Purdue University
Purdue e-Pubs

ECE Technical Reports

Electrical and Computer Engineering

4-28-2008

Covariance Estimation for High Dimensional Data Vectors Using the Sparse Matrix Transform

Guangzhi Cao
Purdue University, guangzhi.cao@ge.com

Charles A. Bouman
Purdue University, bouman@purdue.edu

Cao, Guangzhi and Bouman, Charles A., "Covariance Estimation for High Dimensional Data Vectors Using the Sparse Matrix Transform" (2008). *ECE Technical Reports.* Paper 373.
http://docs.lib.purdue.edu/ecetr/373

Covariance Estimation for High Dimensional Data Vectors Using the Sparse Matrix Transform

Guangzhi Cao and Charles A. Bouman
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907

April 29, 2008

1 Introduction

Many problems in statistical pattern recognition and analysis require the classification and analysis of high dimensional data vectors. However, covariance estimation for high dimensional vectors is a classically difficult problem because the number of coefficients in the covariance grows as the dimension squared [1, 2, 3]. This problem, sometimes referred to as the curse of dimensionality [4], presents a classic dilemma in statistical pattern analysis and machine learning.

In a typical application, one measures M versions of an N dimensional vector. If $M < N$, then the sample covariance matrix will be singular with $N - M$ eigenvalues equal to zero. Over the years, a variety of techniques have been proposed for computing a nonsingular estimate of the covariance. For example, regularized and shrinkage covariance estimators [5, 6, 7, 8, 9, 10] are examples of such techniques.

In this paper, we propose a new approach to covariance estimation, which is based on constrained maximum likelihood (ML) estimation of the covariance. In particular, the covariance is constrained to have an eigen decomposition which can be represented as a sparse matrix transform (SMT) [11]. The SMT is formed by a product of pairwise coordinate rotations known as Givens rotations [12]. Using this framework, the covariance can be efficiently estimated using greedy minimization of the log likelihood function, and the number of Givens rotations can be efficiently computed using a cross-validation procedure. The estimator obtained using this method is always positive definite and well-conditioned even with limited sample size.

In order to validate our model, we perform experiments using a standard set of hyperspectral data [13]. Our experiments show that SMT covariance estimation results in consistently better estimates of the covariance for a variety of different classes and sample sizes. Also, we show that the SMT method has a particular advantage over traditional methods when estimating small eigenvalues and their associated eigenvectors.

2 Covariance Estimation for High Dimensional Vectors

In the general case, we observe a set of M vectors, $y_1, y_2, \cdots, y_M$, where each vector, y_m, is N dimensional. Without loss of generality, we assume y_m has zero mean. We can represent this data as the following $N \times M$ matrix

$$Y = [y_1, y_2, \cdots, y_M] \ .$$

If the vectors y_m are assumed to be identically distributed, then the sample covariance is computed by

$$S = \frac{1}{M} Y Y^t \ ,$$

and S is an unbiased estimate of the true covariance matrix[1]

$$R = \mathrm{E}\left[y_m y_m^t \right] = \mathrm{E}[S] \ .$$

While S is an unbiased estimate of R it is also singular when $M < N$. This is a serious deficiency since as the dimension N grows, the number of vectors needed to estimate R also grows. In practical applications, M may be much smaller than N which means that most of the eigenvalues of R are erroneously estimated as zero.

A variety of methods have been proposed to regularize the estimate of R so that it is not singular. Shrinkage estimators are a widely used class of estimators which regularize the covariance matrix by shrinking it toward some target structures [7, 8, 9, 10]. Shrinkage estimators generally have the form

$$\hat{R} = \alpha D + (1 - \alpha) S \ ,$$

where D is some positive definite matrix. One popular choice for D is the identity matrix or its scalar multiple. Another popular choice for D is diag(S), the matrix formed by the diagonal entries of S. In either case, the parameter α can be estimated using cross-validation or boot-strap methods. However, neither of these regularization procedures full account for the correlation between elements of the vector.

Our approach will be to compute a constrained maximum likelihood (ML) estimate of the covariance R, under the assumption that eigenvectors of R may be represented as a sparse matrix transform (SMT). So we first decompose R as

$$R = E \Lambda E^t \ ,$$

where E is the orthonormal matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues. Then we will estimate the covariance by maximizing the likelihood of the data Y subject to the constraint that E is an SMT. By varying the order, K, of the SMT, we may then reduce or increase the regularizing constraint on the covariance.

[1] If the sample mean is used as an estimate of the mean, then $\frac{M}{M-1} S$ is the unbiased estimate of the covariance.

2.1 Maximum Likelihood Covariance Estimation

If we assume that the columns of Y are independent and identically distributed Gaussian random vectors with mean zero and positive-definite covariance R, then the likelihood of Y given R is given by

$$p_R(Y) = \frac{1}{(2\pi)^{\frac{NM}{2}}} |R|^{-\frac{M}{2}} \exp\left\{ -\frac{1}{2}\mathrm{tr}\{Y^t R^{-1} Y\} \right\} \ .$$

From Appendix A, we see that the log-likelihood of Y is given by

$$\log p_{(E,\Lambda)}(Y) = -\frac{M}{2}\mathrm{tr}\{\mathrm{diag}(E^t S E)\Lambda^{-1}\} - \frac{M}{2}\log|\Lambda| - \frac{NM}{2}\log(2\pi) \ .$$

where $R = E\Lambda E^t$ is specified by the orthonormal eigenvalue matrix E and diagonal eigenvalue matrix Λ. Jointly maximizing the likelihood with respect to E and Λ then results in the ML estimates of E and Λ given by (see Appendix A)

$$\begin{aligned} \hat{E} &= \arg\min_{E\in\Omega} \left\{ \left| \mathrm{diag}(E^t S E) \right| \right\} & (1) \\ \hat{\Lambda} &= \mathrm{diag}(\hat{E}^t S \hat{E}) \ , & (2) \end{aligned}$$

where Ω is the set of allowed orthonormal transforms. So we may compute the ML estimate by first solving the constrained optimization of (1), and then computing the eigenvalues from (2).

An interesting special case occurs when S has full rank and Ω is the set of all orthonormal transforms. In this case, equations (1) and (2) are solved by selecting E and Λ as the eigenvalues and eigenvectors of S (See Appendix B). So this leads to the well known result that when S is non-singular, then the ML estimate of the covariance is given by the sample covariance, $\hat{R} = S$. However, when S is singular and Ω is the set of all orthonormal transforms, then the likelihood is unbounded, with a subset of the estimated eigenvalues tending toward zero.

2.2 ML Estimation of Eigenvectors Using SMT Model

The ML estimate of E can be improved by constraining the feasible set of eigenvectors, Ω, to a smaller set. We will see that by properly constraining the set Ω, we can compute the ML estimate of the covariance of data even when the sample covariance, S, is singular. By constraining Ω, we are effectively regularizing the ML estimate by imposing a model constraint. As with any model-based approach, the key is to select a feasible set, Ω, which is as small as possible while still accurately modeling the behavior of real data.

Our approach is to select Ω to be the set of all orthonormal transforms that can be represented as an SMT of order K [11]. More specifically, a matrix E is an SMT of order K if it can be written as a product of K sparse orthornormal matrices, so that

$$E = \prod_{k=K-1}^{0} E_k = E_{K-1}E_{K-2}\cdots E_0 \ , \qquad (3)$$

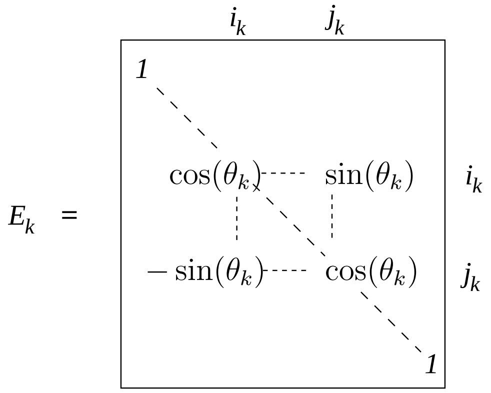


Figure 1: The structure of a pair-wise sparse transform E_k. Here, all the unlabeled diagonal elements are 1's, and all the unlabeled off-diagonal elements are 0's. Each transform E_k is a Givens transform with rotation angle θ_k.

where every sparse matrix, E_k, is a Givens rotation operating on a pair of coordinate indices (i_k, j_k) [12]. Figure 1 illustrates the structure of the Givens rotation. Notice that every Givens rotation E_k is an orthonormal rotation in the plane of the two coordinates, i_k and j_k, which has the form

$$E_k = I + \Theta(i_k, j_k, \theta_k) \ , \tag{4}$$

where $\Theta(m, n, \theta)$ is defined as

$$[\Theta]_{ij} = \begin{cases} \cos(\theta) - 1 & \text{if } i = j = m \text{ or } i = j = n \\ \sin(\theta) & \text{if } i = m \text{ and } j = n \\ -\sin(\theta) & \text{if } i = n \text{ and } j = m \\ 0 & \text{otherwise} \end{cases} \ . \tag{5}$$

Figure 2 shows the flow diagram for the application of an SMT to a data vector y. Notice that each 2D rotation, E_k, plays a role analogous to a "butterfly" used in an traditional Fast Fourier Transform (FFT) [14]. However, unlike an FFT, the organization of the butterflies in an SMT is unstructured, and each butterfly can have an arbitrary rotation angle θ_k. This more general structure allows an SMT to implement a larger set of orthonormal transformations. In fact when $K = \binom{N}{2}$, the SMT can be used to exactly represent any $N \times N$ orthonormal transformation (see Appendix C). Even when $K << \binom{N}{2}$, the SMT can represent a wide range of operations since it captures the dependencies between pairs of coordinates.

Using the SMT model constraint, the ML estimate of E is given by

$$\hat{E} = \arg \min_{E = \prod_{k=K-1}^{0} E_k} \left| \text{diag}(E^t S E) \right| \ . \tag{6}$$

Unfortunately, evaluating the constrained ML estimate of (6) requires the solution of a optimization problem with a nonconvex constraint. So evaluation of the globally optimal solutions is very difficult. Therefore, our approach will be to use greedy minimization to compute an approximate solution to (6). The greedy minimization approach works by selecting each new butterfly E_k to minimize the cost, while fixing the previous butterflies, E_l

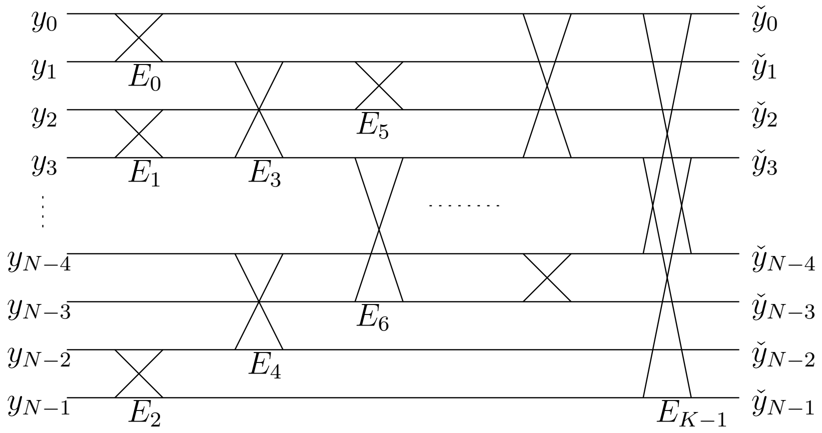


Figure 2: The structure of an SMT implementation of $\check{y} = Ey$. Every E_k plays a role analogous to a "butterfly" used in an traditional FFT. However, the irregular structure of the SMT makes it a generalization of the FFT and allows it to be used to accurately approximate a general orthonormal transform. In fact, an SMT with $N(N-1)/2$ butterflies can be used to exactly compute any $N \times N$ orthonormal transformation.

for $l < k$. The solution to this local optimization can be computed quite easily by determining the two coordinates, i_k and j_k, that are most correlated, and then applying a rotation with angle θ that decorrelates the measurements (See Appendix D).

This greedy optimization algorithm can be implemented with the following simple recursive procedure. We start by setting $S_0 = S$ to be the sample covariance, and initialize $k = 0$. Then we apply the following two steps for $k = 0$ to $K - 1$.

$$E_k^* = \arg\min_{E_k} \left|\text{diag}\left(E_k^t S_k E_k\right)\right| \tag{7}$$

$$S_{k+1} = E_k^{*t} S_k E_k^* \ . \tag{8}$$

The resulting values of E_k^* are the butterflies of the SMT.

The problem remains of how to compute the solution to (7). For a specified coordinate pair (i, j), the cost function in (7) is minimized when the off-diagonal entries S_{ij} and S_{ji} become zero. In this case, the ratio of the minimized cost function to its original value is given by

$$\left(1 - \frac{S_{ij}^2}{S_{ii}S_{jj}}\right) \ , \tag{9}$$

where i and j are the indices corresponding to the pair-wise transform of E_k (See Appendix D). Therefore, with each iteration of the greedy algorithm, we select the coordinate pair (i_k, j_k) that reduces the cost in (7) most among all possible pairs, i.e.,

$$(i_k, j_k) \leftarrow \arg\min_{(i,j)} \left(1 - \frac{S_{ij}^2}{S_{ii}S_{jj}}\right) \ . \tag{10}$$

Once i_k and j_k are determined, the Givens rotation E_k^* is given by

$$E_k^* = I + \Theta(i_k, j_k, \theta_k) \ , \tag{11}$$

where

$$\theta_k = \frac{1}{2}\text{atan}(-2S_{i_k j_k}, S_{i_k i_k} - S_{j_k j_k}) \ . \tag{12}$$

By iterating the (7) and (8) K times, we obtain the constrained ML estimation of E given by

$$\hat{E}^* = \prod_{k=K-1}^{0} E_k^* \ .$$

2.3 Model Order

A simple cross-validation method can be used to determine the model order K. First, we split the sample data Y into three subsets Y_1, Y_2, and Y_3. Then we then compute an order k SMT covariance estimate using the combined data from subsets Y_2 and Y_3. We denote this estimates by $(\hat{E}_{1,k}, \hat{\Lambda}_{1,k})$. Next, we compute the log likelihood of Y_1 given $(\hat{E}_{1,k}, \hat{\Lambda}_{1,k})$, which is given by

$$\log p_{(E_{1,k},\Lambda_{1,k})}(Y_1) = -\frac{M}{2}\text{tr}\{\text{diag}(\hat{E}^t_{1,k} S_1 \hat{E}_{1,k})\hat{\Lambda}^{-1}_{1,k}\} - \frac{M}{2}\log\left|\hat{\Lambda}_{1,k}\right| - \frac{NM}{2}\log(2\pi) \ .$$

In fact, this log likelihood expression can be evaluated on-the-fly as $\hat{E}_{1,k}$ is estimated starting from $k = 0$. This on-the-fly evaluation dramatically reduces the computation.

The cross-validation log likelihood is then evaluated in a similar manner for the subsets Y_2 and Y_3 to yield $(\hat{E}_{2,k}, \hat{\Lambda}_{2,k})$ and $(\hat{E}_{3,k}, \hat{\Lambda}_{3,k})$. The order is then estimated by maximizing the sum of the cross-validation likelihood terms over the three possible partitions.

$$K^* = \arg\max_{k\in\{0,1,2,\dots\}} \sum_{i=1}^{3} \log p_{(E_{i,k},\Lambda_{i,k})}(Y_i) \ .$$

Once K^* is determined, the proposed covariance estimator is re-computed using using all the data and the estimated model order.

3 Experimental Results

The effectiveness of the SMT covariance estimation procedure depends on how effectively the SMT model can capture the behavior of real data vectors. Therefore in this section, we compare the effectiveness of the SMT covariance estimator to commonly used shrinkage estimators describe in Appendix E. We do this comparison using hyperspectral remotely sensed data as our high dimensional data vectors.

The hyperspectral data we use is available with the recently published book by Landgrebe [13]. Figure 3 shows five simulated color IR views of an airborne hyperspectral data flightline over the Washington DC Mall. The sensor system used here measured pixel response in 210 bands in the 0.4 to 2.4 μm region of the visible and infrared spectrum. Bands in the 0.9 and 1.4 μm region where the atmosphere is opaque have been omitted from the data set, leaving 191 bands. The data set contains 1208 scan lines with 307 pixels in each scan line. Each of the images were made using bands 60, 27 and 17 for the red, green and blue colors, respectively. The data set also provides ground truth pixels for each of five classes designated as grass, water, roof, street, and tree. Figures 3 (a) through (e) show images for each of the five ground-truth classes. For each image, the pixels of the designated class are outlined with a white rectangle.

For each class, we computed the "true" covariance by using all the ground truth pixels to calculate the sample covariance. The covariance is computed by first subtracting the sample mean vector for each class, and then computing the sample covariance for the zero mean vectors. The number of pixels for the ground-truth classes of water, grass, roof, street, and tree are 1224, 1928, 3579, 416, and 388, respectively. In each case, the number of ground truth pixels was much larger than 191, so the true covariance matrices are nonsingular, and accurately represent the covariance of the hyperspectral data for that class.[2]

3.1 Gaussian Case

First, we compare how different estimators perform when the data vectors are samples from an ideal multivariate Gaussian distribution. To do this, we first generated zero mean multivariate vectors with the true covariance for each of the five classes. Next we estimated the covariance using each of four methods, SMT covariance estimation, shrinkage method I, shrinkage method II, and shrinkage method III. (See Appendix E for a description of the shrinkage estimators.) In order to determine the effect of sample size, we also performed each experiment for a sample sizes of $M = 80$, 40, and 20.

In order to get an aggregate accessment of the effectiveness of SMT covariance estimation, we compared the estimated covariance for each method to the true covariance using the Kullback-Leibler (KL) distance as derived in Appendix F. The KL distance is a measure of the error between the estimated and true distribution. So a value of 0 indicates a perfect estimate, and larger values indicate greater error. Figures 4-8 show plots of the KL distance as a function of sample size for three of the four estimators.[3] Notice that the values of the Kullback-Leibler distance are consistently and substantially smaller for the SMT covariance estimation method. The error bars indicate the standard deviation of the KL distance due to random variation in the sample statistics.

Figure 9 shows the spectrum of the ground-truth pixels for the water and grass classes. Then Figures 10-12 and 16-18 show more details of eigenvalues estimated for water and grass, with part a) showing the samples used to compute the covariance, and part b) showing plots of the estimated eigenvalues.

For example, consider the case of water with $M = 80$ as shown in Fig. 10(b). Notice that the eigenvalues of the SMT estimator are much closer to the true values than the eigenvalues

[2] We did not estimate the covariance for the classes designated as "path" and "shadow" because the number of ground truth pixels for these two classes was too small to get an accurate estimate of the covariance.

[3] The KL distance for shrinkage estimator I is not shown because it is much larger than the remaining three estimators.

produced by the shrinkage estimators. From the plots, we can see the eigenvectors estimated by the shrinkage estimators I and II tend to degrade beyond the $M = 80^{th}$ eigenvalue. This is reasonable since eigen decomposition of the sample covariance can only estimate the 80 largest eigenvalues. Alternatively, the SMT estimator generates good estimates for the small eigenvalues. This tendency of the SMT estimator to generate better estimates of smaller eigenvalues is consistent across different classes and different sample sizes.

3.2 Non-Gaussian Case

In practice, the sample vectors may not be from an ideal multivariate Gaussian distribution. In order to see the effect of the non-Gaussian statistics on the accuracy of the covariance estimate, we performed a set of experiments in which we randomly sampled vectors from the ground truth pixels. Since these samples are from the actual measured data, their distribution is not precisely Gaussian. Using these samples, we performed the covariance estimation using the four different methods with sample sizes of $M = 80$, 40, and 20.

Plots of the KL distance for this non-Gaussian case are shown in in Figures 4-8; and Figures 13-15 and 19-21 show more details of the estimated eigenvalues and the sampled pixels spectra. We note that the results are similar to those found for the ideal Guassian case. However, in a few cases there is some reduction in the accuracy of the SMT estimate. The worst case seems to appear in Fig. 4 (b) in which the variance of the KL distance for $M = 20$ becomes large. It appears that the model-based SMT method is more sensitive to the non-Guassian statistics, however, it is likely that the variance could be reduced by the use of more than 3 sets for the cross-validation method (the shrinkage methods use true "leave-one-out" cross-validation, which is computationally expensive).

4 Conclusion

We have proposed a novel method for covariance estimation of high dimensional data. The new method is based on constrained maximum likelihood (ML) estimation in which the eigenvector transformation is constrained to be the composition of K Givens rotations. This model seems to capture the essential behavior of the data with a relatively small number of parameters. The constraint set is a K dimensional manifold in the space of orthnormal transforms, but since it is not a linear space, the resulting ML estimation optimization problem does not yield a closed form global optimum. However, we show that a recursive local optimization procedure is simple, intuitive, and yields good results.

We also demonstrate that the proposed SMT covariance estimation method substantially reduces the error in the covariance estimate as compared to current state-of-the-art shrinkage estimates for a standard hyperspectral data set.

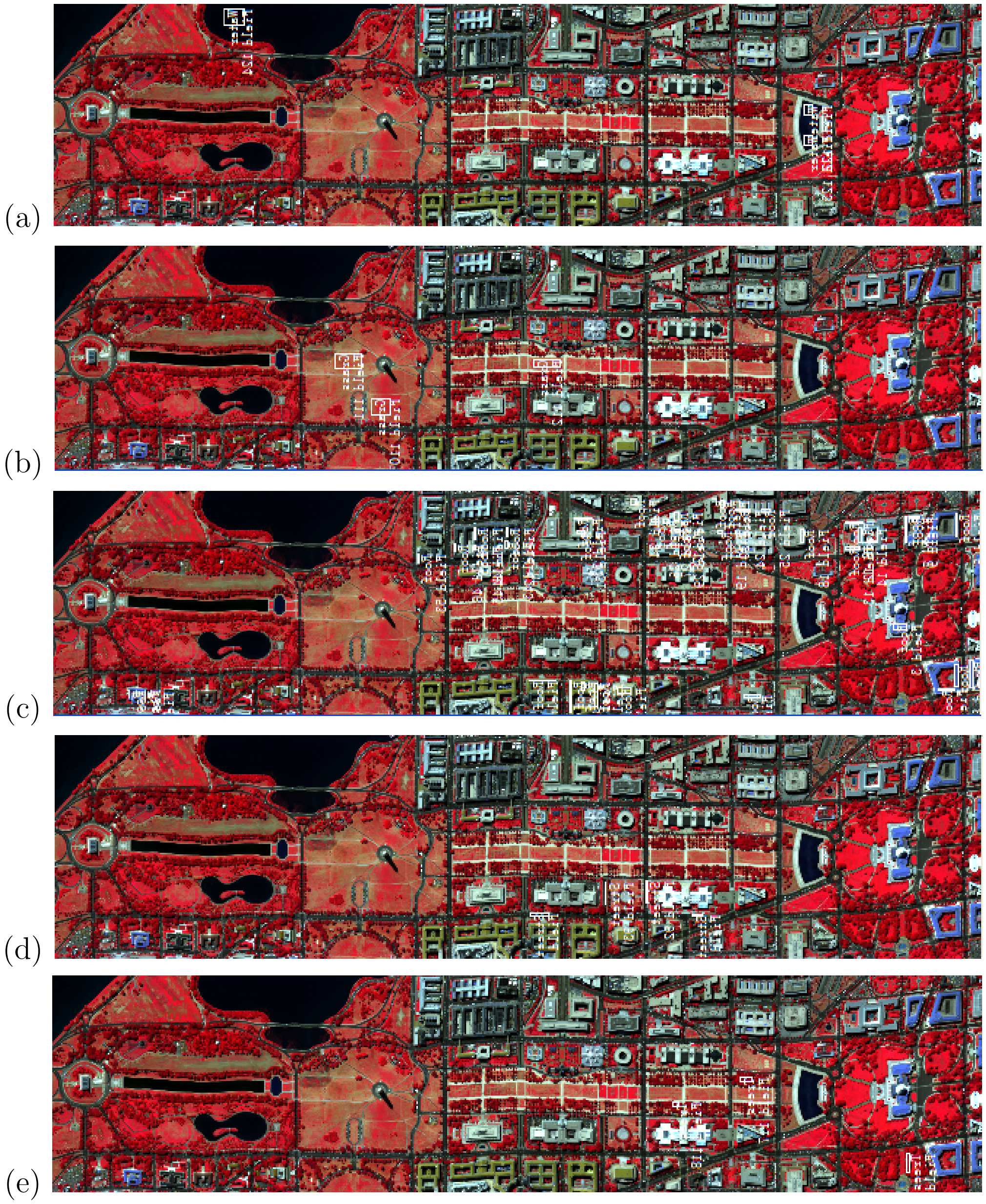
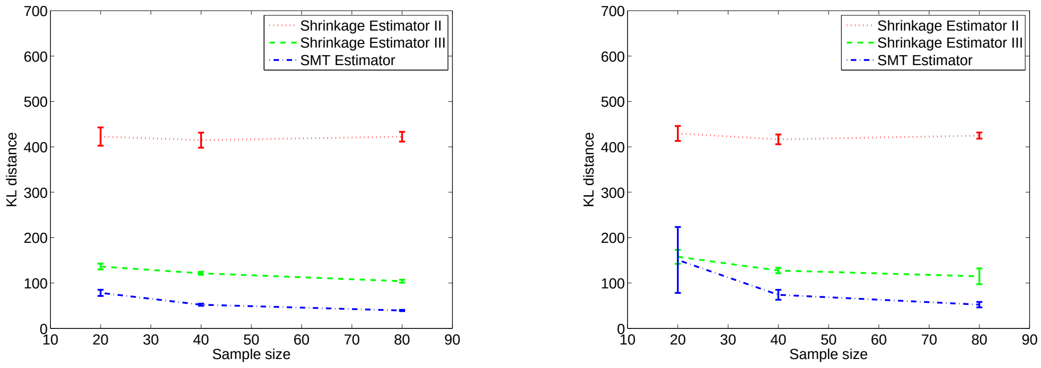
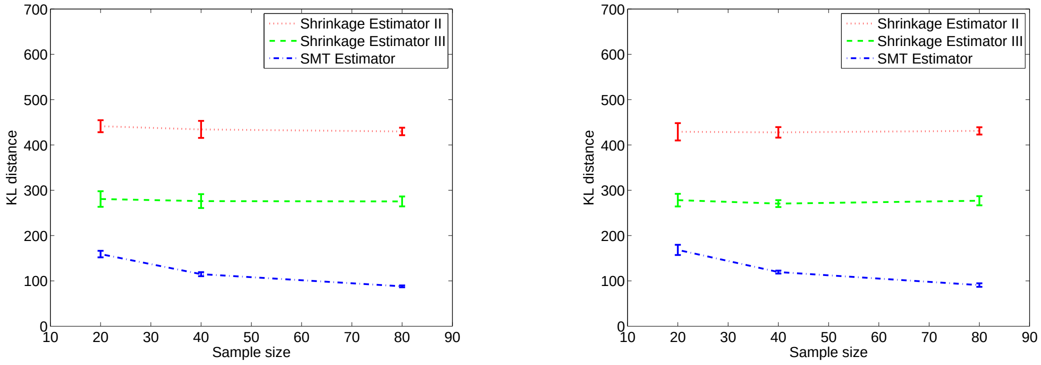


Figure 3: Images showing the location of the pixels for each of the five ground-truth classes denoted by (a) “water”, (b) “grass”, (c) “roof”, (d) “street”, and (e) “tree”. For each image, the locations of the ground-truth pixels are outlined with white rectangles.



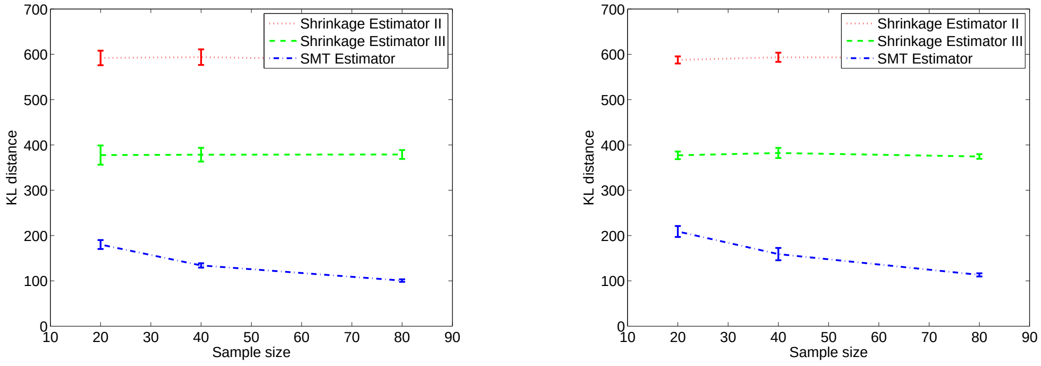
(a) Gaussian Case (b) Non-Gaussian Case

Figure 4: Kullback-Leibler distance from true distribution versus sample size for the water class.



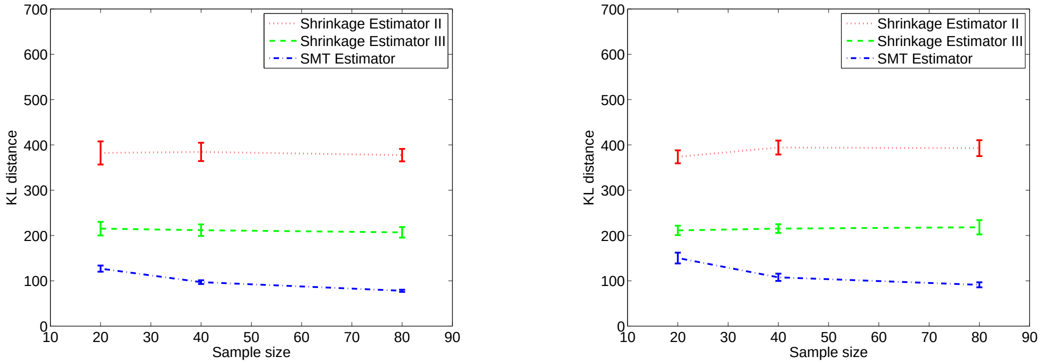
(a) Gaussian Case (b) Non-Gaussian Case

Figure 5: Kullback-Leibler distance from true distribution versus sample size for the grass class.



(a) Gaussian Case (b) Non-Gaussian Case

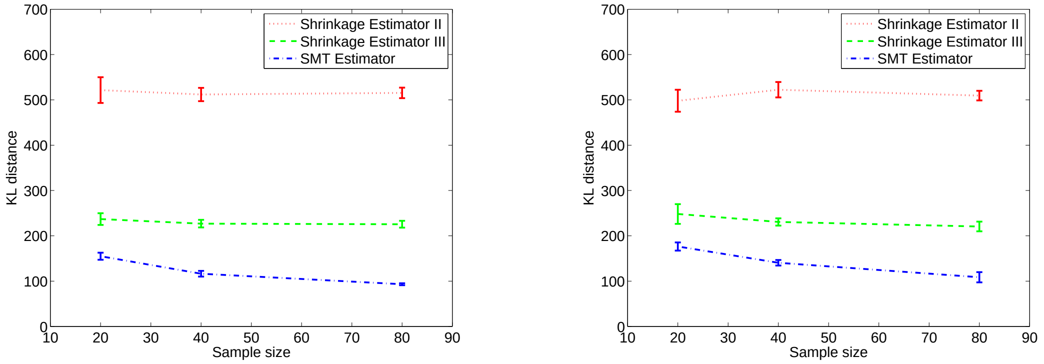
Figure 6: Kullback-Leibler distance from true distribution versus sample size for the roof class.



(a) Gaussian Case

(b) Non-Gaussian Case

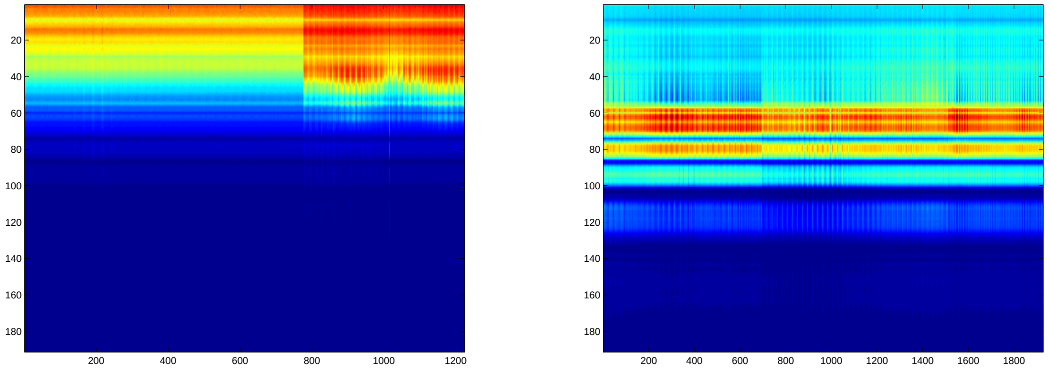
Figure 7: Kullback-Leibler distance from true distribution versus sample size for the street class.



(a) Gaussian Case

(b) Non-Gaussian Case

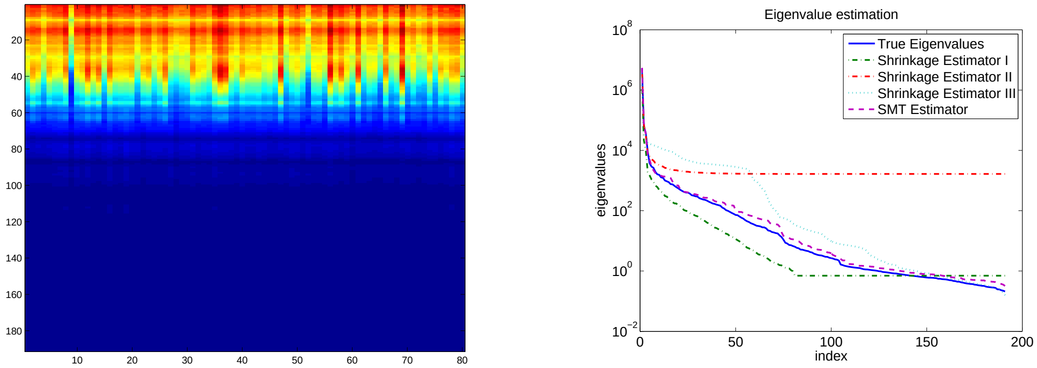
Figure 8: Kullback-Leibler distance from true distribution versus sample size for the tree class.



(a) Water Spectum

(b) Grass Spectum

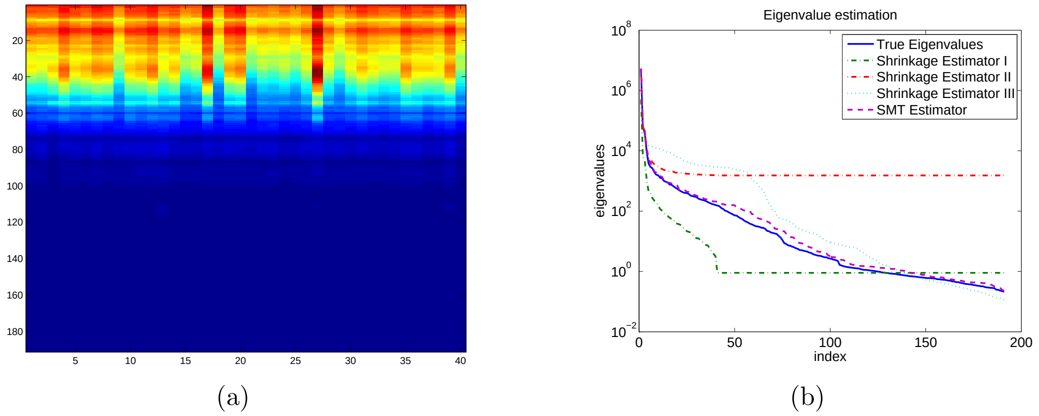
Figure 9: The spectrum of the ground-truth pixels for the (a) water and (b) grass classes. The vertical axis shows the spectral band, and the horizontal axis is the pixel number.



(a)

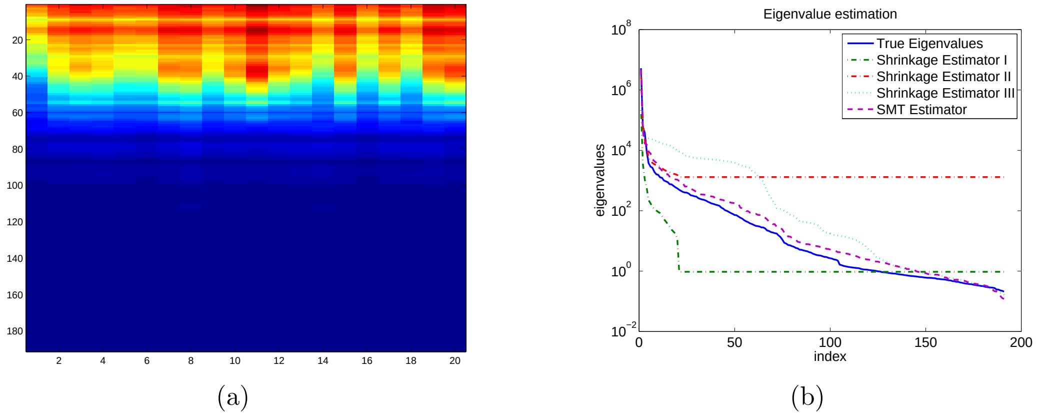
(b)

Figure 10: Gaussian case of water for $M = 80$. (a) Synthesized data spectrum. (b) The distribution of estimated eigenvalues.



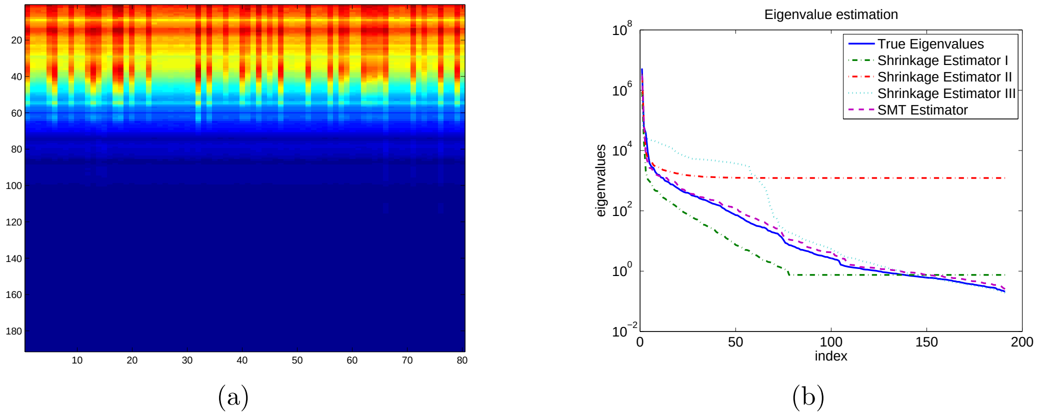
(a) (b)

Figure 11: Gaussian case of water for $M = 40$. (a) Synthesized data spectrum. (b) The distribution of estimated eigenvalues.



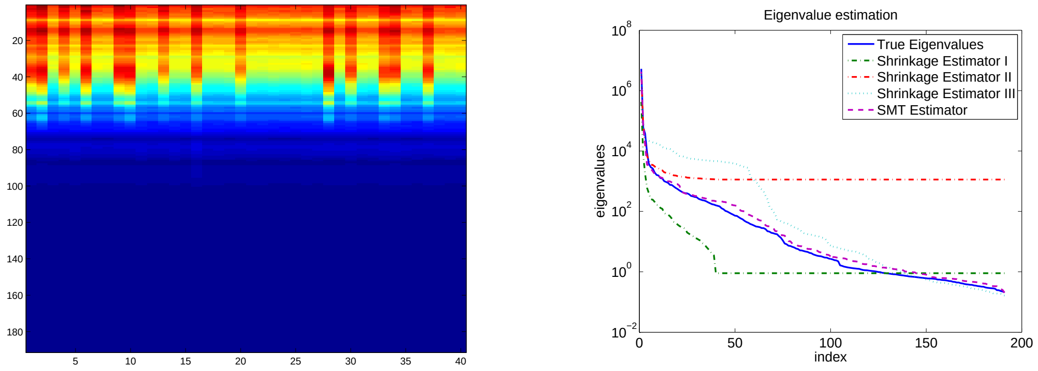
(a) (b)

Figure 12: Gaussian case of water for $M = 20$. (a) Synthesized data spectrum. (b) The distribution of estimated eigenvalues.



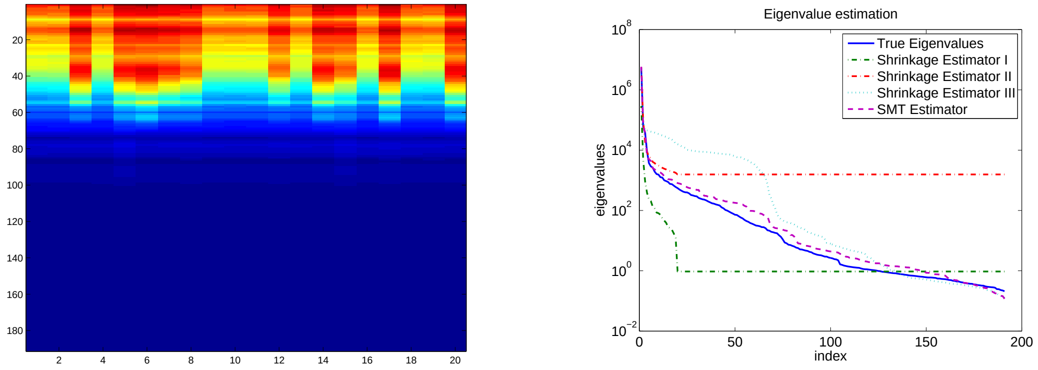
(a) (b)

Figure 13: Non-Gaussian case of water for $M = 80$. (a) Sampled data spectrum. (b) The distribution of estimated eigenvalues.



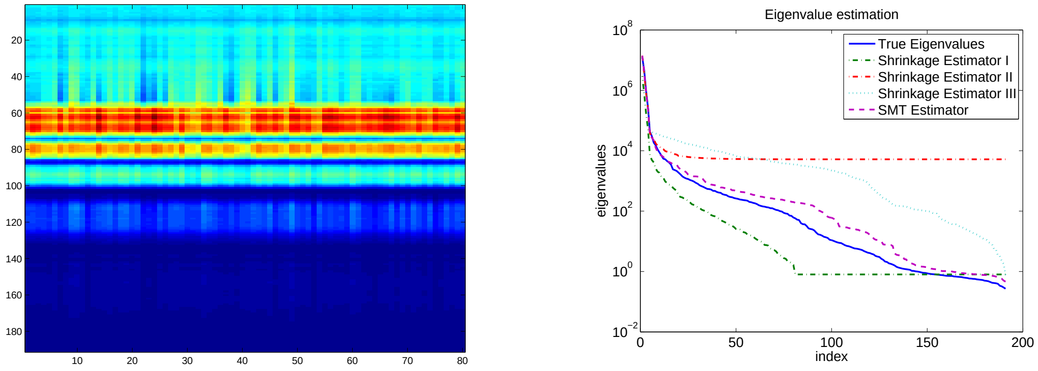
(a) (b)

Figure 14: Non-Gaussian case of water for $M = 40$. (a) Sampled data spectrum. (b) The distribution of estimated eigenvalues.



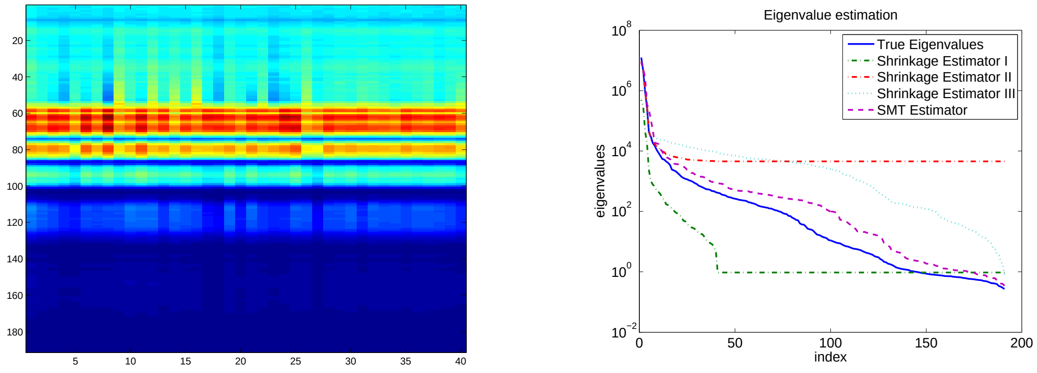
(a) (b)

Figure 15: Non-Gaussian case of water for $M = 20$. (a) Sampled data spectrum. (b) The distribution of estimated eigenvalues.



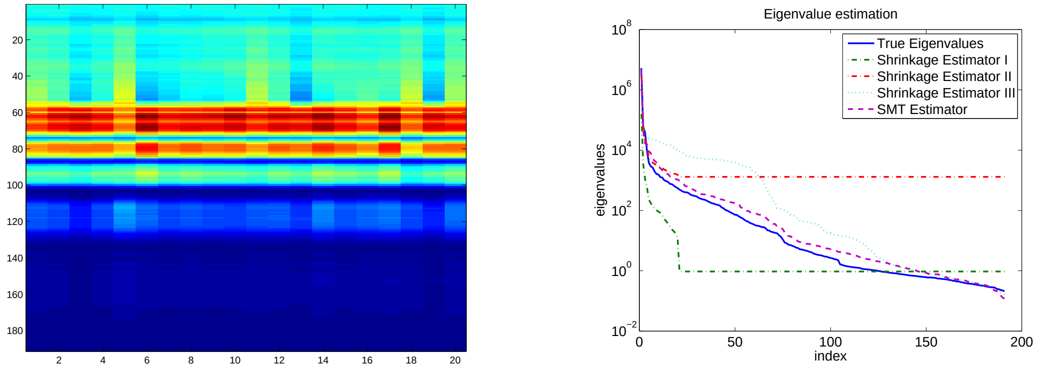
(a) (b)

Figure 16: Gaussian case of grass for $M = 80$. (a) Synthesized data spectrum. (b) The distribution of estimated eigenvalues.



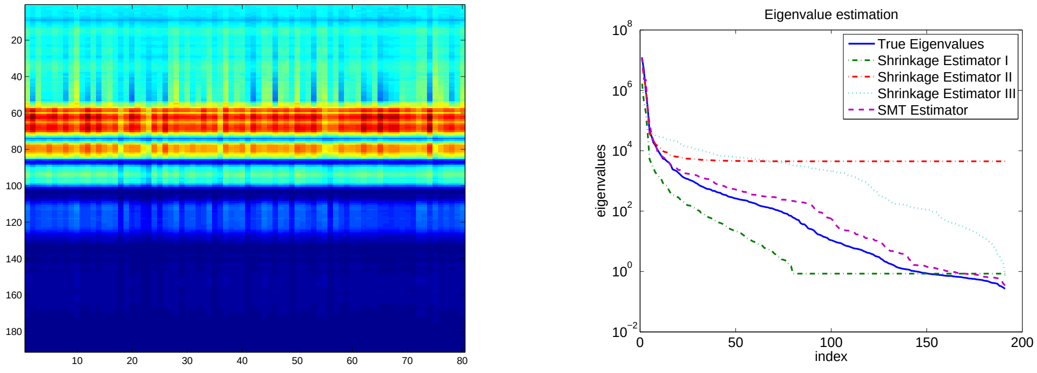
(a) (b)

Figure 17: Gaussian case of grass for $M = 40$. (a) Synthesized data spectrum. (b) The distribution of estimated eigenvalues.



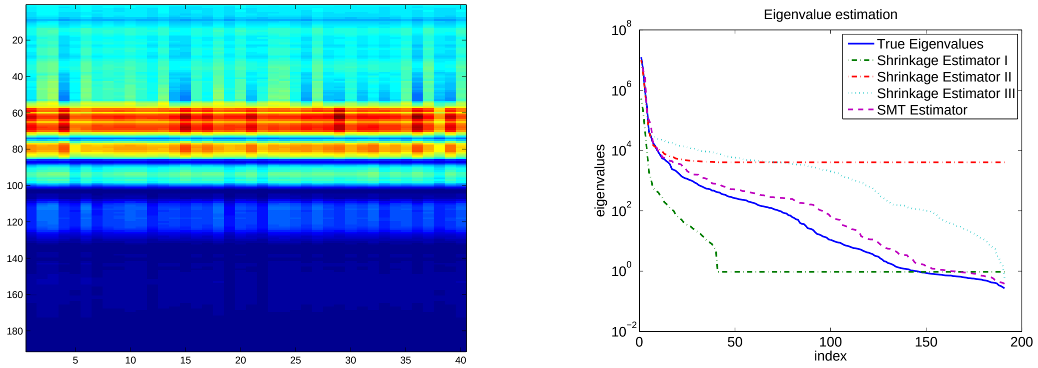
(a) (b)

Figure 18: Gaussian case of grass for $M = 20$. (a) Synthesized data spectrum. (b) The distribution of estimated eigenvalues.



(a) (b)

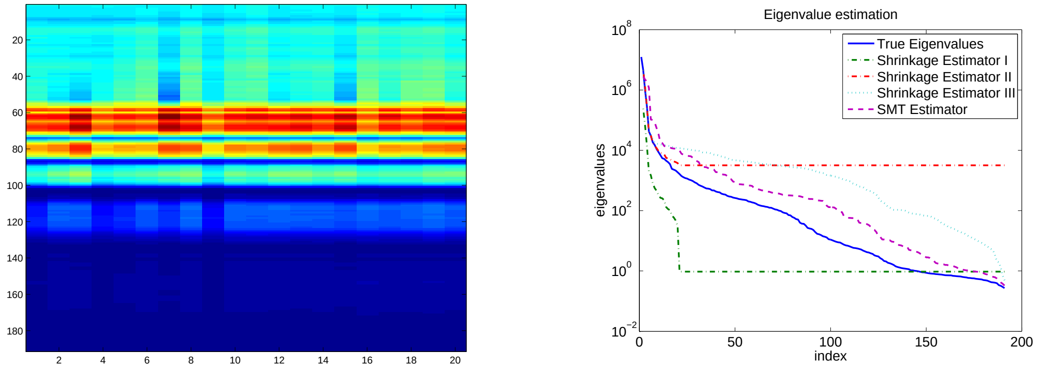
Figure 19: Non-Gaussian case of grass for $M = 80$. (a) Sampled data spectrum. (b) The distribution of estimated eigenvalues.



(a)

(b)

Figure 20: Non-Gaussian case of grass for $M = 40$. (a) Sampled data spectrum. (b) The distribution of estimated eigenvalues.



(a)

(b)

Figure 21: Non-Gaussian case of grass for $M = 20$. (a) Sampled data spectrum. (b) The distribution of estimated eigenvalues.

A Derivation of Log Likelihood

If the columns of Y are independent and identically distributed Gaussian random vectors with mean zero and positive-definite covariance R, then the likelihood of Y given R is given by

$$\begin{aligned} p_{(E,\Lambda)}(Y) &= \frac{1}{(2\pi)^{\frac{NM}{2}}} |R|^{-\frac{M}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\{Y^t R^{-1} Y\}\right\} && (13) \\ &= \frac{1}{(2\pi)^{\frac{NM}{2}}} |\Lambda|^{-\frac{M}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\{Y^t E \Lambda^{-1} E^t Y\}\right\} && (14) \\ &= \frac{1}{(2\pi)^{\frac{NM}{2}}} |\Lambda|^{-\frac{M}{2}} \exp\left\{-\frac{M}{2}\mathrm{tr}\{E^t S E \Lambda^{-1}\}\right\} && (15) \\ &= \frac{1}{(2\pi)^{\frac{NM}{2}}} |\Lambda|^{-\frac{M}{2}} \exp\left\{-\frac{M}{2}\mathrm{tr}\{\mathrm{diag}(E^t S E)\Lambda^{-1}\}\right\} . && (16) \end{aligned}$$

Taking the logarithm yields

$$\log p_{(E,\Lambda)}(Y) = -\frac{M}{2}\mathrm{tr}\{\mathrm{diag}(E^t S E)\Lambda^{-1}\} - \frac{M}{2}\log|\Lambda| - \frac{NM}{2}\log(2\pi) . \qquad (17)$$

Therefore, the maximum likelihood (ML) estimator of (E, Λ) is given by

$$\begin{aligned} (\hat{E}, \hat{\Lambda}) &= \arg\max_{(E,\Lambda)} \log p_{(E,\Lambda)}(Y) && (18) \\ &= \arg\max_{E} \max_{\Lambda} \log p_{(E,\Lambda)}(Y) . && (19) \end{aligned}$$

We first maximize the log-likelihood with respect to Λ. Setting the derivatives of $\log p_{(E,\Lambda)}(Y)$ with respect to all the diagonal entries of Λ to zero, we obtain

$$\hat{\Lambda} = \mathrm{diag}(E^t S E) .$$

Therefore, the ML estimation of E is given by

$$\begin{aligned} \hat{E} &= \arg\max_{E\in\Omega} \log p_{(E,\hat{\Lambda}(E))}(Y) && (20) \\ &= \arg\max_{E\in\Omega} \left\{-\frac{NM}{2}\log(2\pi) - \frac{M}{2}\log\left|\mathrm{diag}(E^t S E)\right| - \frac{MN}{2}\right\} && (21) \\ &= \arg\min_{E\in\Omega} \left\{\left|\mathrm{diag}(E^t S E)\right|\right\} , && (22) \end{aligned}$$

where Ω is the set of allowed orthonormal transforms. So minimization of $|\mathrm{diag}(E^t S E)|$ leads to the ML estimate of E, and hence the ML estimate of Λ whicheq:MLE is given by

$$\hat{\Lambda} = \mathrm{diag}(\hat{E}^t S \hat{E}) . \qquad (23)$$

B Unconstrained ML Estimate

Proposition: Let S be an $N \times N$ positive definite symmetric matrix with eigenvalue decomposition given by $S = E^* \Lambda_S E^{*t}$, and let Ω be the set of all $N \times N$ orthonormal transforms. Then E^* achieves the global minimization of (1), so that

$$\left|\text{diag}(E^{*t} S E^*)\right| = \min_{E \in \Omega} \left\{ \left|\text{diag}(E^t S E)\right| \right\} \ . \tag{24}$$

Proof: first, we show for any symmetric, positive definite matrix S, we have

$$|\text{diag}(S)| \geq |S| \ . \tag{25}$$

We know there exists a unique low triangular $N \times N$ matrix G, such that

$$S = GG^t \ , \tag{26}$$

which is called the Cholesky factorization [15]. Therefore, $|S| = |G|^2 = \prod_{i=1}^n G_{ii}^2$. Clearly, we have $S_{ii} = \sum_{j=1}^n G_{ij}^2 \geq G_{ii}^2$ for $i = 1, 2, \ldots, n$. This gives

$$|\text{diag}(S)| \geq \prod_{i=1}^n G_{ii}^2 = |S| \ . \tag{27}$$

The equality holds if and only if $S_{ii} = G_{ii}^2$ for $i = 1, 2, \ldots, N$, which is equivalent to the fact that S is diagonal. Therefore, we know for any orthonormal transform E,

$$\left|\text{diag}(E^t S E)\right| \geq \left|E^t S E\right| = |S| \ . \tag{28}$$

If $S = E^* \Lambda_S E^{*t}$ is the eigen-decomposition of S, then we know

$$\left|\text{diag}(E^{*t} S E^*)\right| = |\Lambda_S| = |S| \ . \tag{29}$$

Therefore, E^* is the solution of global minimization of (1).

C Exact SMT Factorization of Orthonormal Transforms

We know the Givens QR factorization can be used to find a decomposition of an $N \times N$ matrix into $\binom{N}{2}$ Givens rotations [15]. Let A be an $N \times N$ orthonormal matrix, and let $Q = G_1 G_2 \ldots G_K$ with $K = \binom{N}{2}$, so that

$$A = QR \ ,$$

where every G_k is a Givens rotation and R is upper triangular. Since A and Q are orthonormal, R must be orthonormal. Therefore, the columns of R must be orthogonal. Since R is also upper triangular, this means that it must be diagonal. Therefore, R is a diagonal orthonormal matrix, which means that it is the identity matrix. Hence, we have $A = Q$.

D Solution of (7) for a Specified Coordinate Index Pair

In this appendix, we will find the solution to the optimization problem of (7) for a specified coordinate index pair and the corresponding change of the cost function. Since the coordinate index pair is specified, we can assume all the matrices to be 2×2 without loss of generality.

From Appendix B, we know that E^* minimizes the cost function (7) if and only if E^* is the matrix of eigenvectors of S, i.e. $S = E^* \Lambda_S E^{*t}$. Next we obtain an expression for E^* in terms of a Givens rotation. Let

$$S = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} , \tag{30}$$

and let $E^* = I + \Theta(1, 2, \theta)$ with $\theta = \frac{1}{2}\text{atan}(-2s_{12}, s_{11} - s_{22})$. Then we have

$$E^{*t} S E^* = \begin{bmatrix} s'_{11} & 0 \\ 0 & s'_{22} \end{bmatrix} , \tag{31}$$

where

$$s'_{11} = \frac{1}{2}\left(s_{11} + s_{22} + \sqrt{(s_{11} - s_{22})^2 + 4s_{12}^2} \right) \tag{32}$$

$$s'_{22} = \frac{1}{2}\left(s_{11} + s_{22} - \sqrt{(s_{11} - s_{22})^2 + 4s_{12}^2} \right) . \tag{33}$$

This shows that E^* of the given form is the matrix of eigenvectors of S. Hence E^* must minimize the cost function of (7). Based on (29), we know that the ratio of the cost function before and after the transform of E^* is given as

$$\frac{|\text{diag}\,(E^{*t} S E^*)|}{|\text{diag}\,(S)|} = \frac{|S|}{|\text{diag}\,(S)|} = 1 - \frac{s_{12}^2}{s_{11} s_{22}} . \tag{34}$$

E Review of Commonly Used Shrinkage Estimators

A general approach to the problem of covariance estimation for high dimensional data is the use of shrinkage estimators. Shrinkage estimators work by shrinking the sample covariance matrix toward some target structures

$$\hat{R} = \alpha D + (1 - \alpha) S ,$$

where D is some positive definite matrix and S is the sample covariance matrix.

The first method, which we will refer to as "Shrinkage Estimator I" (SE I) in the experimental results, sets D equal to the indentity matrix [16, 7].

$$\hat{R} = \alpha I + (1 - \alpha) S .$$

This identity regularization is usually used in ridge regression and Tikhonov regularization.

The second method, which we will refer to as "Shrinkage Estimator II" (SE II) in the experimental results, sets D equal to an scaled indentity matrix [5, 17, 9].

$$\hat{R} = \alpha \left(\frac{\mathrm{tr}(S)}{N} \right) I + (1-\alpha) S \ .$$

This approach scales the identity matrix by the average of the eigenvalues, so it tends to decrease the large eigenvalues and increases the small ones.

The third method, which we will refer to as "Shrinkage Estimator III" (SE III) in the experimental results, sets D equal to the diagonal of S [6, 10].

$$\hat{R} = \alpha \mathrm{diag}\left(S\right) + (1-\alpha) S \ .$$

For these three methods, we use the leave-one-out likelihood to choose an appropriate value for shrinkage intensity parameter α. Specifically, the value of α is chosen so that the average likelihood of omitted samples is maximized as suggested in [6],

$$\begin{aligned} \alpha^* &= \arg\max_{\alpha} \left\{ \frac{1}{M} \sum_{k=1}^{M} \log \left(p_{\hat{R}_k(\alpha)}\left(y_k\right) \right) \right\} & (35) \\ &= \arg\max_{\alpha} \left\{ -\frac{1}{2M} \sum_{k=1}^{M} \left[y_k^t \hat{R}_k(\alpha)^{-1} y_k + \log \left| \hat{R}_k(\alpha) \right| + N \log(2\pi) \right] \right\} \ , & (36) \end{aligned}$$

where $\hat{R}_k$ is the estimated covariance matrix without sample y_k. To do this, we search for α^* in the interval $[0.05, 1]$ with a step size of 0.05. Once the value of α^* is determined, $\hat{R}(\alpha^*)$ is computed using all the training samples. However, this leave-one-out cross-validation approach is very computationally expensive since we need to compute the inverse of $\hat{R}_k(\alpha)$ for every k and α. Some approximations can be used to improve the computation speed [5, 6]; however, these approximations still can reduce accuracy. With these approximations, the computation is comparable to the proposed SMT method.

For these three shrinkage estimators, the eigenvector and eigenvalue estimates, $(\hat{E}, \hat{\Lambda})$, are given by the eigen decomposition of the covariance estimator $\hat{R}(\alpha^*)$

$$\hat{R}(\alpha^*) = \hat{E} \hat{\Lambda} \hat{E} \ . \tag{37}$$

F Kullback-Leibler Distance

The Kullback-Leibler distance for two distributions $p_\theta(y)$ and $p_{\hat{\theta}}(y)$ is defined as [18]

$$d(\theta, \hat{\theta}) = E_\theta \left[\log p_\theta(Y) - \log p_{\hat{\theta}}(Y) \right] \ .$$

So if $\theta = (E, \Lambda)$ and $\hat{\theta} = (\hat{E}, \hat{\Lambda})$, then

$$d(\theta, \hat{\theta}) = E_\theta \left[\log p_\theta(Y) - \log p_{\hat{\theta}}(Y) \right]$$

$$
\begin{aligned}
&= -\frac{1}{2}\mathrm{tr}\{\mathrm{diag}(E^tRE)\Lambda^{-1}\} - \frac{1}{2}\log|\Lambda| + \frac{1}{2}\mathrm{tr}\{\mathrm{diag}(\hat{E}^tR\hat{E})\hat{\Lambda}^{-1}\} + \frac{1}{2}\log\left|\hat{\Lambda}\right| \\
&= \frac{1}{2}\mathrm{tr}\{\mathrm{diag}(\hat{E}^tR\hat{E})\hat{\Lambda}^{-1} - I\} + \frac{1}{2}\log\left|\hat{\Lambda}\Lambda^{-1}\right|
\end{aligned}
$$

We will use the Kullback-Leibler distance as one of the measures for the various covariance estimators.

References

[1] C. Stein, B. Efron, and C. Morris, "Improving the usual estimator of a normal covariance matrix," Dept. of Statistics, Stanford University, Report 37, 1972.

[2] T. Anderson, *An introduction to multivariate statistical analysis.* New York: John Wiley & Sons, Inc., 1984, 2nd Ed.

[3] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Boston, MA: Academic Press, 1990, 2nd Ed.

[4] A. K. Jain, R. P. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[5] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.

[6] J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 763–767, 1996.

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

[8] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of Empirical Finance*, vol. 10, no. 5, pp. 603–621, 2003.

[9] ——, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivar. Anal.*, vol. 88, no. 2, pp. 365–411, 2004.

[10] J. S. 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. 1, 2005.

[11] G. Cao, C. A. Bouman, and K. J. Webb, "Non-iterative map reconstruction using sparse matrix representations," *IEEE Trans. on Image Processing*, submitted.

[12] W. Givens, “Computation of plane unitary rotations transforming a general matrix to triangular form,” *Journal of the Society for Industrial and Applied Mathematics*, vol. 6, no. 1, pp. 26–50, March 1958.

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

[14] J. W. Cooley and J. W. Tukey, “An algorithm for the machine calculation of complex Fourier series,” *Mathematics of Computation*, vol. 19, no. 90, pp. 297–301, April 1965.

[15] G. H. Golub and C. F. Van Loan, *Matrix Computations.* Baltimore, MD: The John Hopkins University Press, 1989.

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

[17] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West, “Sparse graphical models for exploring gene expression data,” *J. Multivar. Anal.*, vol. 90, no. 1, pp. 196–212, 2004.

[18] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.