SPARSE MATRIX TRANSFORM FOR FAST PROJECTION TO REDUCED DIMENSION

James Theiler[1]*, Guangzhi Cao*[2]*,* and *Charles A. Bouman*[3]

[1]Los Alamos National Laboratory, Space and Remote Sensing Sciences, Los Alamos, NM 87545
[2]GE Healthcare Technologies, 3000 N Grandview Blvd, W-1180, Waukesha, WI 53188
[3]Purdue University, School of Electrical and Computer Engineering, West Lafayette, IN 47907

ABSTRACT

We investigate three algorithms that use the sparse matrix transform (SMT) to produce variance-maximizing linear projections to a lower-dimensional space. The SMT expresses the projection as a sequence of Givens rotations and this enables computationally efficient implementation of the projection operator. The baseline algorithm uses the SMT to directly approximate the optimal solution that is given by principal components analysis (PCA). A variant of the baseline begins with a standard SMT solution, but prunes the sequence of Givens rotations to only include those that contribute to the variance maximization. Finally, a simpler and faster third algorithm is introduced; this also estimates the projection operator with a sequence of Givens rotations, but in this case, the rotations are chosen to optimize a criterion that more directly expresses the dimension reduction criterion.

1. INTRODUCTION

For a variety of remote sensing detection problems, the covariance matrix is a key statistical quantity for characterizing the variability of the data. Particularly for high-dimensional data (*e.g.,* hyperspectral imagery), it provides a concise characterization of the data distribution that includes *all* pairwise correlations between the spectral bands. For this reason, it remains a cornerstone in the statistical analysis of remote sensing data.

Often this statistical analysis benefits from a projection of the high-dimensional data to a lower-dimensional subspace. Principal components analysis provides the classical solution to this problem; it is an optimal solution, but it requires an accurate estimate of the covariance matrix and it provides a dense projection operator.

Particularly when there are limitations on the number of samples available for estimating the covariance matrix, the sample covariance can over-fit the actual covariance, and lead to reduced performance for detection and regression algorithms that employ the covariance matrix. For this reason, various kinds of regularization have been introduced [1, 2, 3, 4, 5], including, very recently, regularization based on the sparse matrix transform (SMT) [6, 7, 8]. What is "sparse" about the SMT is the number of operations required to express the eigenvector matrix of the covariance matrix. Thus, in addition to its value as a regularizer, the SMT also provides a computationally efficient implementation of signal processing operations that involve the covariance matrix. Among these is the problem of variance-maximizing dimension reduction.

Although the SMT was initially developed as a *regularized* covariance estimate, we will neglect the regularization aspects of the SMT for this paper, and instead will concentrate on the implementation efficiency of three different SMT-based approximations to the covariance matrix for the purpose of fast projection to reduced dimension.

In Section 2, we describe the dimension reduction problem, and in Subsections 2.1, 2.2, and 2.3, we describe three SMT-based approaches to solving this problem. In Section 3, we illustrate these concepts with numerical results applied to hyperspectral imagery, and finally in Section 4, we conclude.

2. DIMENSION REDUCTION

A linear projection $E_q \in \mathbb{R}^{p\times q}$ maps data from $\mathbf{x} \in \mathbb{R}^p$ to $\mathbf{y} = E_q^T\mathbf{x} \in \mathbb{R}^q$ with $q < p$. The projector does not introduce any expansion or contraction, and is constrained to satisfy $E_q^T E_q = I$.

We can write the variance of the projected signal as $\langle \mathbf{y}^T\mathbf{y} \rangle$, where the angle brackets indicate an average over samples (in the case of hyperspectral imagery, the samples are pixels). In particular, we can express this variance as

$$\begin{aligned}\langle \mathbf{y}^T\mathbf{y} \rangle &= \text{trace}\big(\langle \mathbf{y}\mathbf{y}^T \rangle\big) = \text{trace}\big(E_q^T \langle \mathbf{x}\mathbf{x}^T \rangle E_q\big)\\ &= \text{trace}\big(E_q^T R E_q\big) \qquad (1)\end{aligned}$$

where $R = \langle \mathbf{x}\mathbf{x}^T \rangle$ is the covariance matrix associated with the data $\mathbf{x}$. (We have assumed that mean values have been subtracted from the data, so that $\langle \mathbf{x} \rangle = 0$.)

If we want a variance-maximizing projection, then we can express this as a direct optimization problem:

$$E_q = \text{argmax}_{E_q\in\Omega_q} \text{trace}\big(E_q^T R E_q\big)\,, \qquad (2)$$

JT was supported by the Laboratory Directed Research and Development (LDRD) program at Los Alamos National Laboratory.

CAB was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number 56541-CI.

1. Input covariance matrix R, number of rotations K
2. Let $S = R$
3. For $k = 1 \dots K$,
   a. Find pair i, j to maximize $s_{ij}^2/(s_{ii}s_{jj})$
   b. Compute angle $\theta = \frac{1}{2}\text{atan}\,(-2s_{ij}, s_{ii} - s_{jj})$
   c. Let $G_k = I + \Theta(i, j, \theta)$
   d. Apply the Givens rotation: $S \leftarrow G_k^T S G_k$
4. Let $E = G_1 G_2 \cdots G_K$
5. Let $\Lambda = \text{diag}(S)$

Fig. 1. Pseudocode for standard SMT

1. Input R, K, and q
2. Use standard SMT to obtain Λ and E
3. Let $\mathcal{I} = \{i_1, \dots, i_q\}$ index the q largest values of Λ
4. Let E_q be the q columns of E represented by $\mathcal{I}$

Fig. 2. Standard SMT for dimension reduction

where Ω_q is the set of suborthogonal matrices $E_q \in \mathbb{R}^{p\times q}$ satisfying $E_q^T E_q = I$.

An optimal solution (but not a *unique* optimal solution) to the dimension reduction problem is given by principal components analysis (PCA). Write the covariance matrix $R = E\Lambda E^T$, where $E^T E = I$ and Λ is a diagonal matrix with non-negative entries. Write $E_q = EH_q$, where $H_q \in \mathbb{R}^{p\times q}$ is a matrix of ones and zeros that extracts the q columns of E that are associated with the largest values in Λ. Then, this E_q solves the optimization in Eq. (2).

We can cast PCA as an optimization problem. If $R = E\Lambda E^T$, then $\Lambda = E^T R E$ is a diagonal matrix. We can use the fact (*e.g.,* see Eq. 58 of Ref. [8]) that $|\text{diag}(S)| \geq |S|$ to write

$$\left|\text{diag}\left(\hat{E}^T R \hat{E}\right)\right| \geq \left|\hat{E}^T R \hat{E}\right| = |R| = |\Lambda| \qquad (3)$$

with equality holding only when $\hat{E}$ is an eigenvector matrix for R. Thus, the eigenvector matrix is the solution to the optimization problem

$$E = \text{argmin}_{E\in\Omega} \left|\text{diag}\left(E^T R E\right)\right| \qquad (4)$$

where Ω is the set of orthogonal matrices $E \in \mathbb{R}^{p\times p}$ satisfying $E^T E = I$.

Note that this result can also be derived in terms of a maximum likelihood estimate for a Gaussian distribution [6], but here we want to emphasize that we are maximizing variance without making any assumptions about underlying distribution.

1. Input $G_1, G_2, \dots, G_K$; $\{i_1, \dots, i_q\}$; and $\mathbf{x} \in \mathbb{R}^p$.
2. For $k = 1 \dots K$,
   a. Want to compute $\mathbf{x} \leftarrow G_k^T \mathbf{x} = [I + \Theta(i, j, \theta)]^T \mathbf{x}$ (Only the i and j components of $\mathbf{x}$ will be altered.)
   b. Let $\begin{bmatrix} x_i \\ x_j \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_i \\ x_j \end{bmatrix}$
3. Set $y_k = x_{i_k}$ for $k = 1 \dots q$.
4. Return $\mathbf{y} \in \mathbb{R}^q$

Fig. 3. Application of Givens rotations to reduce dimension; shown here is the straightforward implementation that uses four multiplications per rotation. A streamlined variant of this algorithm [8, Appendix B] provides updates with only two multiplications per rotation, followed a final step with q multiplications. Thus, only $2K + q$ (instead of $4K$) multiplications are needed. Recognizing that we don't always need both x_i and x_j, it is possible that the number of multiplications may be reduced even further.

But the PCA solution is not the *only* optimal solution. Any q vectors that span the same space as the q columns of $E_q = EH_q$ will be an optimal solution. Thus, while minimizing $\left|\text{diag}\left(E^T R E\right)\right|$ does provide an optimal solution to the dimension reduction problem, it is an overly restrictive condition. Put another way, PCA solves a more general problem than is actually needed.

2.1. Sparse Matrix Transform (SMT)

The most sparse nontrivial orthogonal transform is the *Givens rotation*, which corresponds to a rotation by an angle θ in the plane of the i and j axes; specifically, it is given by $G = I + \Theta(i, j, \theta)$ where

$$\Theta(i, j, \theta)_{rs} = \begin{cases} \cos(\theta) - 1 & \text{if } r = s = i \text{ or } r = s = j \\ \sin(\theta) & \text{if } r = i \text{ and } s = j \\ -\sin(\theta) & \text{if } r = j \text{ and } s = i \\ 0 & \text{otherwise.} \end{cases} \qquad (5)$$

The Sparse Matrix Transform (SMT) expresses the eigenvector matrix E as a product of Givens rotations; restricting the number of these rotations provides an approximation that achieves two purposes. One is a regularization to resist overfitting; two is a reduction in the computation needed to do signal processing with the covariance matrix.

Let G_k denote a Givens rotation, and note that a product of orthogonal rotations $G_1 G_2 \cdots G_K$ is still orthogonal. Let Ω_K be the set of orthogonal matrices that can be expressed as a product of K Givens rotations. The SMT covariance estimate is then given by $\hat{R} = \hat{E}\hat{\Lambda}\hat{E}^T$ where E is given by Eq. (4) with $\Omega = \Omega_K$ and $\hat{\Lambda} = \text{diag}\left(\hat{E}^T R \hat{E}\right)$.

1. Input R, K, and q
2. Use standard SMT to obtain Λ and $G_1, \ldots, G_K$
3. Let $\mathcal{I} = \mathcal{I}_o = \{i_1, \ldots, i_q\}$ index the q largest values of Λ
4. Let $\mathcal{J} = \{i_{q+1}, \ldots, i_p\}$ be the remaining $p - q$ indices.
5. For $k = K, \ldots, 1$,
   a. Let i, j be the axes associated with G_k
   b. If ($i \in \mathcal{I}$ and $j \in \mathcal{J}$) or ($i \in \mathcal{J}$ and $j \in \mathcal{I}$), then
      - Keep G_k
      - Add $\mathcal{I} \leftarrow \mathcal{I} \cup \{i, j\}$
      - Add $\mathcal{J} \leftarrow \mathcal{J} \cup \{i, j\}$

      Else discard G_k
6. Using only the G_k that were not discarded, recompute $E = G_1 G_2 \ldots G_{K'}$ where $K' \leq K$.
7. Let E_q be the q columns of E indexed by the initial $\mathcal{I}_o$

Fig. 4. Standard SMT for dimension reduction, with pruning

Fig. 1 illustrates the steps needed to generate the standard SMT estimate of a covariance matrix. And Fig. 2 shows how to use E and Λ to achieve dimension reduction. But when E is expressed as a product of Givens rotations, then the computation for applying the dimension reduction operator can be substantially decreased. To compute $\mathbf{y} = E_q \mathbf{x}$ for general E_q requires $O(pq)$ multiplications and additions. But when $E_q = G_1 G_2 \cdots G_K H_q$, then each Givens rotation requires only $O(1)$ multiplications and additions; so that $O(K)$ operations are required. Typically, $K = O(p)$ (e.g., see [9]), so if $q \gg 1$, this can be a significant gain. Fig. 3 shows how the application is done.

In the situation where p is so large that an actual $p \times p$ covariance matrix is never explicitly computed, but instead is characterized in terms of an SMT representation [9], then the algorithm in Fig. 3 can still be used.

2.2. Pruning

Having collected the Givens rotation matrices, one can recognize that not all of them contribute to the ultimate dimension reduction criterion. For instance if the last rotation G_K rotates channels i and j, and both i and j are among the top q channels, then it is simply "moving variance around" within those top channels, and not adding to the total variance in the top q channels. Similarly, if neither i nor j is among the top q, then the rotation will not affect trace$\left(E_q^T R E_q\right)$. This argument only applies to the last rotation, however. The effect of earlier rotations depends on later rotations, so one must work backwards to determine the influence of each Givens rotation. This process is illustrated in Fig. 4.

1. Input: R, K, and q
2. Let $\mathcal{I} = \{i_1, \ldots, i_q\}$ index the q largest values of diag(R)
3. Let $S = R$.
4. For $k = 1 \ldots K$,
   a. Considering only $i \in \mathcal{I}$ and $j \notin \mathcal{I}$, find the pair i, j for which $[(s_{ii} - s_{jj})^2 + 4s_{ij}^2]^{1/2} - [s_{ii} - s_{jj}]$ is maximum.
   b. Compute angle $\theta = \frac{1}{2}\text{atan}\left(-2s_{ij}, s_{ii} - s_{jj}\right)$
   c. Set $G_k = I + \Theta(i, j, \theta)$
   d. Apply the Givens rotation: $S \leftarrow G_k^T S G_k$
5. Let $E = G_1 G_2 \ldots G_K$
6. Let E_q be the q columns of E represented by $\mathcal{I}$

Fig. 5. Pseudocode for SMT-DR

2.3. Direct dimension reduction (SMT-DR)

The use of standard SMT, with or without pruning, requires the optimization of Eq. (4) which is both more complicated and more restrictive than the trace criterion in Eq. (2) that directly describes what it is we want to optimize. With this in mind, we developed a third approach to dimension reduction by effectively replacing Eq. (4) with Eq. (2) in the core of the SMT algorithm. This amounts to choosing axis pairs i, j for which rotation maximally increases Eq. (2); that is,

$$[(s_{ii} - s_{jj})^2 + 4s_{ij}^2]^{1/2} - [s_{ii} - s_{jj}] \tag{6}$$

instead of $s_{ij}^2/(s_{ii}s_{jj})$. The details are shown in Fig. 5.

One practical advantage of this approach is that the high-variance coordinates $\{i_1, \ldots, i_q\}$ remain the same throughout the computation. By contrast, they are identified after the fact when the standard SMT algorithm is employed (see Fig. 2).

3. NUMERICAL ILLUSTRATION

Because this approach is designed for the large p regime, we will consider a hyperspectral image with extended dimensionality induced by spatial operators. Consider the AVIRIS [10] image `f960323t01p02_r04_sc01` of the Florida coastline. This is a 224-channel image, but we will augment that by smoothing the image with kernels of radius one and two; this leads to a $p = 672$ channel image. Fig. 6 shows the projection of this data to a $q = 5$ dimension space. If V is the variance exhibited by the projected data, and if V_o is the variance exhibited by all p channels, then $V_o - V$ is the missing variance, and $(V_o - V)/V_o$ is the relative missing variance. If V_q is the variance exhibited by the q top principal components, then $(V_o - V_q)/V_o$ is a lower bound on the relative missing variance that can be exhibited by a dimension reduction scheme.

What we see in Fig. 6 is that for small K, SMT-DR gets more variance (lower missing variance) with fewer rotations than standard SMT, though that advantage is matched by SMT with pruning. For larger K (in this case, the turnaround point is near $K = 600$), SMT overtakes SMT-DR. In this regime, SMT with pruning gets the best performance.

4. DISCUSSION

Certainly trace$\left(E_q^T R E_q\right)$ is an appropriate criterion, in the sense that if it is optimized, the dimension reduction problem is solved. But minimizing $\left|\text{diag}\left(E^T R E\right)\right|$ and taking E_q as the q columns of E associated with the largest values of diag$\left(E^T R E\right)$, also produces an optimum. The trace condition is less stringent (maximizing trace in Eq. (2) does *not* minimize determinant, but minimizing the determinant in Eq. (4) *will* maximize the trace). Because the trace condition is less stringent, one can imagine that it is somehow easier (specifically, that it can be done with fewer Givens rotations). To some extent, this is reflected in the numerical experiment, which sees SMT-DR achieving better performance than standard SMT for small K. But this advantage is lost for larger K, and in any case, it is retrieved when standard SMT is followed by pruning.

The main advantage of standard SMT is flexibility. One need not choose q beforehand, and one can always prune the result to get a smaller K afterwards.

SMT-DR has some advantages. It is simple to implement and cheap to train. It provides good performance at small K, and the set of top indices $\mathcal{I}$ does not change throughout the computation.

But SMT followed by pruning is a strong competitor. It takes a little longer than SMT-DR to train, though this is only in the small K case, when training is cheap anyway. It gives performance that is as good as or equal to both SMT-DR and standard SMT, over a wide range of K values.

We remark that the alternative SMT implementation in Ref [9] which can be applied for extremely large p, permits both SMT-DR and SMT with pruning.

5. REFERENCES

[1] J. H. Friedman, "Regularized discriminant analysis," *J. Am. Statistical Assoc.*, vol. 84, pp. 165–175, 1989.

[2] C. Lee and D. A. Landgrebe, "Analyzing high-dimensional multispectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 31, pp. 792–800, 1993.

[3] J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, pp. 763–767, 1996.

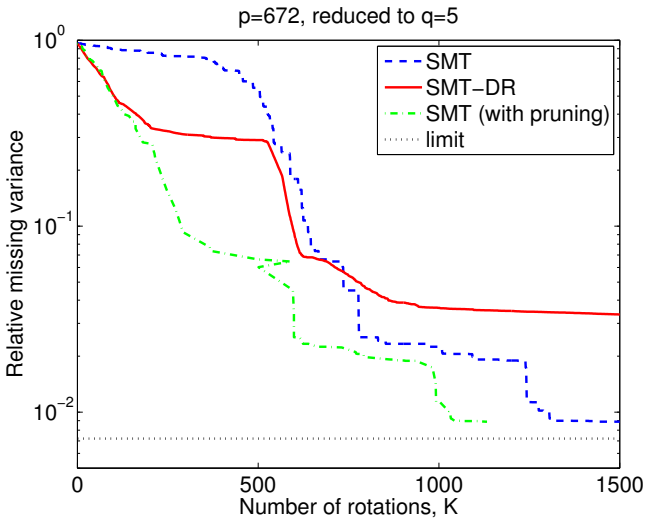

Fig. 6. Relative missing variance as a function of the number K of rotations.

[4] M. J. Daniels and R. E. Kass, "Shrinkage estimators for covariance matrices," *Biometrics*, vol. 57, no. 4, pp. 1173–1184, 2001.

[5] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 32, 2005.

[6] G. Cao and C. A. Bouman, "Covariance estimation for high dimensional data vectors using the sparse matrix transform," in *Advances in Neural Information Processing Systems 21*. 2009, pp. 225–232, MIT Press.

[7] G. Cao, C. A. Bouman, and J. Theiler, "Weak signal detection in hyperspectral imagery using sparse matrix transform (SMT) covariance estimation," in *Proc. WHISPERS (Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing)*. 2009, IEEE.

[8] G. Cao, C. A. Bouman, and K. J. Webb, "Noniterative MAP reconstruction using sparse matrix representations," *IEEE Trans. Image Processing*, vol. 18, pp. 2085–2099, 2009.

[9] L. R. Bachega, G. Cao, and C. A. Bouman, "Fast signal analysis and decomposition on graphs using the sparse matrix transform," *Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 5426 – 5429, 2010.

[10] G. Vane, R. O. Green, T. G. Chrien, H. T. Enmark, E. G. Hansen, and W. M. Porter, "The Airborne Visible/Infrared Imaging Spectrometer (AVIRIS)," *Remote Sensing of the Environment*, vol. 44, pp. 127–143, 1993.