example involving comparative genomic hybridization (CGH) data, which measures DNA copy number changes along a chromosome in cancer samples. It is known that some cancers are characterized by contiguous regions of chromosomal gain or loss. For this reason, the fused lasso criterion has been proposed as a way to denoise CGH data for a single sample (Tibshirani and Wang 2008):

$$\hat{\beta}=\arg\min_{\beta}\left\{\frac{1}{2}\sum_{j=1}^{p}(y_j-\beta_j)^2+\lambda_1\sum_j|\beta_j|+\lambda_2\sum_j|\beta_j-\beta_{j-1}|\right\}. \tag{2.18}$$

In (2.18), $\mathbf{y}$ is a vector of length p corresponding to measured log copy number gain/loss, ordered along the chromosome, and $\hat{\beta}$ is a smoothed estimate of the copy number. Note that $\lambda_1, \lambda_2 \geqslant 0$.

Now, suppose that multiple CGH samples are available. We expect some patterns of gain and loss to be shared between some of the samples, and we wish to identify those patterns and samples. Let $\mathbf{X}$ denote the data matrix; the n rows denote the samples and the p columns correspond to (ordered) CGH spots. In this case, the use of PMD(L_1, FL) is appropriate because we wish to encourage sparsity in $\mathbf{u}$ (corresponding to a subset of samples) and sparsity and smoothness in $\mathbf{v}$ (corresponding to chromosomal regions). The use of PMD(L_1, FL) in this context is related to ongoing work by Nowak *and others* (2009). One could apply PMD(L_1, FL) to all chromosomes together (making sure that smoothness in the fused lasso penalty is not required between chromosomes) or one could apply PMD(L_1, FL) to each chromosome separately.

We demonstrate this method on a simple simulated example. We simulate 12 samples, each of which consists of copy number measurements on 1000 spots on a single chromosome. Five of the 12 samples contain a region of gain from spots 100–500. In Figure 2, we compare the results of PMD(L_1, L_1) to PMD(L_1, FL). It is clear that the latter method precisely uncovers the region of gain and the set of samples in which that gained region is present. Simulation details are given in the Appendix (Section A.3).

3. SPARSE PCA VIA PMD

3.1 *Three methods for sparse PCA*

Here, we begin with an $n \times p$ data matrix $\mathbf{X}$, with centered columns. Several methods have been proposed for estimating sparse principal components, based on either the maximum variance property of principal components or the regression/reconstruction error property. In this section, we present 2 existing methods for sparse PCA from the literature, as well as a new method based on the PMD. We will then go on to show that these 3 methods are closely related to each other. We will use the connection between PMD and one of the other methods to develop a fast algorithm for what was previously a computationally difficult formulation for sparse PCA.

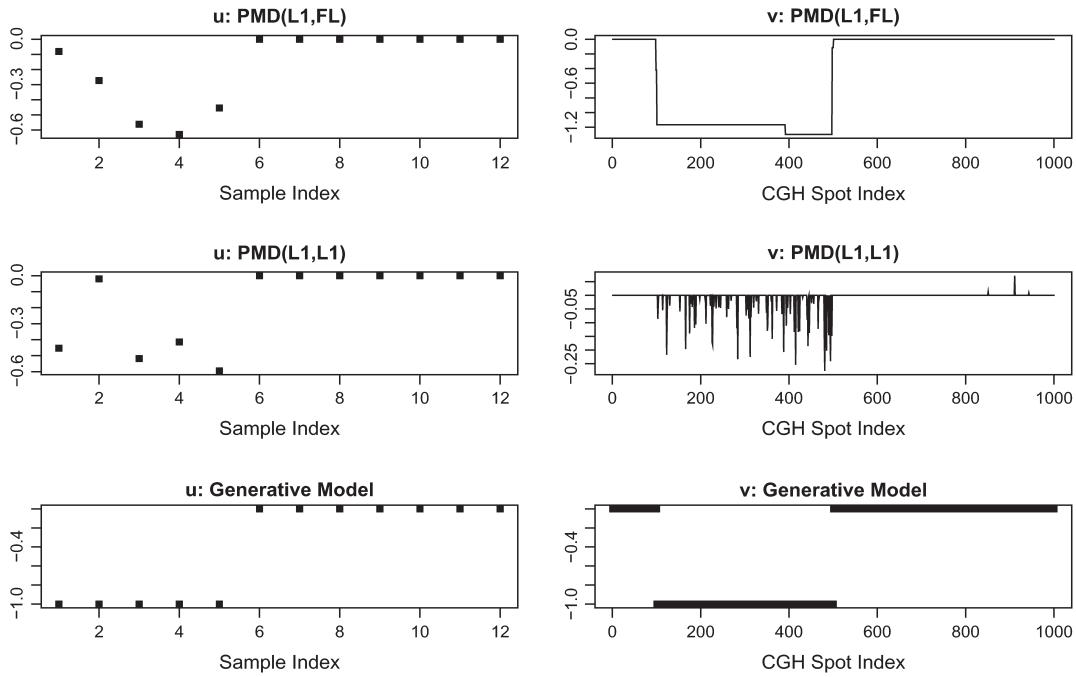

Fig. 2. Simulated CGH data. Top: results of PMD(L_1, FL); middle: results of PMD(L_1, L_1); bottom: generative model. PMD(L_1, FL) successfully identifies both the region of gain and the subset of samples for which that region is present.

The 3 methods for sparse PCA are as follows:

1. SPCA: Zou *and others* (2006) exploit the regression/reconstruction error property of principal components in order to obtain sparse principal components. For a single component, their sparse principal components analysis (SPCA) technique solves

$$\text{minimize}_{\theta,\mathbf{v}} \|\mathbf{X} - \mathbf{X}\mathbf{v}\theta^T\|_F^2 + \lambda_1\|\mathbf{v}\|_2^2 + \lambda_2\|\mathbf{v}\|_1 \quad \text{subject to } \|\theta\|_2 = 1, \tag{3.1}$$

   where $\lambda_1, \lambda_2 \geqslant 0$ and $\mathbf{v}$ and θ are p-vectors. The criterion can equivalently be written with an inequality L_2 bound on θ, in which case it is biconvex in θ and $\mathbf{v}$.
2. SCoTLASS: The SCoTLASS procedure of Jolliffe *and others* (2003) uses the maximal variance characterization for principal components. The first sparse principal component solves the problem

$$\text{maximize}_{\mathbf{v}} \mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v} \quad \text{subject to } \|\mathbf{v}\|_2^2 \leqslant 1, \quad \|\mathbf{v}\|_1 \leqslant c, \tag{3.2}$$

   and subsequent components solve the same problem with the additional constraint that they must be orthogonal to the previous components. This problem is not convex, since a convex objective must be maximized, and the computations are difficult. Trendafilov and Jolliffe (2006) provide a projected gradient algorithm for optimizing (3.2). We will show that this criterion can be optimized much more simply by direct application of Algorithm 3 in Section 2.3.
3. SPC: We propose a new method for sparse PCA. Consider the PMD criterion with $P_2(\mathbf{v}) = \|\mathbf{v}\|_1$, and no P_1-constraint on $\mathbf{u}$. We call this criterion PMD($\cdot$, L_1), and it can be written as follows:

$$\text{maximize}_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}\mathbf{v} \quad \text{subject to } \|\mathbf{v}\|_1 \leqslant c_2, \quad \|\mathbf{u}\|_2^2 \leqslant 1, \ \|\mathbf{v}\|_2^2 \leqslant 1. \tag{3.3}$$

The algorithm for PMD($\cdot$, L_1) is obtained by replacing Step 2(a) of Algorithm 3 (the single-factor PMD(L_1, L_1) algorithm) with the simpler update $\mathbf{u} \leftarrow \frac{\mathbf{Xv}}{\|\mathbf{Xv}\|_2}$. We will refer to this as "sparse principal components," or SPC.

Now, consider the SPC criterion in (3.3). It is easily shown that if $\mathbf{v}$ is fixed and we seek $\mathbf{u}$ to maximize (3.3), then the optimal $\mathbf{u}$ will be $\frac{\mathbf{Xv}}{\|\mathbf{Xv}\|_2}$. Therefore, $\mathbf{v}$ that maximizes (3.3) also maximizes

$$\mathbf{v}^T\mathbf{X}^T\mathbf{Xv} \quad \text{subject to } \|\mathbf{v}\|_1 \leqslant c_2, \ \ \|\mathbf{v}\|_2 \leqslant 1. \tag{3.4}$$

We recognize (3.4) as the SCoTLASS criterion (3.2). Now, since we have a fast iterative algorithm for solving (3.3), this means that we have also developed a fast method to maximize the SCoTLASS criterion. We can extend SPC to find the first K sparse principal components, as in Algorithm 2. Note, however, that only the first component is the solution to the SCoTLASS criterion (since we are not enforcing the constraint that component $\mathbf{v}_k$ be orthogonal to components $\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}$).

It is also not hard to show that PMD applied to a covariance matrix with symmetric L_1-penalties on the rows and columns, as follows,

$$\arg\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Xv} \quad \text{subject to } \|\mathbf{u}\|_2^2 \leqslant 1, \ \ \|\mathbf{u}\|_1 \leqslant c, \ \|\mathbf{v}\|_2^2 \leqslant 1, \ \|\mathbf{v}\|_1 \leqslant c, \tag{3.5}$$

results in solutions $\mathbf{u} = \mathbf{v}$. (This follows from the Cauchy–Schwarz inequality applied to vectors $\mathbf{Xv}$ and $\mathbf{Xu}$.) As a result, these solutions solve the SCoTLASS criterion as well. This also means that SPC can be performed using the covariance matrix instead of the raw data in cases where this is more convenient (e.g. if $n \gg p$ or if the raw data are unavailable).

We have shown that the SPC criterion is equivalent to the SCoTLASS criterion for one component and that the fast algorithm for the former can be used to maximize the latter. It turns out that there also is a connection between the SPCA criterion and the SPC criterion. Consider a modified version of the SPCA criterion (3.1) that uses the bound form, rather than the Lagrange form, of the constraints on $\mathbf{v}$:

$$\text{minimize}_{\theta,\mathbf{v}} \|\mathbf{X} - \mathbf{Xv}\theta^T\|_F^2 \quad \text{subject to } \|\mathbf{v}\|_2^2 \leqslant 1, \ \ \|\mathbf{v}\|_1 \leqslant c, \ \|\theta\|_2^2 = 1. \tag{3.6}$$

With $\|\theta\|_2^2 = 1$, we have

$$\begin{aligned} \|\mathbf{X} - \mathbf{Xv}\theta^T\|_F^2 &= \text{tr}((\mathbf{X} - \mathbf{Xv}\theta^T)^T(\mathbf{X} - \mathbf{Xv}\theta^T)) \\ &= \text{tr}(\mathbf{X}^T\mathbf{X}) - 2\text{tr}(\theta\mathbf{v}^T\mathbf{X}^T\mathbf{X}) + \text{tr}(\theta\mathbf{v}^T\mathbf{X}^T\mathbf{Xv}\theta^T) \\ &= \text{tr}(\mathbf{X}^T\mathbf{X}) - 2\mathbf{v}^T\mathbf{X}^T\mathbf{X}\theta + \mathbf{v}^T\mathbf{X}^T\mathbf{Xv}. \end{aligned} \tag{3.7}$$

So solving (3.6) is equivalent to

$$\text{maximize}_{\theta,\mathbf{v}}\{2\mathbf{v}^T\mathbf{X}^T\mathbf{X}\theta - \mathbf{v}^T\mathbf{X}^T\mathbf{Xv}\} \quad \text{subject to } \|\theta\|_2^2 = 1, \ \ \|\mathbf{v}\|_2^2 \leqslant 1, \ \|\mathbf{v}\|_1 \leqslant c \tag{3.8}$$

or equivalently

$$2\mathbf{v}^T\mathbf{X}^T\mathbf{X}\theta - \mathbf{v}^T\mathbf{X}^T\mathbf{Xv} \quad \text{subject to } \|\theta\|_2^2 \leqslant 1, \ \ \|\mathbf{v}\|_2^2 \leqslant 1, \ \|\mathbf{v}\|_1 \leqslant c. \tag{3.9}$$

Now, suppose we add an additional constraint to (3.6): that is, let us require also that $\|\theta\|_1 \leqslant c$. We maximize

$$2\mathbf{v}^T\mathbf{X}^T\mathbf{X}\theta - \mathbf{v}^T\mathbf{X}^T\mathbf{Xv} \quad \text{subject to } \|\mathbf{v}\|_2^2 \leqslant 1, \ \ \|\mathbf{v}\|_1 \leqslant c, \ \|\theta\|_2^2 \leqslant 1, \ \|\theta\|_1 \leqslant c \tag{3.10}$$

with respect to θ and $\mathbf{v}$. Note that for any vectors $\mathbf{w}$ and $\mathbf{z}$, $\|\mathbf{z}-\mathbf{w}\|_2^2 \geqslant 0$. This means that $\mathbf{w}^T\mathbf{w} \geqslant 2\mathbf{w}^T\mathbf{z} - \mathbf{z}^T\mathbf{z}$. Let $\mathbf{w} = \mathbf{X}\mathbf{v}$ and $\mathbf{z} = \mathbf{X}\theta$; it follows that $\theta^T\mathbf{X}^T\mathbf{X}\theta \geqslant 2\mathbf{v}^T\mathbf{X}^T\mathbf{X}\theta - \mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v}$. So (3.10) is maximized when $\mathbf{v} = \theta$; that is, $\mathbf{v}$ that maximizes (3.10) also maximizes

$$\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v} \quad \text{subject to } \|\mathbf{v}\|_2^2 \leqslant 1, \quad \|\mathbf{v}\|_1 \leqslant c, \tag{3.11}$$

which of course is simply the SCoTLASS criterion (3.2) again. Therefore, we have shown that if a symmetric L_1-constraint on θ is added to the bound form of the SPCA criterion, then the SCoTLASS criterion results. From this argument, it is also clear that the solution to the bound form of SPCA will give lower reconstruction error (defined as $\|\mathbf{X} - \mathbf{X}\mathbf{v}\theta^T\|_F^2$) than the solution to the SCoTLASS criterion.

We compare the proportion of variance explained by SPC and SPCA on a publicly available gene expression data set from http://icbp.lbl.gov/breastcancer/, and described in Chin *and others* (2006), consisting of 19 672 gene expression measurements on 89 samples. (For consistency with Section 4.3, we use the subset of the samples for which both gene expression and CGH measurements are available.) For computational reasons, we use only the subset of the data consisting of the 5% of genes with highest variance. We compute the first 25 sparse principal components for SPC using the constraint on $\mathbf{v}$ that results in an average of 195 genes with nonzero elements per sparse component. We then perform SPCA on the same data, using tuning parameters such that each loading has the same number of nonzero elements obtained using the SPC method. Figure 3 shows the proportion of variance explained by the first k sparse principal components, defined as $\text{tr}(\mathbf{X}_k^T\mathbf{X}_k)$, where $\mathbf{X}_k = \mathbf{X}\mathbf{V}_k(\mathbf{V}_k^T\mathbf{V}_k)^{-1}\mathbf{V}_k^T$ and where $\mathbf{V}_k$ is the matrix that has the first k sparse principal components as its columns. (This definition is proposed in Shen and Huang, 2008.) SPC results in a substantially greater proportion of variance explained, as expected.

Our extension of PMD to the problem of identifying sparse principal components is closely related to the SPCA method of Shen and Huang (2008). They present a method for identifying sparse principal components via a regularized low-rank matrix approximation, as follows:

$$\text{minimize}_{\mathbf{u},\mathbf{v}} \|\mathbf{X} - \mathbf{u}\mathbf{v}^T\|_F^2 + P_\lambda(\mathbf{v}) \quad \text{subject to } \|\mathbf{u}\|_2 = 1. \tag{3.12}$$

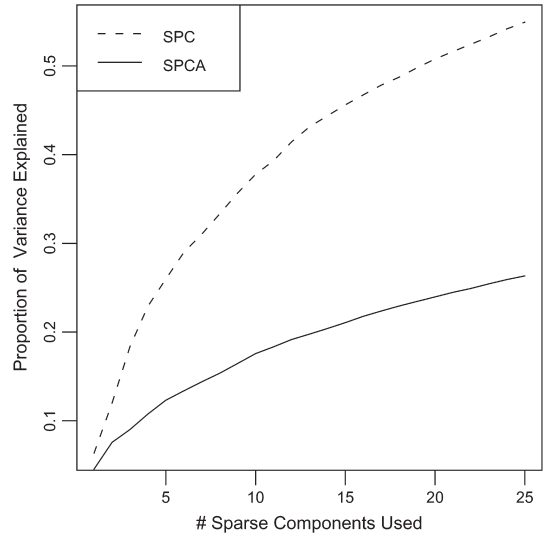

Fig. 3. Breast cancer gene expression data: a greater proportion of variance is explained when SPC is used to obtain the sparse principal components, rather than SPCA. Multiple SPC components were obtained as described in Algorithm 2.

Then, $\mathbf{v}^* = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}$ is the first sparse principal component of their method. They present a number of forms for $P_\lambda(\mathbf{v})$, including $P_\lambda(\mathbf{v}) = \|\mathbf{v}\|_1$. This is very close in spirit to PMD$(\cdot, L_1)$, and in fact the algorithm is almost the same. Since Shen and Huang (2008) use the Lagrange form of the constraint on $\mathbf{v}$, their formulation does not solve the SCoTLASS criterion. Our method unifies the regularized low-rank matrix approximation approach of Shen and Huang (2008) with the maximum variance criterion of Jolliffe *and others* (2003) and the SPCA method of Zou *and others* (2006).

To summarize, in our view, the SCoTLASS criterion (3.2) is the simplest, most natural way to define the notion of sparse principal components. Unfortunately, the criterion is difficult to optimize. Our SPC criterion (3.3) recasts this problem as a biconvex one, leading to an extremely simple algorithm for the solution of the first SCoTLASS component. Furthermore, the SPCA criterion (3.1) is somewhat complex. But we have shown that when a natural symmetric constraint is added to the SPCA criterion (3.1), it is also equivalent to (3.2) and (3.3). Taken as a whole, these arguments point to the SPC criterion (3.3) as the criterion of choice for this problem, at least for a single component.

3.2 *Another option for SPC with multiple factors*

As mentioned in Section 3.1, the first sparse principal component of our SPC method optimizes the SCoTLASS criterion. But subsequent sparse principal components obtained using SPC do not, since we do not enforce that $\mathbf{v}_k$ be orthogonal to $\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}$. It is not obvious that SPC can be extended to achieve orthogonality among subsequent $\mathbf{v}_i$s, or even that orthogonality is desirable. However, SPC can be easily extended to give something similar to orthogonality.

Consider the criterion for the first factor of SPC, given in (3.3). One could extend to multiple factors as proposed in Algorithm 2. (This was done in Figure 3.) Alternatively, one could obtain multiple factors $\mathbf{u}_k, \mathbf{v}_k$ by optimizing the following criterion, for $k > 1$:

$$\text{maximize}_{\mathbf{u}_k, \mathbf{v}_k} \mathbf{u}_k^T \mathbf{X} \mathbf{v}_k \quad \text{subject to } \|\mathbf{v}_k\|_1 \leqslant c_2, \ \ \|\mathbf{u}_k\|_2^2 \leqslant 1, \ \|\mathbf{v}_k\|_2^2 \leqslant 1, \ \mathbf{u}_k \perp \mathbf{u}_1, \ldots, \mathbf{u}_{k-1}. \quad (3.13)$$

With $\mathbf{u}_k$ fixed, one can solve (3.13) for $\mathbf{v}_k$ easily, as has been done throughout this paper (e.g. Step 2(b) of Algorithm 3). With $\mathbf{v}_k$ fixed, the problem is as follows: we must find $\mathbf{u}_k$ that maximizes

$$\mathbf{u}_k^T \mathbf{X} \mathbf{v}_k \quad \text{subject to } \|\mathbf{u}_k\|_2^2 \leqslant 1, \ \ \mathbf{u}_k \perp \mathbf{u}_1, \ldots, \mathbf{u}_{k-1}. \quad (3.14)$$

Let $\mathbf{U}_{k-1}^\perp$ denote an orthogonal basis that is orthogonal to $\mathbf{U}_{k-1}$, the matrix with columns $\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$. It follows that $\mathbf{u}_k$ is in the column space of $\mathbf{U}_{k-1}^\perp$, and so can be written as $\mathbf{u}_k = \mathbf{U}_{k-1}^\perp \theta$. Note also that $\|\mathbf{u}_k\|_2 = \|\theta\|_2$. So (3.14) is equivalent to solving

$$\theta^T {\mathbf{U}_{k-1}^\perp}^T \mathbf{X} \mathbf{v}_k \quad \text{subject to } \|\theta\|_2^2 \leqslant 1, \quad (3.15)$$

and so we find that the optimal θ is

$$\theta = \frac{{\mathbf{U}_{k-1}^\perp}^T \mathbf{X} \mathbf{v}_k}{\|{\mathbf{U}_{k-1}^\perp}^T \mathbf{X} \mathbf{v}_k\|_2}. \quad (3.16)$$

Therefore, the value of $\mathbf{u}_k$ that solves (3.14) is

$$\mathbf{u}_k = \frac{\mathbf{U}_{k-1}^\perp {\mathbf{U}_{k-1}^\perp}^T \mathbf{X} \mathbf{v}_k}{\|{\mathbf{U}_{k-1}^\perp}^T \mathbf{X} \mathbf{v}_k\|_2} = \frac{(\mathbf{I} - \mathbf{U}_{k-1} {\mathbf{U}_{k-1}}^T) \mathbf{X} \mathbf{v}_k}{\|{\mathbf{U}_{k-1}^\perp}^T \mathbf{X} \mathbf{v}_k\|_2}. \quad (3.17)$$

So we can use this update step for $\mathbf{u}_k$ to develop an iterative algorithm to find multiple factors for (3.3), the single-factor SPC criterion, that yields orthogonal $\mathbf{u}_k$s. Though we have not guaranteed that the $\mathbf{v}_k$s will

be exactly orthogonal, they are unlikely to be very correlated since the different $\mathbf{v}_k$s each are associated with orthogonal $\mathbf{u}_k$s. Based on our initial investigation, this appears to be a promising path for obtaining multiple sparse principal components.

4. Penalized CCA via PMD

4.1 *PMD and other approaches to penalized CCA*

Suppose that we have n observations on $p + q$ variables and that the variables are naturally partitioned into 2 sets of size p and q. Let $\mathbf{X}$ denote the $n \times p$ matrix that is comprised of the first set of variables, and let $\mathbf{Y}$ denote the $n \times q$ matrix that is comprised of the remaining variables; assume that the columns of $\mathbf{X}$ and $\mathbf{Y}$ have been centered and scaled. CCA, developed by Hotelling (1936), involves finding $\mathbf{u}$, $\mathbf{v}$ that maximize cor($\mathbf{Xu}$, $\mathbf{Yv}$)—that is, that solve

$$\text{maximize}_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Yv} \quad \text{subject to } \mathbf{u}^T\mathbf{X}^T\mathbf{Xu} \leqslant 1, \;\; \mathbf{v}^T\mathbf{X}^T\mathbf{Xv} \leqslant 1. \tag{4.1}$$

There is a closed-form solution for $\mathbf{u}$ and $\mathbf{v}$ that involves the eigenvectors of some function of the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$. We call $\mathbf{u}$ and $\mathbf{v}$ the canonical variates.

CCA results in vectors $\mathbf{u}$, $\mathbf{v}$ that are not sparse, and these vectors are not unique if p or q exceeds n. In certain applications, especially if p or q is large, one might be interested in finding a linear combination of the variables in $\mathbf{X}$ and $\mathbf{Y}$ that has large correlation but is also sparse in the variables used.

One way to obtain penalized canonical variates would simply be to include penalties in (4.1):

$$\text{maximize}_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Yv} \quad \text{subject to } \mathbf{u}^T\mathbf{X}^T\mathbf{Xu} \leqslant 1, \;\; \mathbf{v}^T\mathbf{X}^T\mathbf{Xv} \leqslant 1, \;\; P_1(\mathbf{u}) \leqslant c_1, \;\; P_2(\mathbf{v}) \leqslant c_2. \tag{4.2}$$

It has been shown that in other high-dimensional problems, treating the covariance matrix as diagonal can yield good results (see, e.g. Dudoit *and others*, 2001; Tibshirani *and others*, 2003). For this reason, rather than using (4.2) as our penalized CCA criterion, we substitute in the identity matrix $\mathbf{I}$ for $\mathbf{X}^T\mathbf{X}$ and $\mathbf{Y}^T\mathbf{Y}$; this gives what could be called "diagonal penalized CCA":

$$\text{maximize}_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Yv} \quad \text{subject to } \|\mathbf{u}\|_2^2 \leqslant 1, \;\; \|\mathbf{v}\|_2^2 \leqslant 1, \;\; P_1(\mathbf{u}) \leqslant c_1, \;\; P_2(\mathbf{v}) \leqslant c_2. \tag{4.3}$$

Of course, this criterion is simply (2.7) with $\mathbf{X}$ replaced with $\mathbf{X}^T\mathbf{Y}$; it can be solved with Algorithm 1. But in practice, it can be solved more efficiently, without computation of $\mathbf{X}^T\mathbf{Y}$. To compute multiple canonical variates, we use Algorithm 2. Following the notation of Section 2.3, we refer to this method as PMD(A, B) if A is the penalty on $\mathbf{u}$ and B is the penalty on $\mathbf{v}$.

PMD(L_1, L_1) yields sparse vectors $\mathbf{u}$ and $\mathbf{v}$ for c_1 and c_2 sufficiently small. Waaijenborg *and others* (2008) also present a sparse CCA method, and their algorithm for finding the sparse canonical variates is quite similar to PMD(L_1, L_1). However, they arrive at their algorithm in a circuitous way, as an approximation to the elastic net, and they do not state the exact criterion that they are solving. Moreover, their method involves the Lagrange form rather than the bound form of the L_1-constraints on $\mathbf{u}$ and $\mathbf{v}$; as a result, it yields a different solution. Parkhomenko *and others* (2007) and Wiesel *and others* (2008) present sparse CCA algorithms that lack exact criteria and biconvexity, respectively. Our method for sparse CCA is very closely related to the method of Parkhomenko *and others* (2009), which we encountered after our paper was submitted.

When L_1-penalties are used for both P_1 and P_2, then the values of c_1 and c_2 can be chosen by cross-validation, where c_1 and c_2 are chosen using a grid search to maximize (across the cross-validation folds) the quantity cor($\mathbf{Xu}$, $\mathbf{Yv}$), where $\mathbf{u}$ and $\mathbf{v}$ are computed on a training set and $\mathbf{X}$ and $\mathbf{Y}$ constitute an independent test set. Alternatively, values of c_1 and c_2 can simply be chosen to result in desired amounts of sparsity of $\mathbf{u}$ and $\mathbf{v}$.

4.2 *Sparse CCA applied to simulated data*

We demonstrate the PMD(L_1, L_1) method on a simple simulated example. In this simulation, $p, q > n$, so classical CCA cannot be used. There are 2 sparse latent factors that generate $\mathbf{X}$ and $\mathbf{Y}$; the factors are orthogonal to each other. Results can be seen in Figure 4. For comparison, we also computed the SVD of $\mathbf{X}^T\mathbf{Y}$. Compared to the SVD, PMD(L_1, L_1) does fairly well at identifying linear combinations of the underlying factors. Details of the simulation are given in Section A.4 of the Appendix.

4.3 *Application of penalized CCA to genomic data*

In genomic research, it is becoming increasingly common for researchers to use multiple assays in order to characterize a single set of samples. For instance, gene expression measurements and DNA copy number data might be available on the same set of tissue samples. The patients might also be genotyped. Examples of studies that combine gene expression and copy number and/or single-nucleotide polymorphism (SNP) data include Hyman *and others* (2002), Pollack *and others* (2002), Morley *and others* (2004), and Stranger *and others* (2005, 2007). While much research has gone into developing methods for the identification of genes and SNPs that are associated with an outcome based on a single gene expression or SNP data set, the question of how to combine the results of multiple assays in order to perform inference across the data sets has not been thoroughly investigated.

If both gene expression data and genotype data are available on the same set of samples, then a natural question is to identify sets of genes that are correlated with sets of SNPs. Both Parkhomenko *and others* (2007, 2009) and Waaijenborg *and others* (2008) demonstrate the use of sparse CCA for this purpose. Similarly, if CGH and gene expression data both are available for a set of cancer samples, then one might wish to identify a set of genes that have expression that is correlated with a set of chromosomal gains or losses.

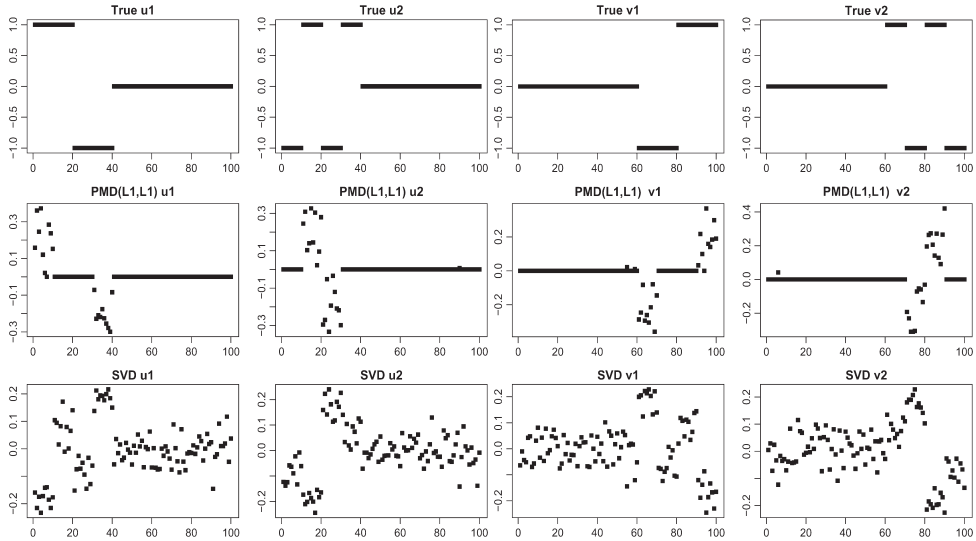

Fig. 4. The efficacy of PMD(L_1, L_1) is demonstrated using a simulation in which $\mathbf{X}$ is generated from 2 sparse latent factors, called $\mathbf{u}_1$ and $\mathbf{u}_2$, and $\mathbf{Y}$ is generated from 2 sparse latent factors, called $\mathbf{v}_1$ and $\mathbf{v}_2$. The PMD(L_1, L_1) method identifies linear combinations of these sparse factors. Details of the simulation are given in Section A.4 of the Appendix.

In the case of gene expression and CGH data, it makes sense to perform penalized CCA with an L_1-penalty on the canonical variate corresponding to genes and a fused lasso penalty on the canonical variate corresponding to copy number—in other words, we might use PMD(L_1, FL). There are 2 ways that this could be done. Penalized CCA could be applied using all available gene expression data and copy number data on all chromosomes (being sure that the fused lasso smoothness penalty is not applied between chromosomes). Alternatively, one could perform penalized CCA once per chromosome, each time using copy number data on that chromosome and all the available gene expression data. (The expression data are not restricted to a particular chromosome.) We choose to pursue this latter approach.

We examine the performance of PMD(L_1, FL) on the breast cancer data set publicly available at http://icbp.lbl.gov/breastcancer/ and described in Chin *and others* (2006). In addition to the gene expression data described earlier, CGH measurements are also available on the same set of 89 samples. There are $p = 19\,672$ gene expression measurements and $q = 2149$ CGH measurements. For convenience and interpretability of the results, we ran PMD(L_1, FL) using values of the tuning parameters that resulted in a list of approximately 25 nonzero genes per chromosome (i.e. 25 nonzero elements of **u**) and very smooth and somewhat sparse **v**. Since we performed PMD(L_1, FL) once for each of the 23 chromosomes, we obtained 23 **v** vectors. The 23 **v**s are shown in the left panel of Figure 5. Nonzero **u**s and **v**s were found for all chromosomes except for chromosome 2. It is clear that PMD(L_1, FL) resulted in both sparsity and smoothness of the **v** vectors.

The genes corresponding to nonzero weights in each **u** vector can also be examined. Consider Table 1, which shows the genes that had nonzero weights when sparse CCA was run using the CGH spots on chromosome 1 and all the available genes. Notably, only genes located on chromosome 1 were given nonzero weights. This is intuitive: a copy number change on chromosome 1 should be correlated with expression changes in the genes that were amplified or deleted. Similar results were seen when PMD(L_1, FL) was run using the CGH spots on other chromosomes. If one is interested only in discovering genes

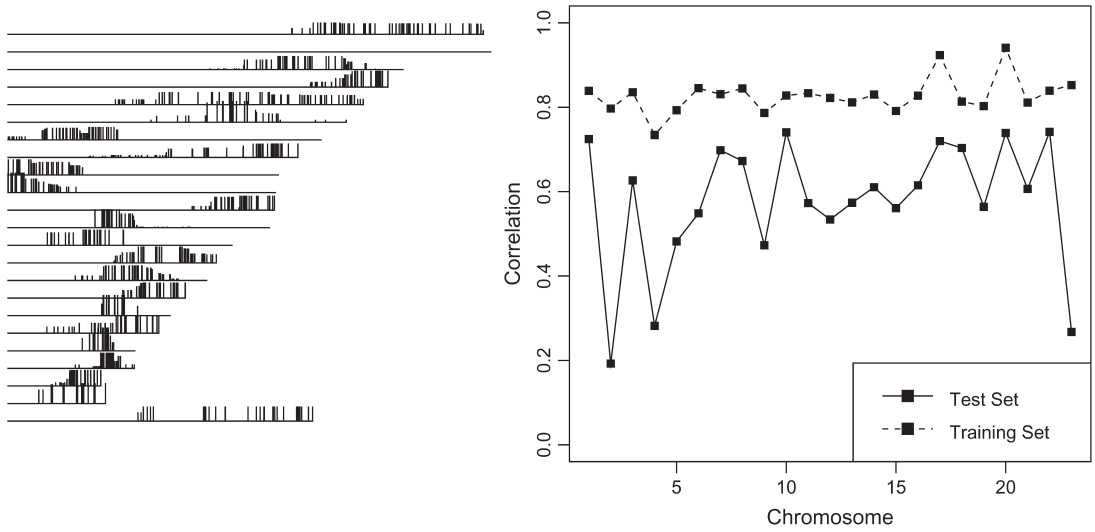

Fig. 5. PMD(L_1, FL) was performed for the breast cancer data set. Left: for each chromosome, the weights of **v** obtained using PMD(L_1, FL) are shown. All the **v** weights shown are positive, but the results would not be affected by flipping the signs of both **v** and **u**. On chromosome 2, **v** has no nonzero elements. Right: for each chromosome, **u** and **v** were computed on a training set consisting of 3/4 of the samples, and cor(**Xu**, **Yv**) is plotted, where **X** and **Y** are the training (dashed) and test (solid) data.

Table 1. *PMD(L_1, FL) was performed using the CGH spots on chromosome* 1 *and gene expression measurements on all chromosomes. The genes corresponding to nonzero elements of* **u** *are shown. Notably, all the genes with nonzero u_is are located on chromosome* 1. *Similar results hold for the other chromosomes*

i	Gene	Chromosome	u_i
1	Jumping translocation break point	1	0.039
2	Translocated promoter region (to activated MET oncogene)	1	0.153
3	Glyceronephosphate O-acyltransferase	1	0.255
4	NADH dehydrogenase (ubiquinone) Fe-S protein 2	1	0.265
5	Nucleoporin 133kD	1	0.007
6	Geranylgeranyl diphosphate synthase 1	1	0.131
7	Rab3 GTPase-activating protein, noncatalytic subunit (150 kD)	1	0.283
8	Peroxisomal biogenesis factor 11B	1	0.154
9	Phosphatidylinositol glycan, class C	1	0.124
10	Tubulin-specific chaperone e	1	0.069
11	Protoporphyrinogen oxidase	1	0.052
12	Tuftelin 1	1	0.037
13	Papillary renal cell carcinoma (translocation associated)	1	0.055
14	Splicing factor 3b, subunit 4, 49 kD	1	0.469
15	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase	1	0.27
16	Hypothetical protein FLJ12671	1	0.229
17	Hypothetical protein HSPC155	1	0.168
18	Mitochondrial ribosomal protein L24	1	0.195
19	HSPC003 protein	1	0.391
20	Hypothetical protein FLJ10876	1	0.091
21	CGI-78 protein	1	0.154
22	Chromosome 1 open reading frame 27	1	0.133
23	Hypothetical protein My014	1	0.278

not on chromosome k that are correlated with copy number change on chromosome k, then one could perform PMD(L_1, FL) using the CGH spots on chromosome k and only genes that are not on chromosome k.

In order to assess whether PMD(L_1, FL) is capturing real structure in the breast cancer data, we computed p-values for the penalized canonical variates. For each chromosome, a p-value for the penalized canonical variates was computed as follows:

1. Let $\mathbf{u}$, $\mathbf{v}$ denote the penalized canonical variates found for this chromosome, and record $c = \text{cor}(\mathbf{Xu}, \mathbf{Yv})$.
2. For $i \in 1, \ldots, B$, permute the samples in $\mathbf{X}$ to obtain $\mathbf{X}^*$; then, compute $\mathbf{u}^*$ and $\mathbf{v}^*$, the penalized canonical variates based on data $(\mathbf{X}^*, \mathbf{Y})$. Record $c_i = \text{cor}(\mathbf{X}^*\mathbf{u}^*, \mathbf{Yv}^*)$.
3. The p-value is given by $\frac{1}{B}\sum_{i=1}^{B} 1_{|c_i| \geqslant |c|}$.

With the exception of chromosome 2, all p-values were significant.

In order to further assess the penalized canonical variates that we obtained, we used a training set/test set approach, as follows:

1. We repeatedly divided the 89 samples into a training set $(\mathbf{X}_{\text{tr}}, \mathbf{Y}_{\text{tr}})$ containing 3/4 of the samples, and a test set $(\mathbf{X}_{\text{te}}, \mathbf{Y}_{\text{te}})$ containing the remaining samples.
2. Penalized CCA was performed on the training set to obtain the vectors $\mathbf{u}_{\text{tr}}$ and $\mathbf{v}_{\text{tr}}$.
3. $\text{cor}(\mathbf{X}_{\text{tr}}\mathbf{u}_{\text{tr}}, \mathbf{Y}_{\text{tr}}\mathbf{v}_{\text{tr}})$ and $\text{cor}(\mathbf{X}_{\text{te}}\mathbf{u}_{\text{tr}}, \mathbf{Y}_{\text{te}}\mathbf{v}_{\text{tr}})$ were computed.

The right panel of Figure 5 shows the average values of $\text{cor}(\mathbf{X}_{\text{tr}}\mathbf{u}_{\text{tr}}, \mathbf{Y}_{\text{tr}}\mathbf{v}_{\text{tr}})$ and $\text{cor}(\mathbf{X}_{\text{te}}\mathbf{u}_{\text{tr}}, \mathbf{Y}_{\text{te}}\mathbf{v}_{\text{tr}})$ for each chromosome. Even on the test set, quite high correlations result for most chromosomes. In the absence of signal, the average value of $\text{cor}(\mathbf{X}_{\text{te}}\mathbf{u}_{\text{tr}}, \mathbf{Y}_{\text{te}}\mathbf{v}_{\text{tr}})$ would be 0.

5. Discussion

We have developed a method for finding a PMD in an efficient manner. This decomposition builds upon a variety of existing matrix decompositions, such as the SVD, the NNMF (Lee and Seung, 1999, 2001), and the plaid model (Lazzeroni and Owen, 2002). We are most interested in obtaining a decomposition made up of sparse vectors. To do this, we use an L_1-penalty on the rows and columns of our decomposition. We also explore the use of an L_1-penalty on the rows and a fused lasso penalty on the columns; this is appropriate if the samples correspond to DNA copy number, ordered by chromosomal location. We exploit the biconvex nature of the PMD criterion in order to minimize it via an alternating algorithm.

We have applied the PMD to give attractive solutions to 2 additional problems: sparse PCA and sparse CCA. We used the resulting sparse CCA method to identify the sets of genes that are correlated with regions of DNA copy number change using a data set consisting of DNA copy number change and gene expression measurements on the same set of samples. In addition, we have established connections between 3 different methods for obtaining sparse principal components.

An R package implementing these methods, called PMA (for penalized multivariate analysis) is available on CRAN.

Acknowledgements

We thank an associate editor and 2 referees for helpful comments, Stephen Boyd for a useful discussion of biconvexity, and Holger Hoefling for the use of his R code for fused lasso regression. *Conflict of Interest:* None declared.

Funding

National Defense Science and Engineering Graduate Fellowship to D.W.; National Science Foundation (DMS-9971405 to R.T., DMS-0505676 to T.H.); National Institutes of Health (N01-HV-28183 to R.T., 2R01 CA 72028-07 to T.H.).

APPENDIX

A.1 *Proof of Theorem* 2.1

Let $\mathbf{u}_k$ and $\mathbf{v}_k$ denote column k of $\mathbf{U}$ and $\mathbf{V}$, respectively. We prove the theorem by expanding out the squared Frobenius norm and rearranging terms:

$$\begin{aligned}\|\mathbf{X} - \mathbf{UDV}^T\|_F^2 &= \text{tr}((\mathbf{X} - \mathbf{UDV}^T)^T(\mathbf{X} - \mathbf{UDV}^T)) \\ &= -2\text{tr}(\mathbf{VDU}^T\mathbf{X}) + \text{tr}(\mathbf{VDU}^T\mathbf{UDV}^T) + \|\mathbf{X}\|_F^2 \\ &= \sum_{k=1}^{K} d_k^2 - 2\text{tr}(\mathbf{DU}^T\mathbf{XV}) + \|\mathbf{X}\|_F^2\end{aligned}$$

$$= \sum_{k=1}^{K} d_k^2 - 2\sum_{k=1}^{K} d_k \mathbf{u}_k^T \mathbf{X}\mathbf{v}_k + \|\mathbf{X}\|_F^2. \tag{A.1}$$

A.2 *Proof of Lemma* 2.2

We seek $\mathbf{u}$ that minimizes

$$-\mathbf{u}^T\mathbf{a} \quad \text{subject to } \|\mathbf{u}\|_2^2 \leqslant 1, \|\mathbf{u}\|_1 \leqslant c_1. \tag{A.2}$$

First, we rewrite the criterion using Lagrange multipliers:

$$-\mathbf{u}^T\mathbf{a} + \lambda\|\mathbf{u}\|_2^2 + \Delta\|\mathbf{u}\|_1, \tag{A.3}$$

and we differentiate, set the derivative to 0, and solve for $\mathbf{u}$:

$$0 = -\mathbf{a} + 2\lambda\mathbf{u} + \Delta\Gamma, \tag{A.4}$$

where $\Gamma_i = \text{sgn}(u_i)$ if $u_i \neq 0$; otherwise, $\Gamma_i \in [-1, 1]$. The Karush–Kuhn–Tucker conditions for optimality consist of (A.4), along with $\lambda(\|\mathbf{u}\|_2^2 - 1) = 0$ and $\Delta(\|\mathbf{u}\|_1 - c_1) = 0$. Now if $\lambda > 0$, we have

$$\mathbf{u} = \frac{S(\mathbf{a}, \Delta)}{2\lambda}. \tag{A.5}$$

In general, we have either $\lambda = 0$ (if this results in a feasible solution) or λ must be chosen such that $\|\mathbf{u}\|_2 = 1$. So we see that

$$\mathbf{u} = \frac{S(\mathbf{a}, \Delta)}{\|S(\mathbf{a}, \Delta)\|_2}. \tag{A.6}$$

Again by the Karush–Kuhn–Tucker conditions, either $\Delta = 0$ (if this results in a feasible solution) or Δ must be chosen such that $\|\mathbf{u}\|_1 = c_1$. So, $\Delta = 0$ if this results in $\|\mathbf{u}\|_1 \leqslant c_1$; otherwise, we choose Δ such that $\|\mathbf{u}\|_1 = c_1$. This completes the proof of the Lemma.

A.3 *Simulation details for Figure* 2

Let $\mathbf{X}$ be a 12×1000 matrix of data. The elements of $\mathbf{X}$ are generated as follows:

1. For $i \in 1, \ldots, 5$ and $j \in 100, \ldots, 500$, $X_{ij} \sim N(1, 1)$.
2. Otherwise, $X_{ij} \sim N(0, 1)$.

In other words, the first 5 patients have a region of gain between positions 100 and 500.

A.4 *Simulation details for Figure* 4

We generate matrices $\mathbf{X}$ and $\mathbf{Y}$, with $n = 50$ and $p = 100$.

1. Let $\mathbf{u}_1$ be a vector of length p, with 20 1s, 20 $-$1s, and 60 0s.
2. Let $\mathbf{u}_2$ be a vector of length p, with 10 $-$1s, 10 1s, 10 $-$1s, 10 1s, and 60 0s.
3. Let $\mathbf{v}_1$ be a vector of length p, with 60 0s, 20 $-$1s, and 20 1s.
4. Let $\mathbf{v}_2$ be a vector of length p, with 60 0s, 10 1s, 10 $-$1s, 10 1s, and 10 $-$1s.
5. Let $\mathbf{w}_1$ and $\mathbf{w}_2$ be orthonormal vectors of length n.
6. Generate the data matrices as follows: $X_{ij} \sim N(w_{1i}u_{1j} + w_{2i}u_{2j}, 0.3^2)$ and $Y_{ij} \sim N(w_{1i}v_{1j} + w_{2i}v_{2j}, 0.3^2)$.

REFERENCES

BOYD, S. AND VANDENBERGHE, L. (2004). *Convex Optimization*. New York: Cambridge University Press.

CHIN, K., DEVRIES, S., FRIDLYAND, J., SPELLMAN, P., ROYDASGUPTA, R., KUO, W.-L., LAPUK, A., NEVE, R., QIAN, Z., RYDER, T. *and others* (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529–541.

DUDOIT, S., FRIDLYAND, J. AND SPEED, T. (2001). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **96**, 1151–1160.

ECKART, C. AND YOUNG, G. (1936). The approximation of one matrix by another of low rank. *Psychometrika* **1**, 211.

FRIEDMAN, J., HASTIE, T., HOEFLING, H. AND TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics* **1**, 302–332.

HOEFLING, H. (2009). A path algorithm for the fused lasso (in preparation).

HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321–377.

HOYER, P. (2002). Non-negative sparse coding. *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, 557–565.

HOYER, P. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research* **5**, 1457–1469.

HYMAN, E., KAURANIEMI, P., HAUTANIEMI, S., WOLF, M., MOUSSES, S., ROZENBLUM, E., RINGNER, M., SAUTER, G., MONNI, O., ELKAHLOUN, A. *and others* (2002). Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Research* **62**, 6240–6245.

JOLLIFFE, I., TRENDAFILOV, N. AND UDDIN, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics* **12**, 531–547.

LAZZERONI, L. AND OWEN, A. (2002). Plaid models for gene expression data. *Statistica Sinica* **12**, 61–86.

LEE, D. D. AND SEUNG, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788.

LEE, D. D. AND SEUNG, H. S. (2001). Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Diatterich, and V. Tresp (eds.), *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, pp. 556–562.

MORLEY, M., MOLONY, C., WEBER, T., DEVLIN, J., EWENS, K., SPIELMAN, R. AND CHEUNG, V. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747.

NOWAK, G., HASTIE, T., POLLACK, J. AND TIBSHIRANI, R. (2009). Identifying copy number changes in CGH data for multiple samples (in preparation).

OWEN, A. B. AND PERRY, P. O. (2009). Bi-cross-validation of the SVD and the non-negative matrix factorization. *Annals of Applied Statistics*.

PARKHOMENKO, E., TRITCHLER, D. AND BEYENE, J. (2007). Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proceedings* **1**, S119.

PARKHOMENKO, E., TRITCHLER, D. AND BEYENE, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology* **8**, 1–34.

POLLACK, J., SORLIE, T., PEROU, C., REES, C., JEFFREY, S., LONNING, P., TIBSHIRANI, R., BOTSTEIN, D., BORRESEN-DALE, A. AND BROWN, P. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12963–12968.

SHEN, H. AND HUANG, J. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis* **101**, 1015–1034.

STRANGER, B., FORREST, M., CLARK, A., MINICHIELLO, M., DEUTSCH, S., LYLE, R., HUNT, S., KAHL, B., ANTONARAKIS, S., TAVARE, S. *and others* (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genetics* **1**, e78.

STRANGER, B., FORREST, M., DUNNING, M., INGLE, C., BEAZLEY, C., THORNE, N., REDON, R., BIRD, C., DE GRASSI, A., LEE, C. *and others* (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. AND CHU, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science* **18**, 104–117.

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. AND KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B* **67**, 91–108.

TIBSHIRANI, R. AND WANG, P. (2008). Spatial smoothing and hotspot detection for CGH data using the fused lasso. *Biostatistics* **9**, 18–29.

TRENDAFILOV, N. AND JOLLIFFE, I. (2006). Projected gradient approach to the numerical solution of the scotlass. *Computational Statistics & Data Analysis* **50**, 242–253.

TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., BROWN, P., HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D. AND ALTMAN, R. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* **16**, 520–525.

WAAIJENBORG, S., VERSELEWEL DE WITT HAMER, P. AND ZWINDERMAN, A. (2008). Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology* **7**. Issue 1, Article 3.

WIESEL, A., KLIGER, M. AND HERO, A. (2008). A greedy approach to sparse canonical correlation analysis (In preparation). Available at http://arxiv.org/abs/0801.2748.

WOLD, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* **20**, 397–405.

ZOU, H., HASTIE, T. AND TIBSHIRANI, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics* **15**, 265–286.

[*Received July 28, 2008; first revision December 23, 2008; second revision February 4, 2009; accepted for publication February 24, 2009*]