Exploration of Regularized Covariance Estimates with Analytical Shrinkage Intensity for Producing Invertible Covariance Matrices in High Dimensional Hyperspectral Data.

S. J. Walsh
M. F. Tardiff

October 2007

 This document was printed on recycled paper.
(9/2003)

Exploration of Regularized Covariance Estimates with Analytical Shrinkage Intensity for Producing Invertible Covariance Matrices in High Dimensional Hyperspectral Data.

S. J. Walsh
M. F. Tardiff

October 2007

Executive Summary

Remote sensing methods are a significant element of nonproliferation programs aimed at detecting activities and materials associated with developing nuclear capabilities. Long wave infrared spectroscopy is capable of detecting materials of interest from substantial distances. The effectiveness of these measurement systems is at least partially controlled by the signal to noise ratio of the image and how well the analysis algorithms remove background clutter and preserve signal. Having a covariance matrix that is invertible is central to performing this separation.

The goal of the Spectral Library Statistical Analysis (SLSA) project is to improve clutter removal using information gleaned from laboratory-collected spectral libraries. Due to the small number of repeated measurements in spectral libraries relative to the number of spectral channels, classical methods for approximating covariance matrices are not adequate.

The goal of this study is to investigate shrinkage estimation as an approach to providing an invertible covariance matrix estimate in the case where the number of spectral measurements is less than the number of spectral channels. The approach is an analytic method for arriving at a target matrix and shrinkage parameter that modify the degenerate covariance matrix obtained from the data to make it invertible. The theory is discussed to develop different estimates. The resulting estimates are computed and inspected on a set of hyperspectral data. This technique shows some promise for arriving at an invertible covariance estimate for small hyperspectral data sets.

Acknowledgements

We thank Larry Chilton and Kevin Anderson their mathematical and technical advice.

Contents

List of Figures

List of Tables

1.0 Introduction

Remote sensing methods are a significant element of nonproliferation programs aimed at detecting activities and materials associated with developing nuclear capabilities. Long wave infrared spectroscopy is capable of detecting materials of interest from substantial distances. The effectiveness of these measurement systems is at least partially controlled by the signal to noise ratio of the image and how well the analysis algorithms remove background clutter and preserve signal. Having a covariance matrix that is invertible is central to performing this separation.

Analyzing hyperspectral images for weak signals is a challenging problem because the spectral structure associated with the background of a scene can overwhelm the signal associated with the material of interest. There are a variety of methods for whitening the image by removing the background or clutter and working with the estimated net signal image. A common approach is to use all the pixels in a scene to estimate a mean vector and covariance matrix and to use these two quantities for the background subtraction. This approach works best with an image that has a homogeneous or stationary background. To the extent that the background is not stationary, i.e., there is a trend in the background or the background consists of a mosaic of different types, the subtraction performed with scene mean and covariance matrix results in biases that can remove the material of interest in some parts of the scene and fail to remove clutter in other parts of the scene.

One alternative to the whole-scene approach to background removal is to identify background types in the scene and attempt to conduct specific clutter removals. This process can be enhanced by developing a library of background types to support the estimation of mean vectors and background-specific covariance matrices.

The new difficulty that emerges in developing such a library is in estimating the covariance matrix or more specifically the inverse of the covariance matrix. Hyperspectra can consist of hundreds of channels. The whitening process requires an estimate of the inverse of the covariance matrix to perform the operations necessary to generate the estimated net signal scene. The original data must have significantly more spectra than spectral channels to use the standard approach for inverting the covariance matrix. So, if the spectra have 250 channels, then there must be many more than 250 measurements to get the inverse directly. It is rare for a spectral library to have that many repeated measurements of background

materials. Consequently, alternative methods are necessary to estimate the inverse of the covariance matrix when the number of observations is much less than the number of spectral channels.

The literature is rich with estimates of the covariance matrix that attempt to mitigate the non-invertibility problem, but most papers that have been reviewed to date deal with this problem in some ad hoc setting. As an example, Landgrebe et al. (Hoffbeck and Landgrebe 1996; Kuo and Landgrebe 2001) have done much work in image segmentation. They were confronted with a singular covariance matrix that can't be used in a discrimination rule when the number of pixels available for a training set was small. They provided many estimates that correct this problem and assessed their usefulness by applying them in the discrimination rules and recording the misclassification rates. While this is common practice, it lacks independent verification that the estimated inverse is trustworthy.

A common theme encountered in this problem is the use of regularization (also known as shrinkage estimation). Regularization is essentially taking a weighted average of some standard estimate and some other (possibly appropriate) estimate known as the target(s). The idea is that because of the ill-posed nature of the problem, the standard estimate lacks some desired property. A target estimate is chosen that has the desired property, and the standard estimate is shrunk toward the target and thus inherits the property. In the case where the number of measurements is less than the spectral dimension, the empirical covariance matrix is not invertible. Therefore, an invertible target matrix is chosen. The weighted average of the empirical covariance and the invertible target matrix is then invertible (though it is now biased).

Following the path of Ledoit and Wolf, Schäfer and Strimmer provide a way of arriving at a regularized estimate of the covariance matrix with the goal that the estimate be positive definite and well conditioned. A well-conditioned estimate of the covariance matrix provides stable calculation of its inverse.

2.0 Formal Statement of the Problem

Let $\mathbf{y}_i : p \times 1$ for $i = 1, 2, \ldots, n$ denote a set of n hyperspectral measurements of dimension p (number of channels) and let $n < p$. That is, the number of observations is far less than the

spectral dimension. In this case, the empirical covariance matrix $\mathbf{S} : p \times p$, though unbiased for the population covariance matrix, is positive semi-definite and hence not invertible.

The approach for finding a suitable and invertible estimate of the covariance matrix considered here involves finding an estimate for the spectral covariance matrix that is not only invertible but also well conditioned. The idea is that providing a well-conditioned estimate of the spectral covariance matrix will also provide a good estimate of the inverse spectral covariance.

The class of regularized estimates considered has the form

$$\mathbf{S}^* = \lambda \mathbf{T} + (1 - \lambda)\mathbf{S} \qquad (2.1)$$

where $\mathbf{S}$ is the empirical covariance, $\mathbf{T} : p \times p$ is the shrinkage target, and $\lambda \in [0, 1]$ is the shrinkage parameter. It is important to note that when $n < p$, $\mathbf{S}$ is only positive semi-definite as a number of eigenvalues are zero. It is required that $\mathbf{T}$ be positive definite and $0 < \lambda \leq 1$ for $\mathbf{S}^*$ to be positive definite and invertible.

A typical issue with regularized estimates is how one arrives at an optimal value for λ in (2.1). Often it has to be approximated using some type of expensive computational algorithm, such as cross validation or Markov Chain Monte Carlo simulation. Schäfer and Strimmer (2005) provide an analytical way of arriving at a value for λ. The domain of their work is functional genomics where the dimension of their data often exceeds the number of observations available. They set out to provide a positive definite and well-conditioned estimate of the population covariance matrix. Their approach is explained in the next section.

3.0 Analytic Estimation of the Shrinkage Parameter

Schäfer and Strimmer's method for finding the optimal λ is based on a theorem set in a decision theoretic perspective. They define a loss function for using the shrinkage estimate. The loss function is a distance metric that quantifies how far the shrinkage estimate is from the true covariance matrix. They use the squared Frobenius norm which is the matrix analogue of squared error loss. The loss function is viewed as a function of λ and can be expressed as

$$\text{L}(\lambda) = \|\mathbf{S}^* - \mathbf{\Sigma}\|_F^2 = \sum_{i=1}^{p}\sum_{j=1}^{p}(\lambda t_{ij} + (1-\lambda)s_{ij} - \sigma_{ij})^2 \qquad (3.1)$$

where t_{ij}, s_{ij}, and σ_{ij} are the ij^{th} entries in $\mathbf{T}$, $\mathbf{S}$, and the true covariance, $\mathbf{\Sigma}$, respectively. Theoretically, we are interested in how much we lose on average when we use the shrinkage estimate. This leads to use of the risk function, which is the expected value of the loss function. Taking the expected value of equation (3.1) yields

$$\text{R}(\lambda) = \text{E}[\text{L}(\lambda)] = \sum_{i=1}^{p}\sum_{j=1}^{p}\text{E}[(\lambda t_{ij} + (1-\lambda)s_{ij} - \sigma_{ij})^2]. \qquad (3.2)$$

We would like to find the value of λ that gives us the minimum risk. Optimizing equation (3.2) with respect to λ yields the solution

$$\lambda^* = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p}\text{Var}(s_{ij}) - \text{Cov}(t_{ij}, s_{ij})}{\sum_{i=1}^{p}\sum_{j=1}^{p}\text{E}[(t_{ij} - s_{ij})^2]}. \qquad (3.3)$$

It can be shown that equation (3.3) is a minimizer of (3.2); thus, λ^* is called the *optimal shrinkage intensity*. Note that (3.3) is optimal regardless of the sample size n, so it may be applied when $n < p$. It is also important to note that (3.3) is a theoretical value for the optimal shrinkage intensity, and it must be estimated from the data. Details are provided in Shäfer and Strimmer (2005).

4.0 The Target Shrinkage Matrix

Estimating a full covariance matrix $\mathbf{\Sigma} : p \times p$ that has $\frac{p(p+1)}{2}$ separate parameters constitutes a high dimensional inference problem. The purpose of the target matrix $\mathbf{T}$ is to reduce the number of estimated parameters and help produce a positive definite and well-conditioned shrinkage estimate $\mathbf{S}^*$.

The suggestion for providing the shrinkage target $\mathbf{T}$ is that it should represent the lower dimensional covariance structure expected in the data. As an example, suppose that there are $p = 3$ spectral bands, and we expect the spectral bands to be uncorrelated but each have a different variance. Then the true covariance matrix could be stated as

$$\Sigma = \begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{pmatrix}$$

and an appropriate estimate and shrinkage target would be

$$\mathbf{T} = \begin{pmatrix} s_{11} & 0 & 0 \\ 0 & s_{22} & 0 \\ 0 & 0 & s_{33} \end{pmatrix}$$

where s_{ii} is the empirical variance of the ith spectral channel. Schäfer and Strimmer assert that providing a correct shrinkage target will lead to the greatest increase of efficiency when using $\mathbf{S}^*$ as an estimate for $\mathbf{\Sigma}$ as apposed to the empirical covariance matrix $\mathbf{S}$. They also state that selecting a suitable target "requires some diligence," which will be explored in the next section.

5.0 Application to Hyperspectral Data

A hyperspectral image was selected to explore the shrinkage estimator approach. A 100×80 subset of the image was selected where it appeared to be homogeneous. Thus, all of the pixels in the subset image are expected to have similar spectral signatures. The data have 258 spectral channels. The covariance matrix over all 8000 sample pixels was inspected to give clues as to what a reasonable shrinkage target may be. Figure 1 shows an image plot of the covariance matrix. Inspection of the full covariance matrix seems to give little or no clues as to what a suitable target matrix may be. Keep in mind that the regularization method being considered works by specifying very few parameters in the target matrix. This being the case, four commonly used targets are considered in Table 1.

The data were sampled to get a 258×25 subset for testing the shrinkage targets in Table 1. All shrinkage targets in Table 1 were positive definite when computed on a sample of 25 pixels. Figure 2 shows an image plot of the shrinkage covariance estimates using each target. Note that all estimates exhibit similar structure as compared to the full covariance matrix, though they may differ in the magnitudes of the entries. Even though the pixels in the image were expected to be homogeneous, the pixels can not be considered exact replicates. To see

Figure 1: Covariance Matrix of all 8000 Pixels.
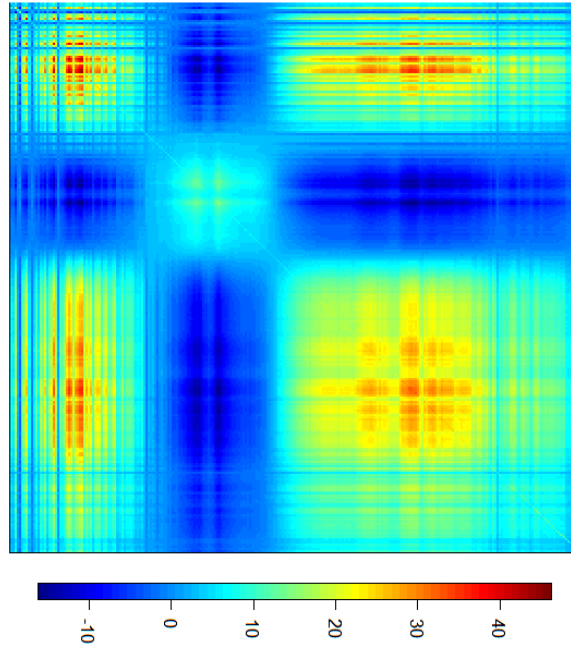
how this may affect the results, 100 random samples of size 25 were taken from the 8000 total pixels. The shrinkage estimates were computed on each sample using each of the four targets. The condition number of each shrinkage estimate was computed for each sample. Figure 3 provides a boxplot of the condition numbers of the four shrinkage estimates over the 100 samples.
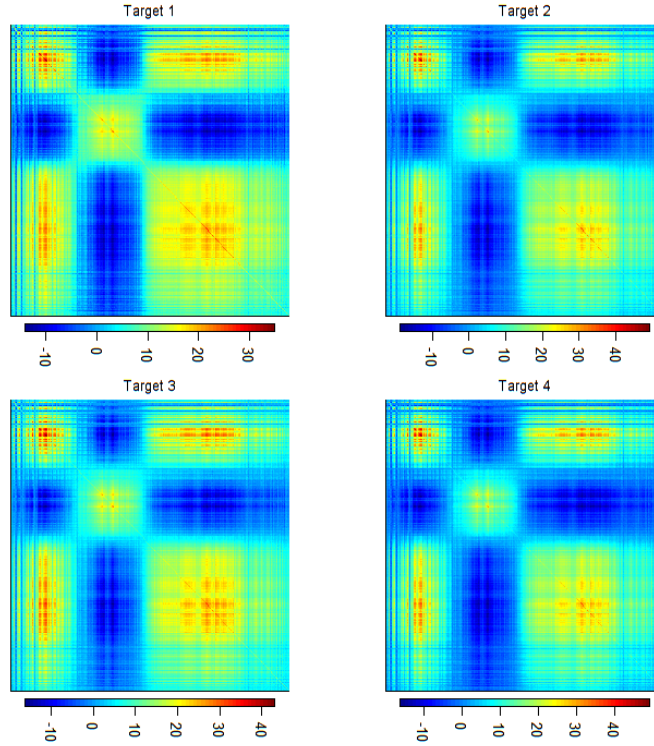
11

Figure 2: Image Plots of each Shrinkage Estimate Using the Four Target Matrices.
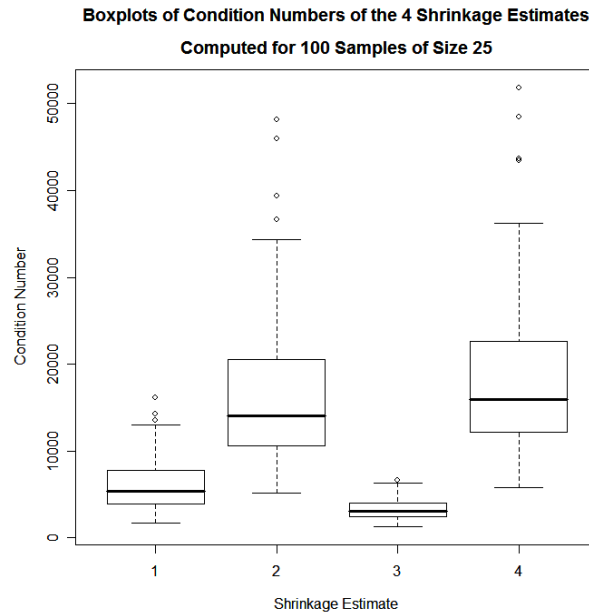

Figure 3: Boxplots of Observed Condition Numbers Computed for the Four Shrinkage Estimates on 100 Samples of Size 25. Note that target 3 seems to produce well-conditioned estimates.

Target 1: "Diagonal Common Variance" 1 estimated parameter	**Target 2:** "Diagonal Unequal Variance" p estimated parameters
$$t_{ij} = \begin{cases} m = median(s_{ii}) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$ $$\hat{\lambda}^* = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) + \sum_i \widehat{\text{Var}}(s_{ii})}{\sum_{i \neq j} s_{ij}^2 + \sum_i (s_{ii} - m)^2}$$	$$t_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$ $$\hat{\lambda}^* = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij})}{\sum_{i \neq j} s_{ij}^2}$$
Target 3: "Common (Co)variance" 2 estimated parameters	**Target 4** "Constant Correlation" $p + 1$ estimated parameter
$$t_{ij} = \begin{cases} v = avg(s_{ii}) & \text{if } i = j \\ c = avg(s_{ij}) & \text{if } i \neq j \end{cases}$$ $$\hat{\lambda}^* = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) + \sum_i \widehat{\text{Var}}(s_{ii})}{\sum_{i \neq j}(s_{ij} - c)^2 + \sum_{i=j}(s_{ii} - v)^2}$$	$$t_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ \bar{r}\sqrt{s_{ii}s_{jj}} & \text{if } i \neq j \end{cases}$$ $$f_{ij} = \tfrac{1}{2}\left\{ \sqrt{\tfrac{s_{jj}}{s_{ii}}}\widehat{\text{Cov}}(s_{ii}, s_{ij}) + \sqrt{\tfrac{s_{ii}}{s_{ij}}}\widehat{\text{Cov}}(s_{jj}, s_{ij}) \right\}$$ $$\hat{\lambda}^* = \frac{\sum_{i \neq j} \widehat{\text{Var}}(s_{ij}) - \bar{r}f_{ij}}{\sum_{i \neq j}(s_{ij} - \bar{r}\sqrt{s_{ii}s_{jj}})^2}$$

Table 1: Four Shrinkage Targets Explored on the Hyperspectral Data and Their Associated Estimators of the Optimal Shrinkage Intensity.

6.0 Conclusions and Suggestions for Further Study

The covariance matrix in Figure 1 was computed on 8000 pixels each with 258 spectral channels. This matrix was well conditioned and thus would lend itself to computing a stable inverse. The resulting inverse would then be used for whitening or discrimination analyses. The regularized estimates in Figure 2 show similar structure to that of Figure 1. These covariance matrices were, however, computed using only 25 pixels each with 258 spectral channels. As a result of using regularization, each of the estimates in Figure 2 is positive definite and invertible, which is not a property of the empirical covariance matrix **S** in this case. Thus, while **S** cannot be used for whitening, a well-chosen regularized estimate could be both invertible and well conditioned and therefore suitable for whitening.

The conditioning of the regularized estimates is an important consideration since a well-conditioned estimate will yield stable calculation of its inverse. One way to measure matrix conditioning is the condition number–the smaller the condition number, the better conditioned the matrix. Figure 3 gives an example of the variability of the condition number of

13

the regularized estimates when computed over 100 samples of size 25. It is apparent that the estimate using shrinkage target 3 is able to produce well-conditioned estimates and that its condition number exhibits the smallest variability of the four estimates considered for these data. Hence the regularized estimate using target 3 may have the potential to produce the most reliable results in a whitening or discrimination analysis.

This study was only exploratory in nature, and it proved difficult to try and quantify how "good" any of the four shrinkage estimates were for approximating the true covariance (the true covariance is not known).

Further study is warranted and may consist of the following:

1. Literature search to elucidate commonly found structures of covariance matrices of hyperspectral data to in turn give the correct target matrix.

2. Build a statistical model that produces data similar to what is observed. A "true" covariance matrix can be specified to generate the data, and then the shrinkage estimates can be compared to the truth.

3. Explore other types of shrinkage estimates as in Hoffbeck and Landgrebe (1996) and Kuo and Landgrebe (2001).

4. Explore the methods proposed by Berge et al. (2007) and Levina et al. (2007) to obtain an estimate of the inverse covariance matrix directly.

References

[1] Berge,A., Jensen, A. C, and Solberg, A. H. S. (2007) Sparse inverse covariance estimates for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 45*, 5.

[2] Hoffbeck, J.P. and Landgrebe,D.A. (1996) Covariance matrix estimation and classification with limited training data. *IEEE Transactions on Pattern Analysis and Machine Intelligence,18*, 7, 763-767.

[3] Kuo, Bor-Chen and Landgrebe, David (2001) Improved statistics estimation and feature extraction for hyperspectral data classification. *Electrical and Computer Engineering Technical Reports*. Purdue University. Available at: http://docs.lib.purdue.edu/ecetr/10, accessed October 20, 2007.

[4] Ledoit, O. and Wolf, M. (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance 10,* 603-621.

[5] Ledoit, O. and Wolf,M. (2004) A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88,* 365-411.

[6] Levina, E., Rothma, A., and Zhu, J. (2007). Sparse estimation of large covariance matrices via a nested lasso penalty. To appear in the *Annals of Applied Statistics.*

[7] Opgen-Rhein, R and Strimmer, K. (2007) Accurate ranking of differentially expressed genes by a distribution free shrinkage approach. *Statistical Applications in Genetics and Molecular Biology. 6,* 9.

[8] Schäfer, J. and Strimmer,K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology. 4,* 32.