HIGH DIMENSIONAL REGRESSION USING THE SPARSE MATRIX TRANSFORM (SMT)

Guangzhi Cao

GE Healthcare Technologies
3000 N. Grandview Blvd, W-1180
Waukesha, WI 53188

*Yandong Guo, and Charles A. Bouman**

School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907

ABSTRACT

Regression from high dimensional observation vectors is particularly difficult when training data is limited. More specifically, if the number of sample vectors n is less than dimension of the sample vectors p, then accurate regression is difficult to perform without prior knowledge of the data covariance.

In this paper, we propose a novel approach to high dimensional regression for application when $n < p$. The approach works by first decorrelating the high dimensional observation vector using the sparse matrix transform (SMT) estimate of the data covariance. Then the decorrelated observations are used in a regularized regression procedure such as Lasso or shrinkage. Numerical results demonstrate that the proposed regression approach can significantly improve the prediction accuracy, especially when n is small and the signal to be predicted lies in the subspace of the observations corresponding to the small eigenvalues.

***Index Terms*—** High dimensional regression, covariance estimation, sparse matrix transform

1. INTRODUCTION

Regression from high dimensional observation vectors is particularly difficult when training data is limited. Traditional regression methods that use the sample covariance, such as the ordinary least squares (OLS) approach, perform poorly in this situation. This is because, if the sample number n is less than the data dimension p, then the sample covariance is singular, with at least $p - n$ of the smallest eigenvalues estimated to be zero. In this case, the sample covariance does not accurately characterize any signal that falls in the subspaces corresponding to the smallest eigenvalues of the observations.

In the past decades, regression methods that adopt regularization have been introduced, such as ridge regression [1], subset selection, and Lasso [2]. More recently, there has also been increasing interest in replacing the sample covariance with some sparse estimates of the true covariance or its inverse for high dimensional regression problems [3, 4].

*This work was supported by the National Science Foundation under Contract CCR-0431024 and the Army Research Offi ce.

In this paper, we propose a novel regression approach that first decorrelates the high dimensional observation vector using the sparse matrix transform (SMT) estimate of the covariance [5]. To improve the prediction accuracy, model selection is then performed by regularizing the regression in the domain of the decorrelated data. In particular, we explore the use of both Lasso and shrinkage methods for this regularized regression step. While the technique we propose can be used with other estimates of the covariance, we have found that SMT covariance estimation results in relatively good estimates particularly when $n < p$ [5, 6]. The SMT covariance estimate achieves this improved accuracy by imposing a constraint that the eigenvector transformation should be formed by a product of sparse rotations.

Our numerical results demonstrate that the both the SMT-Lasso and SMT-Shrinkage regression methods can significantly improve the prediction performance when $n < p$, and that, for our experiments, the SMT-Lasso method yields better results than the SMT-Shrinkage method, but at the cost of greater computational cost.

2. REGRESSION MODEL

Without loss of generality, let $y \in \Re^{n\times 1}$ be a vector of n i.i.d. zero-mean Gaussian random variables which we would like to estimate. Our observations are $X \in \Re^{n\times p}$, a matrix containing n independent zero mean Gaussian random row vectors, each of dimension p. The minimum mean square error (MMSE) estimate of y given X has the form

$$\hat{y} = Xb \tag{1}$$

where b is a vector of regression coefficients given by

$$b = R_x^{-1}\rho\,, \tag{2}$$

where $R_x = \frac{1}{n}\mathrm{E}[X^tX]$ is the covariance of the observations X, and $\rho = \frac{1}{n}\mathrm{E}[X^ty]$ is the correlation between the observations X and y.

Of course, in practice R_x and ρ are often unknown, so that b must be estimated from training data (y, X). This problem has been widely studied over the years, and most recently has

become of particular interest in the challenging case when $n < p$. The traditional method for solving the regression problem is ordinary least squares (OLS). However, OLS becomes ill-posed when $n < p$, so partial least squares (PLS), ridge regression [1], and Lasso [2] have been proposed as alternatives.

3. SMT REGRESSION FOR HIGH DIMENSIONAL DATA

Our approach will be to estimate y based on the assumption that we can accurately estimate the covariance R_x. Our approach is motivated by a variety of new methods for estimating high dimensional covariance matrices through the use of sparsity constraints [5, 7]. In particular, we will use the recently introduced SMT covariance estimation method, which has been shown to produce substantially more accurate covariance estimates for certain physical data through the introduction of a covariance model [5, 6]. Importantly, the SMT covariance estimate can accurately produce all the eigenvalues of the covariance even when $n < p$, and the resulting estimate is typically full rank.

Perhaps surprisingly, we find that even when the exact value of the covariance is used in (2), the resulting regression coefficients may yield estimates that are inferior to established methods. Intuitively, this is because the correlation ρ must also be accurately determined. Therefore, having an accurate estimate of R_x does not insure success.

Our proposed regression approach is based on two steps. In the first step, we decorrelate the observations using the estimate of R_x. In the second step, we estimate the regression coefficients in this decorrelated domain. We propose three possible methods for estimating these regression coefficients. The first method, which we refer to as SMT-Lasso, applies the Lasso regression method in the decorrelated domain. The second method, which we refer to as SMT-Shrinkage, shrinks the regression coefficients toward zero; and the third method, which we refer to as SMT-subset selection, simply selects the coordinates which are most correlated with the y.

For all the three methods, we use the SMT covariance estimate to decorrelate the observation data in the first step. Let $\hat{R}_x$ be the covariance estimate, and let the eigen decomposition of the covariance estimate be given by

$$\hat{R}_x = \hat{E}\hat{\Lambda}\hat{E}^t \,, \tag{3}$$

where $\hat{E}$ is the orthonormal matrix of eigenvectors and $\hat{\Lambda}$ is a diagonal matrix of eigenvalues. Using these estimated quantities, we can approximately decorrelate and whiten the observed data using the transformation

$$\tilde{X} = X\hat{E}\hat{\Lambda}^{-\frac{1}{2}} \,. \tag{4}$$

For the SMT covariance estimate, the entries of the diagonal matrix $\hat{\Lambda}$ are generally positive, so the matrix is invertible.

Using the whitened observations $\tilde{X}$, the estimate of y can now be expressed as

$$\hat{y} = \tilde{X}\beta \,. \tag{5}$$

Since $\tilde{X}$ is a linear transformation of the observations X, it does not change the OLS estimate. However, it can change other regression results based on nonlinear estimators of the regression coefficients.

An important special case occurs if the observations are perfectly whitened. In this case, $\frac{1}{n}E\left[\tilde{X}^t\tilde{X}\right] = I$, and the regression parameters for MMSE estimation are given by

$$\beta = \frac{1}{n}\mathrm{E}[\tilde{X}^t y] \,. \tag{6}$$

The question remains of how to compute an effective estimate of β. For this purpose, we propose three methods.

3.1. SMT-Lasso

The first method we propose for estimating β, which we call SMT-Lasso, is based on the use of least squares minimization with an L_1 norm constrain on β [2]. The SMT-Lasso estimate is given by

$$\hat{\beta} = \arg\min_{\beta}\left\{\parallel y - \tilde{X}\beta \parallel^2 + \lambda \parallel \beta \parallel_1\right\} \,, \tag{7}$$

where λ is a regularization parameter that control the intensity of L_1 constrain.

Notice that the solution to (7) depends on the specific whitening transformation used in (4) because the L_1 norm is not invariant to orthonormal rotations. Therefore, SMT-Lasso will produce a different solution from conventional Lasso performed in the native coordinates of X. Since the columns of the matrix $\tilde{X}$ are not exactly orthogonal, the SMT-Lasso regression coefficients are computed as the solution to a quadratic programing problem. As with conventional Lasso, this optimization problem can be computed using a variety of efficient techniques [2, 8].

3.2. SMT-Shrinkage

The second method we propose for estimating β, which we call SMT-Shrinkage, is based on the approximation that the columns of $\tilde{X}$ are orthogonal, and have an L_2 norm of n. More specifically, we assume that

$$\frac{1}{n}\tilde{X}^t\tilde{X} \approx I \,. \tag{8}$$

In fact, the form of the SMT covariance estimate ensures that $\frac{1}{n}\mathrm{diag}\left\{\tilde{X}^t\tilde{X}\right\} = I$. So the columns of $\tilde{X}$ have a constant L_2 norm of n. In addition, the columns are approximately orthogonal because SMT covariance estimation attempts to minimize correlation between columns subject to a constraint on the required number of sparse rotations.

Using this approximation, the solution to (7) can be computed by first solving the unconstrained optimization problem to yield

$$\hat{\beta} = \frac{1}{n}\tilde{X}^t y \,, \tag{9}$$

and then applying the soft shrinkage operator to yield

$$\hat{\beta}_{\gamma i} = \text{sign}(\hat{\beta}_i)(\mid \hat{\beta}_i \mid -\gamma)^+ \,, \text{for } i = 1, 2, \ldots, p \,, \tag{10}$$

where γ is a regularization parameter and the operator $(\cdot)^+$ returns the positive portion of the argument. Notice that this soft shrinkage operator has the same form that has been proposed for shrinkage of wavelet coefficients [9].

3.3. SMT-Subset Selection

The third method we propose for estimating β, which we call SMT-Subset Selection, is similar to SMT-Shrinkage, but it selects a subset of decorrelated components for the regression, rather than shrinking the components toward zero.

For SMT-Subset selection, we first compute $\hat{\beta}$ of (9). We then apply the following operation to each component of $\hat{\beta}$

$$\hat{\beta}_{\gamma i} = \begin{cases} \hat{\beta}_i & \text{if } \mid \hat{\beta}_i \mid > \gamma \\ 0 & \text{otherwise} \end{cases} \,. \tag{11}$$

Notice that this operation selects the components that have the largest correlation with the observations to be predicted.

The values of the regularization parameters, λ in (7), and γ in (10) and (11), can be estimated using cross validation.

4. NUMERICAL EXPERIMENTS

In this section, we compare the accuracy of various regression methods using the following model for $X \in \Re^{n \times p}$

$$X = y \cdot \tau^t + W \,, \tag{12}$$

where $\tau \in \Re^{p \times 1}$ is a deterministic but unknown signal, $y \in \Re^{n \times 1}$ is a vector of n independent Gaussian random variables to be estimated, and each row of $W \in \Re^{n \times p}$ is an independent p-dimensional Gaussian random vector of correlated noise or clutter. Without loss of generality, we assume that W, y, and X are all zero mean, and that elements of y have unit variance.

For all numerical experiments, the assumed clutter covariance, $R_w = \frac{1}{n}\text{E}[W^t W]$, is computed from the real hyperspectral data shown in Fig. 1 with dimension $p = 191$ [10]. We use two covariance matrices in our experiments corresponding to "grass" and "water" classes, and for these two classes, $\frac{1}{p}\text{trace}\{R_w\}$ is equal to 9.68×10^4 and 2.83×10^4, respectively. Accordingly, we know that

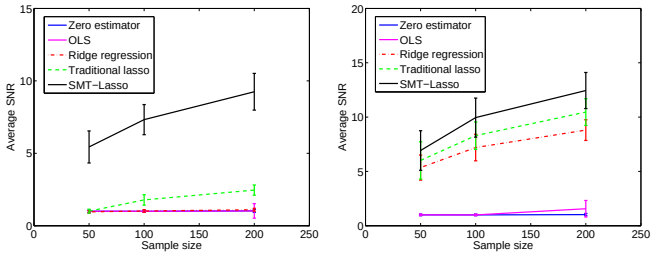
$$R_x = R_w + \tau \cdot \tau^t \,. \tag{13}$$

All comparisons are done in terms of the signal-to-noise ratio,

$$SNR = \frac{\| y \|^2}{\| y - X\hat{b} \|^2} \,, \tag{14}$$

(a) (b) (c)

Fig. 1. (a) Simulated color IR view of an airborne hyperspectral data over the Washington DC Mall [10]. (b) Ground-truth pixel spectrum of grass. (c) Ground-truth pixel spectrum of water.



(a) grass (b) water

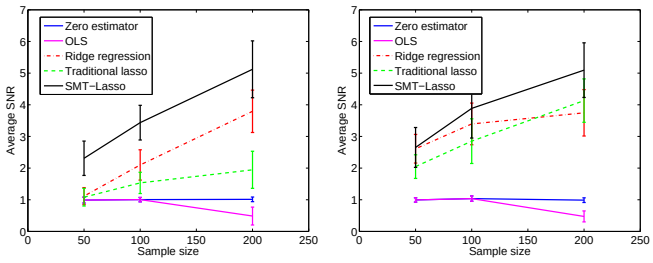
Fig. 2. Plots of average SNR when τ is the 170-th eigenvector of R_w. Notice that SMT-Lasso regression results in the highest SNR in the range of $n < p$. (a) Clutter W is generated using hyperspectral grass data. (b) Clutter W is generated using hyperspectral water data.

which is computed for a set of 300 data vectors that are generated independently of the training data. Also, in each case, the SNR is averaged over 30 simulations, each of which uses a different realization of the cluster and signal.

4.1. When τ Is a Small Eigenvector

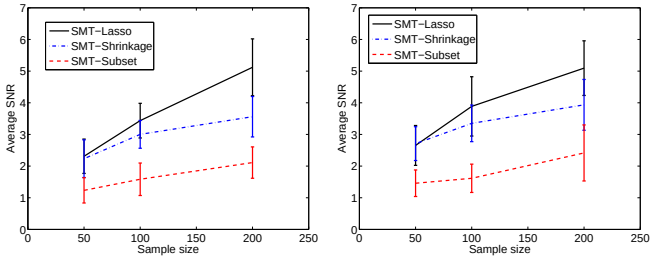
Here, we investigate the case when the signal, τ, falls in a subspace corresponding to small eigenvalues of the clutter/noise. To do this, we choose τ to be the 170-th eigenvector of R_w (with eigenvalues sorted from largest to smallest). Furthermore, the data is scaled so that $||\tau||^2 = 3^2$. The experiments are run for $n = 50, 100$, and 200, and the results are compared with the zero estimator ($\hat{b} = 0$), OLS, ridge regression, and traditional Lasso regression.

Figure 2 shows the plots of the average SNR as a function of the sample size n for both the "grass" and "water" covariance matrices. In Fig. 2(a) we can see the traditional regularized regression methods perform poorly; however, they do improve the prediction accuracy in some cases as shown in Fig. 2(b). Notice that the SMT-Lasso regression results in the highest SNRs in the whole range of $n < p$ for both cases. This is because the SMT covariance estimate is able to represent the signal that exists in small eigenvector subspaces of the clutter.



(a) grass (b) water

Fig. 3. Plots of average SNR when τ is a random Gaussian signal. Notice that SMT-Lasso regression results in consistently higher SNR in the range of $n < p$ compared to the other regression methods. (a) Clutter W is generated using hyperspectral grass data. (b) Clutter W is generated using hyperspectral water data.



(a) grass (b) water

Fig. 4. SMT-Lasso versus SMT-Shrinkage versus SMT-Subset. SMT-Lasso works best, but is more computationally expensive. (a) Clutter W is generated using hyperspectral grass data. (b) Clutter W is generated using hyperspectral water data.

4.2. When τ Is a Random Signal

Here, τ is generated as a Gaussian random vector with zero mean and covariance I. The amplitude of τ is then scaled so that $\| \tau \|^2 = 4^2$.

Figure 3 shows the plots of the average SNRs as a function of the sample size n for the "grass" and "water" classes. Notice again that the SMT-Lasso regression results in consistently higher SNR in the range of $n < p$ compared to the other methods. Figure 4 shows the SNR for each of the methods SMT-Lasso, SMT-Shrinkage, and SMT-Subset selection. SMT-Lasso performs best but is more computationally expensive than the others because of the optimization step in (7) .

5. CONCLUSIONS

In this paper, we proposed a novel regression approach for high dimensional data. In this approach, the SMT covariance estimate is computed and used to decorrelate the data, and then different model selection methods are used to obtain good estimates of regression coefficients in the decorrelated data domain. Numerical examples show that the proposed approach can significantly improve the SNR of the regression models, especially for "small n, large p" problems.

6. REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2001.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[3] X. J. Jeng, Z. J. Daye, and Y. Zhu, "Sparse covariance thresholding for high-dimensional variable selection with the lasso," Purdue University, Statistics Department, Technical Report TR 08-07, 2008.

[4] D. M. Witten and R. Tibshirani, "Covariance-regularized regression and classification for high dimensional problems," *Journal of the Royal Statistical Society: Series B*, vol. 71, no. 3, pp. 615–636, 2009.

[5] G. Cao and C. Bouman, "Covariance estimation for high dimensional data vectors using the sparse matrix transform," in *Advances in Neural Information Processing Systems*. MIT Press, 2008, pp. 225–232.

[6] G. Cao, C. A. Bouman, and J. Theiler, "Weak signal detection in hyperspectral imagery using sparse matrix transformation (SMT) covariance estimation." First Workshop on Hyperspectral Image and Signal Processing, 2009.

[7] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Annals of Statistics*, vol. 36, no. 1, pp. 199–227, 2008.

[8] S. Alliney and S. Ruzinsky, "An algorithm for the minimization of mixed l(1) and l(2) norms with application to bayesian estimation," *IEEE Transactions on Signal Processing*, vol. 42, no. 3, pp. 618–627, 1994.

[9] D. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, pp. 1200–1224, 1995.

[10] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. New York: Wiley-Interscience, 2005.