CUR from a Sparse Optimization Viewpoint

Jacob Bien[*]
Department of Statistics
Stanford University
Stanford, CA 94305
jbien@stanford.edu

Ya Xu[*]
Department of Statistics
Stanford University
Stanford, CA 94305
yax.stanford@gmail.com

Michael W. Mahoney
Department of Mathematics
Stanford University
Stanford, CA 94305
mmahoney@cs.stanford.edu

Abstract

The CUR decomposition provides an approximation of a matrix $\mathbf{X}$ that has low reconstruction error and that is sparse in the sense that the resulting approximation lies in the span of only a few columns of $\mathbf{X}$. In this regard, it appears to be similar to many sparse PCA methods. However, CUR takes a randomized algorithmic approach, whereas most sparse PCA methods are framed as convex optimization problems. In this paper, we try to understand CUR from a sparse optimization viewpoint. We show that CUR is implicitly optimizing a sparse regression objective and, furthermore, cannot be directly cast as a sparse PCA method. We also observe that the sparsity attained by CUR possesses an interesting structure, which leads us to formulate a sparse PCA method that achieves a CUR-like sparsity.

1 Introduction

CUR decompositions are a recently-popular class of randomized algorithms that approximate a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ by using only a small number of actual columns of $\mathbf{X}$ [12, 4]. CUR decompositions are often described as SVD-like low-rank decompositions that have the additional advantage of being easily interpretable to domain scientists. The motivation to produce a more interpretable low-rank decomposition is also shared by sparse PCA (SPCA) methods, which are optimization-based procedures that have been of interest recently in statistics and machine learning.

Although CUR and SPCA methods start with similar motivations, they proceed very differently. For example, most CUR methods have been randomized, and they take a purely algorithmic approach. By contrast, most SPCA methods start with a combinatorial optimization problem, and they then solve a relaxation of this problem. Thus far, it has not been clear to researchers how the CUR and SPCA approaches are related. It is the purpose of this paper to understand CUR decompositions from a sparse optimization viewpoint, thereby elucidating the connection between CUR decompositions and the SPCA class of sparse optimization methods.

To do so, we begin by putting forth a combinatorial optimization problem (see (6) below) which CUR is implicitly approximately optimizing. This formulation will highlight two interesting features of CUR: first, CUR attains a distinctive pattern of sparsity, which has practical implications from the SPCA viewpoint; and second, CUR is implicitly optimizing a regression-type objective. These two observations then lead to the three main contributions of this paper: (a) first, we formulate a non-randomized optimization-based version of CUR (see Problem 1: GL-REG in Section 3) that is based on a convex relaxation of the CUR combinatorial optimization problem; (b) second, we show that, in contrast to the original PCA-based motivation for CUR, CUR's implicit objective cannot be directly expressed in terms of a PCA-type objective (see Theorem 3 in Section 4); and (c) third, we propose an SPCA approach (see Problem 2: GL-SPCA in Section 5) that achieves the sparsity structure of CUR within the PCA framework. We also provide a brief empirical evaluation of our two proposed objectives. While our proposed GL-REG and GL-SPCA methods are promising in and of themselves, our purpose in this paper is not to explore them as alternatives to CUR; instead, our goal is to use them to help clarify the connection between CUR and SPCA methods.

[*]Jacob Bien and Ya Xu contributed equally.

We conclude this introduction with some remarks on notation. Given a matrix $\mathbf{A}$, we use $\mathbf{A}_{(i)}$ to denote its ith row (as a row-vector) and $\mathbf{A}^{(i)}$ its ith column. Similarly, given a set of indices $\mathcal{I}$, $\mathbf{A}_{\mathcal{I}}$ and $\mathbf{A}^{\mathcal{I}}$ denote the submatrices of $\mathbf{A}$ containing only these $\mathcal{I}$ rows and columns, respectively. Finally, we let $\mathcal{L}_{\mathrm{col}}(\mathbf{A})$ denote the column space of $\mathbf{A}$.

2 Background

In this section, we provide a brief background on CUR and SPCA methods, with a particular emphasis on topics to which we will return in subsequent sections. Before doing so, recall that, given an input matrix $\mathbf{X}$, Principal Component Analysis (PCA) seeks the k-dimensional hyperplane with the lowest reconstruction error. That is, it computes a $p \times k$ orthogonal matrix $\mathbf{W}$ that minimizes

$$\mathrm{ERR}(\mathbf{W}) = ||\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^T||_F. \tag{1}$$

Writing the SVD of $\mathbf{X}$ as $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, the minimizer of (1) is given by $\mathbf{V}_k$, the first k columns of $\mathbf{V}$. In the data analysis setting, each column of $\mathbf{V}$ provides a particular linear combination of the columns of $\mathbf{X}$. These linear combinations are often thought of as latent factors. In many applications, interpreting such factors is made much easier if they are comprised of only a small number of actual columns of $\mathbf{X}$, which is equivalent to $\mathbf{V}_k$ only having a small number of nonzero elements.

2.1 CUR matrix decompositions

CUR decompositions were proposed by Drineas and Mahoney [12, 4] to provide a low-rank approximation to a data matrix $\mathbf{X}$ by using only a small number of actual columns and/or rows of $\mathbf{X}$. Fast randomized variants [3], deterministic variants [5], Nyström-based variants [1, 11], and heuristic variants [17] have also been considered. Observing that the best rank-k approximation to the SVD provides the best set of k linear combinations of all the columns, one can ask for the best set of k *actual* columns. Most formalizations of "best" lead to intractable combinatorial optimization problems [12], but one can take advantage of oversampling (choosing slightly more than k columns) and randomness as computational resources to obtain strong quality-of-approximation guarantees.

Theorem 1 (Relative-error CUR [12])**.** *Given an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and an integer k, there exists a randomized algorithm that chooses a random subset $\mathcal{I} \subset \{1, \dots, p\}$ of size $c = O(k \log k \log(1/\delta)/\epsilon^2)$ such that $\mathbf{X}^{\mathcal{I}}$, the $n \times c$ submatrix containing those c columns of $\mathbf{X}$, satisfies*

$$||\mathbf{X} - \mathbf{X}^{\mathcal{I}}\mathbf{X}^{\mathcal{I}+}\mathbf{X}||_F = \min_{\mathbf{B} \in \mathbb{R}^{c \times p}} ||\mathbf{X} - \mathbf{X}^{\mathcal{I}}\mathbf{B}||_F \leq (1+\epsilon)||\mathbf{X} - \mathbf{X}_k||_F, \tag{2}$$

with probability at least $1 - \delta$, where $\mathbf{X}_k$ is the best rank k approximation to $\mathbf{X}$.

The algorithm referred to by Theorem 1 is very simple:

1) Compute the *normalized statistical leverage scores*, defined below in (3).
2) Form $\mathcal{I}$ by randomly sampling c columns of $\mathbf{X}$, using these normalized statistical leverage scores as an importance sampling distribution.
3) Return the $n \times c$ matrix $\mathbf{X}^{\mathcal{I}}$ consisting of these selected columns.

The key issue here is the choice of the importance sampling distribution. Let the $p \times k$ matrix $\mathbf{V}_k$ be the top-k right singular vectors of $\mathbf{X}$. Then the *normalized statistical leverage scores* are

$$\pi_i = \frac{1}{k}||\mathbf{V}_{k(i)}||_2^2, \tag{3}$$

for all $i = 1, \dots, p$, where $\mathbf{V}_{k(i)}$ denotes the i-th row of $\mathbf{V}_k$. These scores, proportional to the Euclidean norms of the *rows* of the top-k right singular vectors, define the relevant nonuniformity structure to be used to identify good (in the sense of Theorem 1) columns. In addition, these scores are proportional to the diagonal elements of the projection matrix onto the top-k right singular subspace. Thus, they generalize the so-called hat matrix [8], and they have a natural interpretation as capturing the "statistical leverage" or "influence" of a given column on the best low-rank fit of the data matrix [8, 12].

2.2 Regularized sparse PCA methods

SPCA methods attempt to make PCA easier to interpret for domain experts by finding sparse approximations to the *columns* of $\mathbf{V}$.[1] There are several variants of SPCA. For example, Jolliffe *et al.* [10]

[1]For SPCA, we only consider sparsity in the right singular vectors $\mathbf{V}$ and not in the left singular vectors $\mathbf{U}$. This is similar to considering only the choice of columns and not of both columns and rows in CUR.

and Witten *et al.* [19] use the maximum variance interpretation of PCA and provide an optimization problem which explicitly encourages sparsity in $\mathbf{V}$ based on a Lasso constraint [18]. d'Aspremont *et al.* [2] take a similar approach, but instead formulate the problem as an SDP.

Zou *et al.* [21] use the minimum reconstruction error interpretation of PCA to suggest a different approach to the SPCA problem; this formulation will be most relevant to our present purpose. They begin by formulating PCA as the solution to a regression-type problem.

Theorem 2 (Zou *et al.* [21])**.** *Given an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and an integer k, let $\mathbf{A}$ and $\mathbf{W}$ be $p \times k$ matrices. Then, for any $\lambda > 0$, let*

$$(\mathbf{A}^*, \mathbf{V}_k^*) = \mathrm{argmin}_{\mathbf{A},\mathbf{W}\in\mathbb{R}^{p\times k}} ||\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{A}^T||_F^2 + \lambda||\mathbf{W}||_F^2 \quad \text{s.t. } \mathbf{A}^T\mathbf{A} = \mathbf{I}_k. \tag{4}$$

Then, the minimizing matrices $\mathbf{A}^*$ and $\mathbf{V}_k^*$ satisfy $\mathbf{A}^{*(i)} = s_i\mathbf{V}^{(i)}$ and $\mathbf{V}_k^{*(i)} = s_i\frac{\mathbf{\Sigma}_{ii}^2}{\mathbf{\Sigma}_{ii}^2+\lambda}\mathbf{V}^{(i)}$, where $s_i = 1$ or -1.

That is, up to signs, $\mathbf{A}^*$ consists of the top-k right singular vectors of $\mathbf{X}$, and $\mathbf{V}_k^*$ consists of those same vectors "shrunk" by a factor depending on the corresponding singular value. Given this regression-type characterization of PCA, Zou *et al.* [21] then "sparsify" the formulation by adding an L_1 penalty on $\mathbf{W}$:

$$(\mathbf{A}^*, \mathbf{V}_k^*) = \mathrm{argmin}_{\mathbf{A},\mathbf{W}\in\mathbb{R}^{p\times k}} ||\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{A}^T||_F^2 + \lambda||\mathbf{W}||_F^2 + \lambda_1||\mathbf{W}||_1 \quad \text{s.t. } \mathbf{A}^T\mathbf{A} = \mathbf{I}_k, \tag{5}$$

where $||\mathbf{W}||_1 = \sum_{ij}|\mathbf{W}_{ij}|$. This regularization tends to sparsify $\mathbf{W}$ element-wise, so that the solution $\mathbf{V}_k^*$ gives a sparse approximation of $\mathbf{V}_k$.

3 Expressing CUR as an optimization problem

In this section, we present an optimization formulation of CUR. Recall, from Section 2.1, that CUR takes a purely algorithmic approach to the problem of approximating a matrix in terms of a small number of its columns. That is, it achieves sparsity indirectly by randomly selecting c columns, and it does so in such a way that the reconstruction error is small with high probability (Theorem 1). By contrast, SPCA methods are generally formulated as the exact solution to an optimization problem.

From Theorem 1, it is clear that CUR seeks a subset $\mathcal{I}$ of size c for which $\min_{\mathbf{B}\in\mathbb{R}^{c\times p}} ||\mathbf{X} - \mathbf{X}^{\mathcal{I}}\mathbf{B}||_F$ is small. In this sense, CUR can be viewed as a randomized algorithm for approximately solving the following combinatorial optimization problem:

$$\min_{\mathcal{I}\subset\{1,\ldots,p\}} \min_{\mathbf{B}\in\mathbb{R}^{c\times p}} ||\mathbf{X} - \mathbf{X}^{\mathcal{I}}\mathbf{B}||_F \quad \text{s.t. } |\mathcal{I}| \leq c. \tag{6}$$

In words, this objective asks for the subset of c columns of $\mathbf{X}$ which best describes the entire matrix $\mathbf{X}$. Notice that relaxing $|\mathcal{I}| = c$ to $|\mathcal{I}| \leq c$ does not affect the optimum. This optimization problem is analogous to all-subsets multivariate regression [7], which is known to be NP-hard.

However, by using ideas from the optimization literature we can approximate this combinatorial problem as a regularized regression problem that is convex. First, notice that (6) is equivalent to

$$\min_{\mathbf{B}\in\mathbb{R}^{p\times p}} ||\mathbf{X} - \mathbf{X}\mathbf{B}||_F \quad \text{s.t. } \sum_{i=1}^{p} \mathbb{1}_{\{||\mathbf{B}_{(i)}||_2\neq 0\}} \leq c, \tag{7}$$

where we now optimize over a $p \times p$ matrix $\mathbf{B}$. To see the equivalence between (6) and (7), note that the constraint in (7) is the same as finding some subset $\mathcal{I}$ with $|\mathcal{I}| \leq c$ such that $\mathbf{B}_{\mathcal{I}^c} = \mathbf{0}$.

The formulation in (7) provides a natural entry point to proposing a convex optimization approach corresponding to CUR. First notice that (7) uses an L_0 norm on the rows of $\mathbf{B}$, which is not convex. However, we can approximate the L_0 constraint by a *group lasso* penalty, which uses a well-known convex heuristic proposed by Yuan *et al.* [20] that encourages prespecified *groups* of parameters to be simultaneously sparse. Thus, the combinatorial problem in (6) can be approximated by the following convex (and thus tractable) problem:

Problem 1 (Group lasso regression: GL-REG). *Given an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, let $\mathbf{B} \in \mathbb{R}^{p\times p}$ and $t > 0$. The* GL-REG *problem is to solve*

$$\mathbf{B}^* = \mathrm{argmin}_{\mathbf{B}} ||\mathbf{X} - \mathbf{X}\mathbf{B}||_F \quad \text{s.t. } \sum_{i=1}^{p} ||\mathbf{B}_{(i)}||_2 \leq t, \tag{8}$$

where t is chosen to get c nonzero rows in $\mathbf{B}^*$.

Since the rows of $\mathbf{B}$ are grouped together in the penalty $\sum_{i=1}^{p} ||\mathbf{B}_{(i)}||_2$, the row vector $\mathbf{B}_{(i)}$ will tend to be either dense or entirely zero. Note also that the algorithm to solve Problem 1 is a special case of Algorithm 1 (see below), which solves the GL-SPCA problem, to be introduced later. (Finally, as a side remark, note that our proposed GL-REG is strikingly similar to a recently proposed method for sparse inverse covariance estimation [6, 15].)

4 Distinguishing CUR from SPCA

Our original intention in casting CUR in the optimization framework was to understand better whether CUR could be seen as an SPCA-type method. So far, we have established CUR's connection to regression by showing that CUR can be thought of as an approximation algorithm for the sparse regression problem (7). In this section, we discuss the relationship between regression and PCA, and we show that CUR cannot be directly cast as an SPCA method.

To do this, recall that regression, in particular "self" regression, finds a $\mathbf{B} \in \mathbb{R}^{p\times p}$ that minimizes

$$||\mathbf{X} - \mathbf{X}\mathbf{B}||_F. \tag{9}$$

On the other hand, PCA-type methods find a set of directions $\mathbf{W}$ that minimize

$$\textsc{err}(\mathbf{W}) := ||\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^+||_F. \tag{10}$$

Here, unlike in (1), we do not assume that $\mathbf{W}$ is orthogonal, since the minimizer produced from SPCA methods is often not required to be orthogonal (recall Section 2.2).

Clearly, with no constraints on $\mathbf{B}$ or $\mathbf{W}$, we can trivially achieve zero reconstruction error in both cases by taking $\mathbf{B} = \mathbf{I}_p$ and $\mathbf{W}$ any $p \times p$ full-rank matrix. However, with additional constraints, these two problems can be very different. It is common to consider sparsity and/or rank constraints. We have seen in Section 3 that CUR effectively requires $\mathbf{B}$ to be row-sparse; in the standard PCA setting, $\mathbf{W}$ is taken to be rank k (with $k < p$), in which case (10) is minimized by $\mathbf{V}_k$ and obtains the optimal value $\textsc{err}(\mathbf{V}_k) = ||\mathbf{X} - \mathbf{X}_k||_F$; finally, for SPCA, $\mathbf{W}$ is further required to be sparse.

To illustrate the difference between the reconstruction errors (9) and (10) when extra constraints are imposed, consider the 2-dimensional toy example in Figure 1. In this example, we compare regression with a row-sparsity constraint to PCA with both rank and sparsity constraints. With $\mathbf{X} \in \mathbb{R}^{n\times 2}$, we plot $\mathbf{X}^{(2)}$ against $\mathbf{X}^{(1)}$ as the solid points in both plots of Figure 1. Constraining $\mathbf{B}_{(2)} = 0$ (giving row-sparsity, as with CUR methods), (9) becomes $\min_{B_{12}} ||\mathbf{X}^{(2)} - \mathbf{X}^{(1)}B_{12}||_2$, which is a simple linear regression, represented by the black thick line and minimizing the sum of squared vertical errors as shown. The red line (left plot) shows the first principal component direction, which minimizes $\textsc{err}(\mathbf{W})$ among all rank-one matrices $\mathbf{W}$. Here, $\textsc{err}(\mathbf{W})$ is the sum of squared projection distances (red dotted lines). Finally, if $\mathbf{W}$ is further required to be sparse in the $\mathbf{X}^{(2)}$ direction (as with SPCA methods), we get the rank-one, sparse projection represented by the green line in Figure 1 (right). The two sets of dotted lines in each plot clearly differ, indicating that their corresponding reconstruction errors are different as well. Since we have shown that CUR is minimizing a regression-based objective, this toy example suggests that CUR may not in fact be optimizing a PCA-type objective such as (10). Next, we will make this intuition more precise.

The first step to showing that CUR is an SPCA method would be to produce a matrix $\mathbf{V}_{\text{CUR}}$ for which $\mathbf{X}^{\mathcal{I}}\mathbf{X}^{\mathcal{I}+}\mathbf{X} = \mathbf{X}\mathbf{V}_{\text{CUR}}\mathbf{V}_{\text{CUR}}^+$, *i.e.* to express CUR's approximation in the form of an SPCA approximation. However, this equality implies $\mathcal{L}_{\text{col}}(\mathbf{X}\mathbf{V}_{\text{CUR}}\mathbf{V}_{\text{CUR}}^+) \subseteq \mathcal{L}_{\text{col}}(\mathbf{X}^{\mathcal{I}})$, meaning that $(\mathbf{V}_{\text{CUR}})_{\mathcal{I}^c} = \mathbf{0}$. If such a $\mathbf{V}_{\text{CUR}}$ existed, then clearly $\textsc{err}(\mathbf{V}_{\text{CUR}}) = ||\mathbf{X} - \mathbf{X}^{\mathcal{I}}\mathbf{X}^{\mathcal{I}+}\mathbf{X}||_F$, and so CUR could be regarded as implicitly performing sparse PCA in the sense that (a) $\mathbf{V}_{\text{CUR}}$ is sparse; and (b) by Theorem 1 (with high probability), $\textsc{err}(\mathbf{V}_{\text{CUR}}) \le (1+\epsilon)\textsc{err}(\mathbf{V}_k)$. Thus, the existence of such a $\mathbf{V}_{\text{CUR}}$ would cast CUR directly as a randomized approximation algorithm for SPCA. However, the following theorem states that unless an unrealistic constraint on $\mathbf{X}$ holds, there does not exist a matrix $\mathbf{V}_{\text{CUR}}$ for which $\textsc{err}(\mathbf{V}_{\text{CUR}}) = ||\mathbf{X} - \mathbf{X}^{\mathcal{I}}\mathbf{X}^{\mathcal{I}+}\mathbf{X}||_F$. The larger implication of this theorem is that CUR cannot be directly viewed as an SPCA-type method.

Theorem 3. *Let $\mathcal{I} \subset \{1,\ldots,p\}$ be an index set and suppose $\mathbf{W} \in \mathbb{R}^{p\times p}$ satisfies $\mathbf{W}_{\mathcal{I}^c} = \mathbf{0}$. Then,*

$$||\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^+||_F > ||\mathbf{X} - \mathbf{X}^{\mathcal{I}}\mathbf{X}^{\mathcal{I}+}\mathbf{X}||_F,$$

unless $\mathcal{L}_{\text{col}}(\mathbf{X}^{\mathcal{I}}) \perp \mathcal{L}_{\text{col}}(\mathbf{X}^{\mathcal{I}^c})$, in which case "$\ge$" holds.

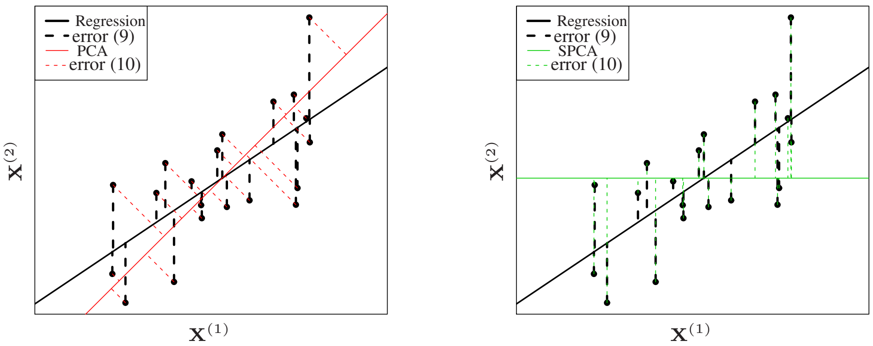

Figure 1: Example of the difference in reconstruction errors (9) and (10), when additional constraints imposed. Left: regression with row-sparsity constraint (black) compared with PCA with low rank constraint (red). Right: regression with row-sparsity constraint (black) compared with PCA with low rank and sparsity constraint (green). In both plots, the corresponding errors are represented by the dotted lines.

Proof.

$$\begin{aligned}||\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{W}^{+}||_F^2 &= ||\mathbf{X}-\mathbf{X}^{\mathcal{I}}\mathbf{W}_{\mathcal{I}}\mathbf{W}^{+}||_F^2 = ||\mathbf{X}-\mathbf{X}^{\mathcal{I}}\mathbf{W}_{\mathcal{I}}(\mathbf{W}_{\mathcal{I}}^T\mathbf{W}_{\mathcal{I}})^{-1}\mathbf{W}^T||_F^2 \\ &= ||\mathbf{X}^{\mathcal{I}}-\mathbf{X}^{\mathcal{I}}\mathbf{W}_{\mathcal{I}}\mathbf{W}_{\mathcal{I}}^{+}||_F^2 + ||\mathbf{X}^{\mathcal{I}^c}||_F^2 \geq ||\mathbf{X}^{\mathcal{I}^c}||_F^2 \\ &= ||\mathbf{X}^{\mathcal{I}^c}-\mathbf{X}^{\mathcal{I}}\mathbf{X}^{\mathcal{I}+}\mathbf{X}^{\mathcal{I}^c}||_F^2 + ||\mathbf{X}^{\mathcal{I}}\mathbf{X}^{\mathcal{I}+}\mathbf{X}^{\mathcal{I}^c}||_F^2 \\ &= ||\mathbf{X}-\mathbf{X}^{\mathcal{I}}\mathbf{X}^{\mathcal{I}+}\mathbf{X}||_F^2 + ||\mathbf{X}^{\mathcal{I}}\mathbf{X}^{\mathcal{I}+}\mathbf{X}^{\mathcal{I}^c}||_F^2 \geq ||\mathbf{X}-\mathbf{X}^{\mathcal{I}}\mathbf{X}^{\mathcal{I}+}\mathbf{X}||_F^2.\end{aligned}$$

The last inequality is strict unless $\mathbf{X}^{\mathcal{I}}\mathbf{X}^{\mathcal{I}+}\mathbf{X}^{\mathcal{I}^c} = \mathbf{0}$. $\square$

5 CUR-type sparsity and the group lasso SPCA

Although CUR cannot be directly cast as an SPCA-type method, in this section we propose a sparse PCA approach (which we call the group lasso SPCA or GL-SPCA) that accomplishes something very close to CUR. Our proposal produces a $\mathbf{V}^*$ that has rows that are entirely zero, and it is motivated by the following two observations about CUR. First, following from the definition of the leverage scores (3), CUR chooses columns of $\mathbf{X}$ based on the norm of their corresponding rows of $\mathbf{V}_k$. Thus, it essentially "zeros-out" the rows of $\mathbf{V}_k$ with small norms (in a probabilistic sense). Second, as we have noted in Section 4, if CUR could be expressed as a PCA method, its principal directions matrix "$\mathbf{V}_{\text{CUR}}$" would have $p-c$ rows that are entirely zero, corresponding to removing those columns of $\mathbf{X}$.

Recall that Zou *et al.* [21] obtain a sparse $\mathbf{V}^*$ by including in (5) an additional L_1 penalty from the optimization problem (4). Since the L_1 penalty is on the entire matrix viewed as a vector, it encourages only unstructured sparsity. To achieve the CUR-type row sparsity, we propose the following modification of (4):

Problem 2 (Group lasso SPCA: GL-SPCA). *Given an arbitrary matrix $\mathbf{X}\in\mathbb{R}^{n\times p}$ and an integer k, let $\mathbf{A}$ and $\mathbf{W}$ be $p\times k$ matrices, and let $\lambda,\lambda_1>0$. The GL-SPCA problem is to solve*

$$(\mathbf{A}^*,\mathbf{V}^*) = \text{argmin}_{\mathbf{A},\mathbf{W}}||\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{A}^T||_F^2 + \lambda||\mathbf{W}||_F^2 + \lambda_1\sum_{i=1}^{p}||\mathbf{W}_{(i)}||_2 \ \text{ s.t. } \mathbf{A}^T\mathbf{A}=\mathbf{I}_k. \quad (11)$$

Thus, the lasso penalty $\lambda_1||\mathbf{W}||_1$ in (5) is replaced in (11) by a group lasso penalty $\lambda_1\sum_{i=1}^{p}||\mathbf{W}_{(i)}||_2$, where rows of $\mathbf{W}$ are grouped together so that each row of $\mathbf{V}^*$ will tend to be either dense or entirely zero.

Importantly, the GL-SPCA problem is not convex in $\mathbf{W}$ and $\mathbf{A}$ together; it is, however, convex in $\mathbf{W}$, and it is easy to solve in $\mathbf{A}$. Thus, analogous to the treatment in Zou *et al.* [21], we propose an iterative alternate-minimization algorithm to solve GL-SPCA. This is described in Algorithm 1; and the justification of this algorithm is given in Section 7. Note that if we fix $\mathbf{A}$ to be $\mathbf{I}$ throughout, then Algorithm 1 can be used to solve the GL-REG problem discussed in Section 3.

Algorithm 1: Iterative algorithm for solving the GL-SPCA (and GL-REG) problems.
(For the GL-REG problem, fix $\mathbf{A} = \mathbf{I}$ throughout this algorithm.)

Input: Data matrix $\mathbf{X}$ and initial estimates for $\mathbf{A}$ and $\mathbf{W}$
Output: Final estimates for $\mathbf{A}$ and $\mathbf{W}$
repeat
1 Compute SVD of $\mathbf{X}^T\mathbf{X}\mathbf{W}$ as $\mathbf{U}\mathbf{D}\mathbf{V}^T$ and then $\mathbf{A} \leftarrow \mathbf{U}\mathbf{V}^T$;
$\mathcal{S} \leftarrow \{i : \ ||\mathbf{W}_{(i)}||_2 \neq 0\}$;
for $i \in \mathcal{S}$ **do**
2 Compute $\mathbf{b}_i = \sum_{j \neq i} \left(\mathbf{X}^{(j)T}\mathbf{X}^{(i)}\right) \mathbf{W}_{(j)}^T$;
if $||\mathbf{A}^T\mathbf{X}^T\mathbf{X}^{(i)} - \mathbf{b}_i||_2 \leq \lambda_1/2$ **then**
3 $\mathbf{W}_{(i)}^T \leftarrow \mathbf{0}$;
else
4 $\mathbf{W}_{(i)}^T \leftarrow \frac{2}{2||\mathbf{X}^{(i)}||_2^2 + \lambda + \lambda_1/||\mathbf{W}_{(i)}||_2} \left(\mathbf{A}^T\mathbf{X}^T\mathbf{X}^{(i)} - \mathbf{b}_i\right)$;
until *convergence*;

We remark that such row-sparsity in $\mathbf{V}^*$ can have either advantages or disadvantages. Consider, for example, when there are a small number of informative columns in $\mathbf{X}$ and the rest are not important for the task at hand [12, 14]. In such a case, we would expect that enforcing entire rows to be zero would lead to better identification of the signal columns; and this has been empirically observed in the application of CUR to DNA SNP analysis [14]. The unstructured $\mathbf{V}^*$, by contrast, would not be able to "borrow strength" across all columns of $\mathbf{V}^*$ to differentiate the signal columns from the noise columns. On the other hand, requiring such structured sparsity is more restrictive and may not be desirable. For example, in microarray analysis in which we have measured p genes on n patients, our goal may be to find several underlying factors. Biologists have identified "pathways" of interconnected genes [16], and it would be desirable if each sparse factor could be identified with a different pathway (that is, a different set of genes). Requiring all factors of $\mathbf{V}^*$ to exclude the same $p - c$ genes does not allow a different sparse subset of genes to be active in each factor.

We finish this section by pointing out that while most SPCA methods only enforce unstructured zeros in $\mathbf{V}^*$, the idea of having a structured sparsity in the PCA context has very recently been explored [9]. Our GL-SPCA problem falls within the broad framework of this idea.

6 Empirical Comparisons

In this section, we evaluate the performance of the four methods discussed above on both synthetic and real data. In particular, we compare the randomized CUR algorithm of Mahoney and Drineas [12, 4] to our GL-REG (of Problem 1), and we compare the SPCA algorithm proposed by Zou *et al.* [21] to our GL-SPCA (of Problem 2). We have also compared against the SPCA algorithm of Witten *et al.* [19], and we found the results to be very similar to those of Zou *et al.*

6.1 Simulations

We first consider synthetic examples of the form $\mathbf{X} = \widehat{\mathbf{X}} + \mathbf{E}$, where $\widehat{\mathbf{X}}$ is the underlying signal matrix and $\mathbf{E}$ is a matrix of noise. In all our simulations, $\mathbf{E}$ has i.i.d. $\mathcal{N}(0, 1)$ entries, while the signal $\widehat{\mathbf{X}}$ has one of the following forms:

Case I) $\widehat{\mathbf{X}} = [\mathbf{0}_{n\times(p-c)}; \widehat{\mathbf{X}}^*]$ where the $n \times c$ matrix $\widehat{\mathbf{X}}^*$ is the nonzero part of $\widehat{\mathbf{X}}$. In other words, $\widehat{\mathbf{X}}$ has c nonzero columns and does not necessarily have a low-rank structure.

Case II) $\widehat{\mathbf{X}} = \mathbf{U}\mathbf{V}^T$ where $\mathbf{U}$ and $\mathbf{V}$ each consist of $k < p$ orthogonal columns. In addition to being low-rank, $\mathbf{V}$ has entire rows equal to zero (*i.e.* it is row-sparse).

Case III) $\widehat{\mathbf{X}} = \mathbf{U}\mathbf{V}^T$ where $\mathbf{U}$ and $\mathbf{V}$ each consist of $k < p$ orthogonal columns. Here $\mathbf{V}$ is low-rank and sparse, but the sparsity is not structured (*i.e.* it is scattered-sparse).

A successful method attains low reconstruction error of the true signal $\widehat{\mathbf{X}}$ and has high precision in identifying correctly the zeros in the underlying model. As previously discussed, the four methods optimize for different types of reconstruction error. Thus, in comparing CUR and GL-REG, we use the regression-type reconstruction error $\mathrm{ERR}_{\mathrm{reg}}(\mathcal{I}) = ||\widehat{\mathbf{X}} - \mathbf{X}^{\mathcal{I}}\mathbf{X}^{\mathcal{I}+}\mathbf{X}||_F$, whereas for the comparison of SPCA and GL-SPCA, we use the PCA-type error $\mathrm{ERR}(\mathbf{V}) = ||\widehat{\mathbf{X}} - \mathbf{X}\mathbf{V}\mathbf{V}^{+}||_F$.

Table 1 presents the simulation results from the three cases. All comparisons use $n = 100$ and $p = 1000$. In Case II and III, the signal matrix has rank $k = 10$. The underlying sparsity level is 20%, *i.e.* 80% of the entries of $\widehat{\mathbf{X}}$ (Case I) and $\mathbf{V}$ (Case II&III) are zeros. Note that all methods except for GL-REG require the rank k as an input, and we always take it to be 10 even in Case I. For easy comparison, we have tuned each method to have the correct total number of zeros. The results are averaged over 5 trials.

	Methods	Case I	Case II	Case III
$\mathrm{ERR}_{\mathrm{reg}}(\mathcal{I})$	CUR	316.29 (0.835)	315.28 (0.797)	315.64 (0.166)
	GL-REG	316.29 (0.989)	315.28 (0.750)	315.64 (0.107)
$\mathrm{ERR}(\mathbf{V})$	SPCA	177.92 (0.809)	44.388 (0.799)	44.995 (0.792)
	GL-SPCA	141.85 (0.998)	37.310 (0.767)	45.500 (0.804)

Table 1: Simulation results: The reconstruction errors and the percentages of correctly identified zeros (in parentheses).

We notice in Table 1 that the two regression-type methods CUR and GL-REG have very similar performance. As we would expect, since CUR only uses information in the top k singular vectors, it does slightly worse than GL-REG in terms of precision when the underlying signal is not low-rank (Case I). In addition, both methods perform poorly if the sparsity is not structured as in Case III. The two PCA-type methods perform similarly as well. Again, the group lasso method seems to work better in Case I. We note that the precisions reported here are based on element-wise sparsity—if we were measuring row-sparsity, methods like SPCA would perform poorly since they do not encourage entire rows to be zero.

6.2 Microarray example

We next consider a microarray dataset of soft tissue tumors studied by Nielsen *et al.* [13]. Mahoney and Drineas [12] apply CUR to this dataset of $n = 31$ tissue samples and $p = 5520$ genes. As with the simulation results, we use two sets of comparisons: we compare CUR with GL-REG, and we compare SPCA with GL-SPCA. Since we do not observe the underlying truth $\widehat{\mathbf{X}}$, we take $\mathrm{ERR}_{\mathrm{reg}}(\mathcal{I}) = ||\mathbf{X} - \mathbf{X}^{\mathcal{I}}\mathbf{X}^{\mathcal{I}+}\mathbf{X}||_F$ and $\mathrm{ERR}(\mathbf{V}) = ||\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^{+}||_F$. Also, since we do not observe the true sparsity, we cannot measure the precision as we do in Table 1. The left plot in Figure 2 shows $\mathrm{ERR}_{\mathrm{reg}}(\mathcal{I})$ as a function of $|\mathcal{I}|$. We see that CUR and GL-REG perform similarly. (However, since CUR is a randomized algorithm, on every run it gives a different result. From a practical standpoint, this feature of CUR can be disconcerting to biologists wanting to report a single set of important genes. In this light, GL-REG may be thought of as an attractive non-randomized alternative to CUR.) The right plot of Figure 2 compares GL-SPCA to SPCA (specifically, Zou *et al.* [21]). Since SPCA does not explicitly enforce row-sparsity, for a gene to be not used in the model requires *all* of the ($k = 4$) columns of $\mathbf{V}^*$ to exclude it. This likely explains the advantage of GL-SPCA over SPCA seen in the figure.

7 Justification of Algorithm 1

The algorithm alternates between minimizing with respect to $\mathbf{A}$ and $\mathbf{B}$ until convergence.

Solving for $\mathbf{A}$ given $\mathbf{B}$: If $\mathbf{B}$ is fixed, then the regularization penalty in (11) can be ignored, in which case the optimization problem becomes $\min_{\mathbf{A}} ||\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^T||_F^2$ subject to $\mathbf{A}^T\mathbf{A} = I$. This problem was considered by Zou *et al.* [21], who showed that the solution is obtained by computing the SVD of $(\mathbf{X}^T\mathbf{X})\mathbf{B}$ as $(\mathbf{X}^T\mathbf{X})\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ and then setting $\widehat{\mathbf{A}} = \mathbf{U}\mathbf{V}^T$. This explains step 1 in Algorithm 1.

Solving for $\mathbf{B}$ given $\mathbf{A}$: If $\mathbf{A}$ is fixed, then (11) becomes an unconstrained convex optimization problem in $\mathbf{B}$. The subgradient equations (using that $\mathbf{A}^T\mathbf{A} = \mathbf{I}_k$) are

$$2\mathbf{B}^T\mathbf{X}^T\mathbf{X}^{(i)} - 2\mathbf{A}^T\mathbf{X}^T\mathbf{X}^{(i)} + 2\lambda\mathbf{B}_{(i)}^T + \lambda_1\mathbf{s}_i = \mathbf{0}; \quad i = 1, \ldots, p, \tag{12}$$

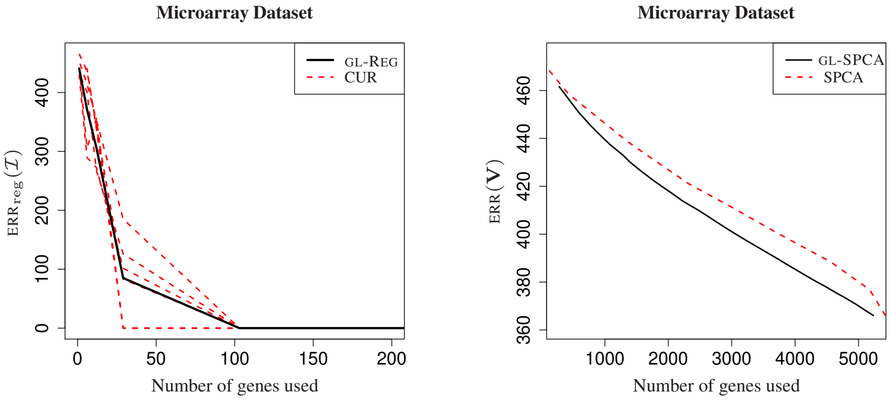

Figure 2: Left: Comparison of CUR, multiple runs, with GL-REG; Right: Comparison of GL-SPCA with SPCA (specifically, Zou *et al.* [21]).

where the subgradient vectors $\mathbf{s}_i = \mathbf{B}_{(i)}^T/||\mathbf{B}_{(i)}||_2$ if $\mathbf{B}_{(i)} \neq \mathbf{0}$, or $||\mathbf{s}_i||_2 \leq 1$ if $\mathbf{B}_{(i)} = \mathbf{0}$. Let us define $\mathbf{b}_i = \sum_{j \neq i}(\mathbf{X}^{(j)T}\mathbf{X}^{(i)})\mathbf{B}_{(j)}^T = \mathbf{B}^T\mathbf{X}^T\mathbf{X}^{(i)} - ||\mathbf{X}^{(i)}||_2^2\mathbf{B}_{(i)}^T$, so that the subgradient equations can be written as

$$\mathbf{b}_i + (||\mathbf{X}^{(i)}||_2^2 + \lambda)\mathbf{B}_{(i)}^T - \mathbf{A}^T\mathbf{X}^T\mathbf{X}^{(i)} + (\lambda_1/2)\mathbf{s}_i = \mathbf{0}. \tag{13}$$

The following claim explains Step 3 in Algorithm 1.

Claim 1. $\mathbf{B}_{(i)} = \mathbf{0}$ *if and only if* $||\mathbf{A}^T\mathbf{X}^T\mathbf{X}^{(i)} - \mathbf{b}_i||_2 \leq \lambda_1/2$.

Proof. First, if $\mathbf{B}_{(i)} = \mathbf{0}$, the subgradient equations (13) become $\mathbf{b}_i - \mathbf{A}^T\mathbf{X}^T\mathbf{X}^{(i)} + (\lambda_1/2)\mathbf{s}_i = \mathbf{0}$. Since $||\mathbf{s}_i||_2 \leq 1$ if $\mathbf{B}_{(i)} = \mathbf{0}$, we have $||\mathbf{A}^T\mathbf{X}^T\mathbf{X}^{(i)} - \mathbf{b}_i||_2 \leq \lambda_1/2$. To prove the other direction, recall that $\mathbf{B}_{(i)} \neq \mathbf{0}$ implies $\mathbf{s}_i = \mathbf{B}_{(i)}^T/||\mathbf{B}_{(i)}||_2$. Substituting this expression into (13), rearranging terms, and taking the norm on both sides, we get $2||\mathbf{A}^T\mathbf{X}^T\mathbf{X}^{(i)} - \mathbf{b}_i||_2 = \left(2||\mathbf{X}^{(i)}||_2^2 + 2\lambda + \lambda_1/||\mathbf{B}_{(i)}||_2\right)||\mathbf{B}_{(i)}||_2 > \lambda_1$. □

By Claim 1, $||\mathbf{A}^T\mathbf{X}^T\mathbf{X}^{(i)} - \mathbf{b}_i||_2 > \lambda_1/2$ implies that $\mathbf{B}_{(i)} \neq 0$ which further implies $\mathbf{s}_i = \mathbf{B}_{(i)}^T/||\mathbf{B}_{(i)}||_2$. Substituting into (13) gives Step 4 in Algorithm 1.

8 Conclusion

In this paper, we have elucidated several connections between two recently-popular matrix decomposition methods that adopt very different perspectives on obtaining interpretable low-rank matrix decompositions. In doing so, we have suggested two optimization problems, GL-REG and GL-SPCA, that highlight similarities and differences between the two methods. In general, SPCA methods obtain interpretability by modifying an existing intractable objective with a convex regularization term that encourages sparsity, and then *exactly* optimizing that modified objective. On the other hand, CUR methods operate by using randomness and approximation as computational resources to optimize *approximately* an intractable objective, thereby implicitly incorporating a form of regularization into the steps of the approximation algorithm. Understanding this concept of *implicit regularization via approximate computation* is clearly of interest more generally, in particular for applications where the size scale of the data is expected to increase.

Acknowledgments

We would like to thank Art Owen and Robert Tibshirani for encouragement and helpful suggestions. Jacob Bien was supported by the Urbanek Family Stanford Graduate Fellowship, and Ya Xu was supported by the Melvin and Joan Lane Stanford Graduate Fellowship. In addition, support from the NSF and AFOSR is gratefully acknowledged.

References

[1] M.-A. Belabbas and P.J. Wolfe. Fast low-rank approximation for covariance matrices. In *Second IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pages 293–296, 2007.

[2] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007.

[3] P. Drineas, R. Kannan, and M.W. Mahoney. Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition. *SIAM Journal on Computing*, 36:184–206, 2006.

[4] P. Drineas, M.W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30:844–881, 2008.

[5] S.A. Goreinov and E.E. Tyrtyshnikov. The maximum-volume concept in approximation by low-rank matrices. *Contemporary Mathematics*, 280:47–51, 2001.

[6] T. Hastie, R. Tibshirani, and J. Friedman. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Manuscript. Submitted. 2010.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, 2003.

[8] D.C. Hoaglin and R.E. Welsch. The hat matrix in regression and ANOVA. *The American Statistician*, 32(1):17–22, 1978.

[9] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. Technical report. Preprint: arXiv:0909.1440 (2009).

[10] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

[11] S. Kumar, M. Mohri, and A. Talwalkar. Ensemble Nyström method. In *Annual Advances in Neural Information Processing Systems 22: Proceedings of the 2009 Conference*, 2009.

[12] M.W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. USA*, 106:697–702, 2009.

[13] T. Nielsen, R.B. West, S.C. Linn, O. Alter, M.A. Knowling, J. O'Connell, S. Zhu, M. Fero, G. Sherlock, J.R. Pollack, P.O. Brown, D. Botstein, and M. van de Rijn. Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet*, 359(9314):1301–1307, 2002.

[14] P. Paschou, E. Ziv, E.G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M.W. Mahoney, and P. Drineas. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics*, 3:1672–1686, 2007.

[15] J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104:735–746, 2009.

[16] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, 102(43):15545–15550, 2005.

[17] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos. Less is more: Compact matrix decomposition for large sparse graphs. In *Proceedings of the 7th SIAM International Conference on Data Mining*, 2007.

[18] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

[19] D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[20] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

[21] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):262–286, 2006.