The incredible shrinking covariance estimator

James Theiler

Space Data Systems Group,
Los Alamos National Laboratory, Los Alamos, NM, USA

ABSTRACT

Covariance estimation is a key step in many target detection algorithms. To distinguish target from background requires that the background be well-characterized. This applies to targets ranging from the precisely known chemical signatures of gaseous plumes to the wholly unspecified signals that are sought by anomaly detectors. When the background is modelled by a (global or local) Gaussian or other elliptically contoured distribution (such as Laplacian or multivariate-t), a covariance matrix must be estimated. The standard sample covariance overfits the data, and when the training sample size is small, the target detection performance suffers.

Shrinkage addresses the problem of overfitting that inevitably arises when a high-dimensional model is fit from a small dataset. In place of the (overfit) sample covariance matrix, a linear combination of that covariance with a fixed matrix is employed. The fixed matrix might be the identity, the diagonal elements of the sample covariance, or some other underfit estimator. The idea is that the combination of an overfit with an underfit estimator can lead to a well-fit estimator. The coefficient that does this combining, called the shrinkage parameter, is generally estimated by some kind of cross-validation approach, but direct cross-validation can be computationally expensive.

This paper extends an approach suggested by Hoffbeck and Landgrebe, and presents efficient approximations of the leave-one-out cross-validation (LOOC) estimate of the shrinkage parameter used in estimating the covariance matrix from a limited sample of data.

Keywords: covariance matrix, shrinkage, regularization, cross-validation, LOOC

1. INTRODUCTION

Models that "try too hard" to fit a limited set of data samples are often victims of their own success. They may fit the available data all too well, but then fail to fit data that wasn't available when the model was being constructed. The new data just doesn't seem to exhibit the exquisite structure that was painstakingly extracted from the original dataset! This is the overfitting problem, and it is a staple of modern statistics. In adapting a model to fit data, one must at the same time adapt the model's *complexity* to be appropriate to the data. A model that is too complex will *find* structure that isn't there and will *overfit* the data. A model that is too simple will miss structure that *is* there and will *underfit* the data. Shrinkage is a statistical scheme that combines a nominally overfit model (papa bear's too-hot soup) with one that is manifestly underfit (mama bear's too-cold soup) to produce a model that is (for Goldilocks' purposes) "just right."

The covariance matrix is a particularly useful candidate for shrinkage (also called "regularization") because the sample covariance – defined explicitly in Eq. (2) below – provides a natural overfit estimator. Often just a small amount of shrinkage (which alters the individual matrix elements by only a tiny fraction of a percent) is enough to dramatically improve the quality of regression, classification, and detection problems. This is because the sample covariance can be singular or near-singular, and it is actually the inverse covariance matrix that is the active ingredient in many of these algorithms.

Efforts at finding better covariance estimators have appeared in the statistical,[1–4] financial,[5] genomics,[6] and remote sensing[7–14] literature. While shrinkage is the most common regularization scheme, sparse[15,16] and sparse transform[17–20] methods have also been proposed. To deal with the non-Gaussian nature of most data, both robust[21] and anti-robust[22] estimators have been proposed.

2. COVARIANCE MATRIX ESTIMATION

Given n samples, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, from a p-dimensional multivariate Gaussian distribution, our goal is to estimate the mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $R \in \mathbb{R}^{p \times p}$ that parameterize that distribution:

$$p(\mathbf{x}; \boldsymbol{\mu}, R) = (2\pi)^{-p/2} |R|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T R^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]. \tag{1}$$

In what follows, we will assume a zero mean Gaussian; *i.e.,* that $\boldsymbol{\mu} = 0$. Although this is often not realistic, it permits a simplified exposition that still contains the important part of the problem. Appendix A derives the more common (but more complicated and no more illuminating) case in which the mean is estimated from the data.

We will estimate the covariance maxtrix R with a linear combination of overfit and and underfit estimators. The overfit estimator is the sample covariance

$$S = (1/n) \sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^T. \tag{2}$$

It can be shown (*e.g.,* see Ref. [17] for a clear derivation) that the sample covariance is the maximum likelihood estimator for R. In the limit of large n, the sample covariance approaches the true covariance. Several choices are available for the underfit estimator, but at this point we will simply call our underfit estimator T without yet specifying what it is. We can now write a shrinkage estimator:

$$R_\alpha = (1 - \alpha) S + \alpha T. \tag{3}$$

The "trick" is to find an appropriate value of α. When n is large, then S is a very good estimator, and α should be small. But if n is small, then S will be more subject to the small-number fluctuations in its estimate, and a larger value of α will be appropriate. If we knew the true R, then we could pick the best α in terms of how closely R_α approximated R. An appropriate way to compare closeness[†] of covariance matrices is in terms of the likelihood function in Eq. (1). If a large number of samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are drawn from the distribution in Eq. (1), then their likelihood under the assumption that the covariance is given by R_α is

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n; R_\alpha) = (2\pi)^{-np/2} |R_\alpha|^{-n/2} \exp\left[-\frac{1}{2} \sum_{k=1}^{n} \mathbf{x}_k^T R_\alpha^{-1} \mathbf{x}_k \right]. \tag{4}$$

Note that the scalar value $\mathbf{x}^T R^{-1} \mathbf{x}$ can be written $\mathrm{trace}\big(\mathbf{x}^T R^{-1} \mathbf{x}\big) = \mathrm{trace}\big(R^{-1} \mathbf{x}\mathbf{x}^T\big)$ which can be used to re-express Eq. (4) as

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n; R_\alpha) = (2\pi)^{-np/2} |R_\alpha|^{-n/2} \exp\left[-\frac{n}{2} \mathrm{trace}\big(R_\alpha^{-1} S\big) \right] \tag{5}$$

with S given by Eq. (2). In the large n limit, $S \to R$, and we can express the dissimilarity of R and R_α with the average of the negative logarithm of this likelihood:

$$\mathcal{L}(R, R_\alpha) = \lim_{n \to \infty} \frac{-\log p(\mathbf{x}_1, \ldots, \mathbf{x}_n; R_\alpha)}{n} = \frac{1}{2} \left[p \log(2\pi) + \log |R_\alpha| + \mathrm{trace}\big(R_\alpha^{-1} R\big) \right]. \tag{6}$$

It follows that the best shrinkage estimator is given by

$$\alpha^* = \mathrm{argmin}_\alpha \ \left[\log |R_\alpha| + \mathrm{trace}\big(R_\alpha^{-1} R\big) \right]. \tag{7}$$

[†]Another way to compare closeness of covariance matrices is with the Kullback-Leibler distance between the two Gaussian distributions that correspond to the two covariance matrices. It is shown in Ref. [17] that this distance is related to the negative log likelihood approach developed in the text. In particular, the minimization in Eq. (7) produces the same result as minimizing the Kullback-Leibler distance.

2.1 Leave-one-out cross-validation

The problem with trying to use Eq. (7) in a practical setting is that the true covariance R is not available. If we try to use S in place of R in Eq. (7), we obtain the trival (and incorrect) result that $\alpha = 0$. That's because we use the same data both to estimate the true covariance *and* to estimate the likelihood. The idea of cross-validation is to partition the data into two independent sets of samples. We use one set to estimate R and the other set to estimate the likelhood.

In particular, for leave-one-out cross-validation, we will estimate R from all but one of the samples, and estimate the likelihood from the remaining sample. We will do this for each of the samples, and average. Let S_k correspond to the sample covariance that would be estimated with all but $\mathbf{x}_k$:[†]

$$S_k = \frac{1}{n-1} \sum_{j=1, j\neq k}^{n} \mathbf{x}_j \mathbf{x}_j^T = \frac{n}{n-1} S - \frac{1}{n-1} \mathbf{x}_k \mathbf{x}_k^T. \tag{8}$$

Several choices are available for the shrinkage target T. Perhaps the most popular of these choices is the ridge regularizer[1,2] which takes $T = \sigma^2 I$, where σ^2 is a scalar constant corresponding to the overall variance in the data. An appropriate value is $\sigma^2 = \mathrm{trace}(S)/p$; that way S and $T = \sigma^2 I$ have the same overall variance. Another choice, popularized by Hoffbeck and Landgrebe,[7] takes the diagonal components of the sample covariance: $T = \mathrm{diag}(S)$.[‡] Other choices have also been considered. The target detection performance of the sparse matrix transform (SMT)[18] was found to be improved if, instead of using the SMT as a replacement for the sample covariance, it was used as a shrinkage target T in conjunction with the sample covariance.[19]

All of these choices depend on the sample covariance S, and for that reason we ought to in principle consider T_k as the shrinkage target that would be used when all but the kth sample is available. For this exposition, however, we will follow Hoffbeck and Landgrebe[7] who argue on empirical grounds that it can be treated as a constant. It also makes sense, theoretically, that a T_k chosen specifically to be an underfit estimator should exhibit only minor dependence on k (certainly much less dependence than the overfit sample covariance S_k).

The likelihood of observing $\mathbf{x}_k$, given $R_{\alpha,k} = (1-\alpha)S_k + \alpha T$, is given by

$$p(\mathbf{x}_k, R_{\alpha,k}) = (2\pi)^{-p/2} |R_{\alpha,k}|^{-1/2} \exp\left[-\frac{1}{2} \mathbf{x}_k^T R_{\alpha,k}^{-1} \mathbf{x}_k\right] \tag{9}$$

The LOOC scheme averages the negative log likelihood of this expression:

$$\mathcal{L}(\alpha, \mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{1}{n} \sum_{k=1}^{n} -\log(x_k \mid R_{\alpha,k}) \tag{10}$$

$$= \frac{p \log(2\pi)}{2} + \frac{1}{2n} \sum_{k=1}^{n} \left(\log |R_{\alpha,k}| + \mathbf{x}_k^T R_{\alpha_k}^{-1} \mathbf{x}_k\right). \tag{11}$$

One computes this quantity for a range of α values, and chooses the α for which this is minimized. To compute $\mathcal{L}(\alpha, \mathbf{x}_1, \ldots, \mathbf{x}_n)$ directly, for any given α, requires one to compute both the determinant and the inverse of $R_{\alpha,k}$ for each k. This requires $O(np^3)$ effort.

But Hoffbeck and Landgrebe[7] showed that a clever manipulation of the expression in Eq. (11) leads to an equivalent expression that can be much more efficiently evaluated. To begin, write

$$R_{\alpha,k} = (1-\alpha)S_k + \alpha T = G_\alpha - \beta \mathbf{x}_k \mathbf{x}_k^T \tag{12}$$

where

$$G_\alpha = n\beta S + \alpha T \tag{13}$$

[†]See Eq. (50) in the Appendix for the adjustments that are made to the coefficients in the case that the mean is not assumed to be zero, but instead is estimated from the data.

[‡]Here, the function 'diag' takes an input matrix and sets its nondiagonal elements to zero; in Matlab, that would be written `T=diag(diag(S))`.

and

$$\beta = \frac{1-\alpha}{n-1}. \tag{14}$$

Next, two two useful identities are introduced. First, use the Sherman-Morrison-Woodbury formula to write

$$(G - \beta \mathbf{x}\mathbf{x}^T)^{-1} = G^{-1} + \beta \frac{G^{-1}\mathbf{x}\mathbf{x}^T G^{-1}}{1 - \beta \mathbf{x}^T G^{-1}\mathbf{x}}. \tag{15}$$

Multiply by $\mathbf{x}^T$ on the left and by $\mathbf{x}$ on the right to obtain

$$\mathbf{x}^T (G - \beta \mathbf{x}\mathbf{x}^T)^{-1}\mathbf{x} = \mathbf{x}^T \left[G^{-1} + \beta \frac{G^{-1}\mathbf{x}\mathbf{x}^T G^{-1}}{1 - \beta \mathbf{x}^T G^{-1}\mathbf{x}} \right] \mathbf{x} \tag{16}$$

$$= \mathbf{x}^T G^{-1}\mathbf{x} + \beta \frac{\mathbf{x}^T G^{-1}\mathbf{x}\mathbf{x}^T G^{-1}\mathbf{x}}{1 - \beta \mathbf{x}^T G^{-1}\mathbf{x}} \tag{17}$$

$$= r + \beta \frac{r^2}{1 - \beta r} = \frac{r}{1 - \beta r} \tag{18}$$

where we have written $r = \mathbf{x}^T G^{-1}\mathbf{x}$.

Second, we will write

$$\left| G - \beta \mathbf{x}\mathbf{x}^T \right| = |G| \cdot \left| I - \beta G^{-1}\mathbf{x}\mathbf{x}^T \right|. \tag{19}$$

Sylvester's determinant theorem says that $|I - AB| = |I - BA|$, and if we take $A = \beta G^{-1}\mathbf{x}$ and $B = \mathbf{x}^T$, then we can write

$$\left| I - \beta G^{-1}\mathbf{x}\mathbf{x}^T \right| = \left| 1 - \mathbf{x}^T \beta G^{-1}\mathbf{x} \right| = 1 - \beta r. \tag{20}$$

Thus:

$$\left| G - \beta \mathbf{x}\mathbf{x}^T \right| = |G| \, (1 - \beta r). \tag{21}$$

Combining these two identities enables us to express the negative log likelihood for observing the data point $\mathbf{x}$, given the estimated covariance $G - \beta \mathbf{x}\mathbf{x}^T$:

$$-\log p(\mathbf{x} \mid G - \beta \mathbf{x}\mathbf{x}^T) = \frac{1}{2} \left[p \log(2\pi) + \log |G| + \log(1 - \beta r) + \frac{r}{1 - \beta r} \right]. \tag{22}$$

In particular,

$$-\log p(\mathbf{x}_k \mid R_{\alpha,k}) = \frac{1}{2} \left[p \log(2\pi) + \log |G_\alpha| + \log(1 - \beta r_k) + \frac{r_k}{1 - \beta r_k} \right] \tag{23}$$

where $r_k = \mathbf{x}_k^T G_\alpha^{-1} \mathbf{x}_k$.

We can now express the LOOC in Eq. (11) in a much more convenient form, and obtain the remarkable result of Hoffbeck and Landgrebe:[7]

$$\begin{aligned} \mathcal{L}_{\mathrm{HL}}(\alpha, \mathbf{x}_1, \ldots \mathbf{x}_n) &= \frac{1}{n} \sum_{k=1}^{n} -\log p(\mathbf{x}_k \mid R_{\alpha,k}) \\ &= \frac{1}{n} \sum_{k=1}^{n} \frac{1}{2} \left[p \log(2\pi) + \log |G_\alpha| + \log(1 - \beta r_k) + \frac{r_k}{1 - \beta r_k} \right] \\ &= \frac{1}{2} \left[p \log(2\pi) + \log |G_\alpha| \right] + \frac{1}{2n} \sum_{k=1}^{n} \left[\log(1 - \beta r_k) + \frac{r_k}{1 - \beta r_k} \right]. \end{aligned} \tag{24}$$

Although this expression is equivalent to Eq. (11), the computation of determinant and inverse are only done once, to G_α. Once G_α^{-1} is precomputed, the computation of $r_k = \mathbf{x}_k^T G_\alpha^{-1} \mathbf{x}_k$ is only $O(p^2)$ and so the full computation requires only $O(p^3) + O(np^2)$. Further, the np^2 term has a very small coefficient because the dominant computation is given by the n matrix-vector multiplies $G_\alpha^{-1}\mathbf{x}_k$, one for each of n samples.

Table 1. Computational complexity for various estimators of LOOC log likelihood.

Estimator	Complexity
Nominal: Eq. (11)	$O(np^3)$
Hoffbeck and Landgrebe:[7] Eq. (24)	$O(p^3) + O(np^2)$
Monte-Carlo: Eq. (26)	$O(p^3) + O(n'p^2)$, with $1 \ll n' \ll n$
Mean Mahalanobis: Eq. (30)	$O(p^3)$

3. PROPOSED APPROXIMATIONS

Although the Hoffbeck-Landgrebe result in Eq. (24) is strictly equivalent to the LOOC defined in Eq. (11), it is still an approximations to the "true" negative log likelihood given in Eq. (6). What follows are two strategies for further approximating the LOOC, with the benefit of further reducing the computational effort in estimating the negative log likelihood.

3.1 Monte-Carlo approximation

First, we note that one way to interpret the sum in Eq. (24) is as a Monte-Carlo approximation to the expected value of

$$f(r) = \log(1 - \beta r) + \frac{r}{1 - \beta r}. \tag{25}$$

Thus, one easy way to approximate it is to take a smaller sample of $n' \ll n$ points and average over them.

$$\mathcal{L}_{\mathrm{MC}}(\alpha, \mathbf{x}_1, \ldots \mathbf{x}_n) \approx \frac{1}{2}\left(p \log(2\pi) + \log|G_\alpha|\right) + \frac{1}{2n'} \sum_{k=1}^{n'} \left[\log(1 - \beta r_k) + \frac{r_k}{1 - \beta r_k}\right]. \tag{26}$$

In particular, if we take $n' = O(p)$, we will be able to evaluate Eq. (26) in $O(p^3)$ time.

Further, Eq. (26) is an unbiased estimator, which means that on average (over many choices of the n' subsample points) it will equal the exact LOOC in Eq. (24). On the other hand, there will be some variance in this estimator due to the limited subsampling. In general, that variance will scale as $O(1/n')$. This variance is independent of the total number of samples n, so even as $n \to \infty$, the magnitude of this variance will not vanish.

3.2 Mean Mahalanobis approximation

We can estimate the mean of the Mahalanobis distance $r = \mathbf{x}^T G^{-1} \mathbf{x}$ by recognizing that $\mathbf{x}$ is generated from a Gaussian distribution with covariance S:

$$r_o = \langle\, r \,\rangle = \left\langle\, \mathbf{x}^T G^{-1} \mathbf{x} \,\right\rangle = \left\langle\, \mathrm{trace}\big(\mathbf{x}^T G^{-1} \mathbf{x}\big) \,\right\rangle = \left\langle\, \mathrm{trace}\big(G^{-1} \mathbf{x}\mathbf{x}^T\big) \,\right\rangle = \mathrm{trace}\big(G^{-1} \left\langle\, \mathbf{x}\mathbf{x}^T \,\right\rangle\big) = \mathrm{trace}\big(G^{-1} S\big) \tag{27}$$

The key idea is that a function of this mean r_o can be used to approximate the mean of the function. That is:

$$\langle\, f(r) \,\rangle \approx f\left(\langle\, r \,\rangle\right) = f(r_o). \tag{28}$$

Then we can write

$$\frac{1}{2n} \sum_{k=1}^{n} \left[\log(1 - \beta r_k) + \frac{r_k}{1 - \beta r_k}\right] \approx \frac{1}{2} \left[\log(1 - \beta r_o) + \frac{r_o}{1 - \beta r_o}\right], \tag{29}$$

from which

$$\mathcal{L}_{\mathrm{MM}}(\alpha, \mathbf{x}_1, \ldots \mathbf{x}_n) \approx \frac{1}{2} \left[p \log(2\pi) + \log|G_\alpha| + \log(1 - \beta r_o) + \frac{r_o}{1 - \beta r_o}\right]. \tag{30}$$

This enables us to evaluate the average leave-one-out negative log likelihood in $O(p^3)$ time, independent of n. Table 1 compares this computational complexity to the other estimators of LOOC log likelihood.

We can investigate the validity of the approximation in Eq. (28) by first writing the Taylor series expansion:

$$f(r) = f(r_o) + (r - r_o)f'(r_o) + \frac{1}{2}(r - r_o)^2 f''(r_o) + \cdots \tag{31}$$

The average of $f(r)$ can be written

$$\langle\, f(r)\,\rangle = f(r_o) + \langle\, (r - r_o)\,\rangle\, f'(r_o) + \frac{1}{2}\left\langle\, (r - r_o)^2\,\right\rangle f''(r_o) + \cdots \tag{32}$$

Since $r_o = \langle\, r\,\rangle$, then the second (linear) term vanishes:

$$\langle\, f(r)\,\rangle = f(r_o) + \frac{1}{2}\left\langle\, (r - r_o)^2\,\right\rangle f''(r_o) + \cdots \tag{33}$$

So the validity of this approximation depends on the size of the second term in Eq. (33). In particular, we have

$$f(r) = \log(1 - \beta r) + \frac{r}{1 - \beta r} \tag{34}$$

from which

$$f'(r) = \frac{-\beta}{1 - \beta r} + \frac{1}{(1 - \beta r)^2} = \frac{(1 - \beta) + \beta^2 r}{(1 - \beta r)^2}, \tag{35}$$

and

$$f''(r) = \frac{\beta(\beta^2 r - \beta + 2)}{(1 - \beta r)^3}. \tag{36}$$

A first observation is that $f''(r) > 0$, so from Eq. (33), we have that $\langle\, f(r)\,\rangle > f(r_o)$, which means that the mean Mahalanobis approximation always under-estimates the negative log likelihood. The magnitude of the error in negative log likelihood is well approximated by the second order term:

$$\mathcal{L}_{\text{HL}} - \mathcal{L}_{\text{MM}} = \langle\, f(r)\,\rangle - f(r_o) \approx \frac{1}{2}\left\langle\, (r - r_o)^2\,\right\rangle f''(r_o) = \frac{1}{2}\left\langle\, (r - r_o)^2\,\right\rangle \frac{\beta(\beta^2 r_o - \beta + 2)}{(1 - \beta r_o)^3} \tag{37}$$

To estimate the mean and variance of r, write $\mathbf{x} = S^{1/2}\mathbf{u}$, where $\mathbf{u}$ is a random variable from a normal distribution with covariance equal to the identity. Then, $r = \mathbf{u}^T H \mathbf{u}$, where $H = S^{1/2} G^{-1} S^{1/2}$. Now, if H is nearly the identity, which we expect in the $n \gg p$ limit, then r is nearly a chi-squared distribution with p degrees of freedom. In that case, $r_o = \langle\, r\,\rangle \approx p$ and $\left\langle\, (r - r_o)^2\,\right\rangle \approx 2p$. Also, for large n, we expect $\alpha \ll 1$ and $\beta \approx 1/n$. Thus, $1 - \beta r_o \approx 1$ and $\beta^2 r_o - \beta + 2 \approx 2$, and so we can simplify Eq. (37) in the large n limit:

$$\mathcal{L}_{\text{HL}} - \mathcal{L}_{\text{MM}} \approx \frac{2p}{n} \tag{38}$$

Unlike the Monte-Carlo estimator, the mean Mahalanobis estimator is biased, and in particular it is biased toward smaller values of negative log likelihood. The size of that bias, however, vanishes in the $n \to \infty$ limit.

4. NUMERICAL EXPERIMENTS

Motivated by issues in hyperspectral target detection, this section will illustrate the use of these shrinkage estimators in the context of a covariance matrix obtained from an AVIRIS (Airborne Visible/InfraRed Imaging Spectrometer)[23,24] image. This imagery has $p = 224$ spectral channels. To avoid confounding results with issues of non-Gaussian behavior, data samples are drawn from a Gaussian distribution with zero mean and covariance given by the AVIRIS data. In fact, this is the same covariance that was used in Ref. [20].

Fig. 1 and Fig. 2 plot negative log likelihood $\mathcal{L}$ as a function of α. The "true" negative log likelihood, given by Eq. (6), is plotted along with the Hoffbeck-Landgrebe LOOC estimator and the mean Mahalanobis and Monte-Carlo approximations. These plots illustrate that the approximations quite closely follow both the

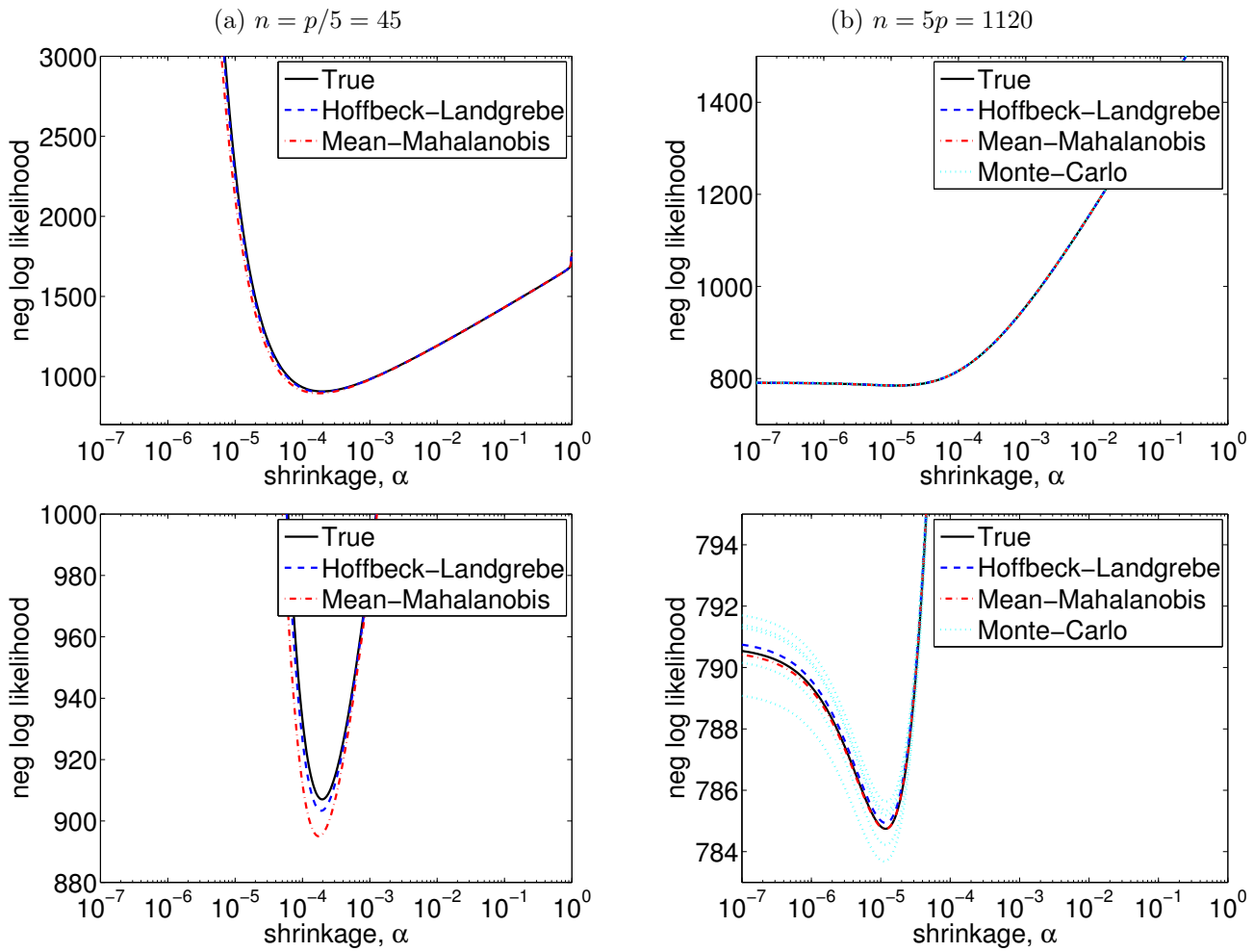

Figure 1. Negative log likelihood is plotted for the shrinkage estimated covariance matrix R_α as a function of the shrinkage parameter α. The sample covariance S is computed from n samples, and the target covariance is $T = \sigma^2 I$ for all of these plots. The bottom plots show the same data as the top plots but over a narrower range of negative log likelihood values to emphasize the minima. It is seen that both the mean Mahalanobis and the Monte-Carlo estimators do roughly as well as the Hoffbeck-Landgrebe estimator at following the curve of the true negative log likelihood, and in particular they do well at estimating a good choice of shrinkage parameter α. As n increases, the sample covariance gets better, and the need for shrinkage decreases; the optimal α thus gets smaller as n gets larger.

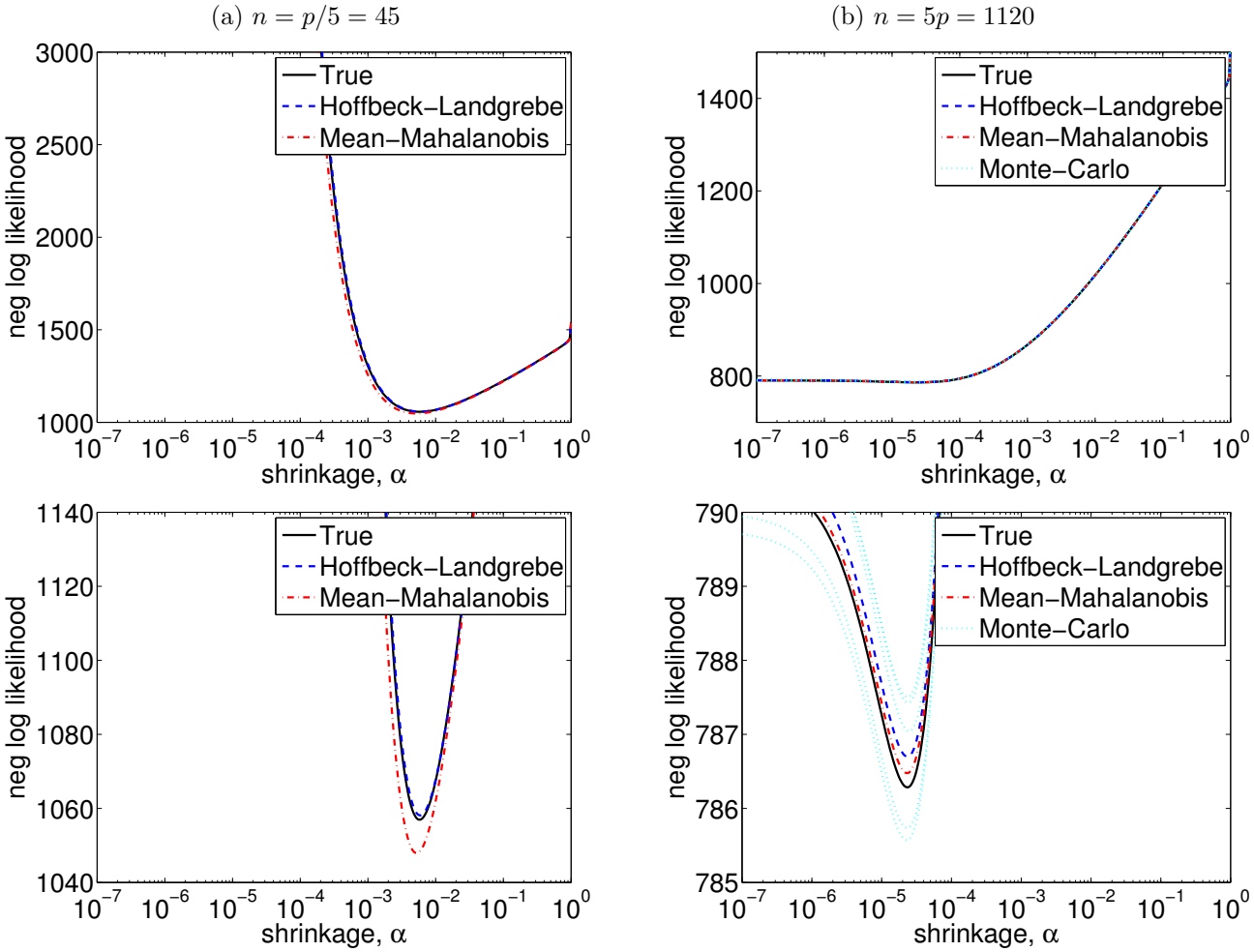

Figure 2. Same as Fig. 1 but with $T = \mathrm{diag}(S)$ as the target covariance.

Hoffbeck-Landgrebe and the true values over a broad range of α. These plots also show how much the negative log likelihood varies with α, and how valuable a good estimate of the optimal α can be.

In Fig. 3, direct estimates of this α are plotted over a set of twenty trials. Each trial corresponds to a different draw of n samples from the Gaussian distribution. Again, we see that the approximations both follow the Hoffbeck-Landgrebe estimates. For the mean Mahalanobis estimator, we do see a consistent bias, but it is much smaller than the trial-to-trial variation in the estimate of optimum α, and this behavior is observed over a wide range of n. For the Monte-Carlo estimator, it appears roughly competitive with the mean Mahalanobis estimator when n is small, but for $n \gg p$, it exhibits much greater variance than either the mean Mahalanobis bias or the trial-to-trial variability.

In Fig. 4, this behavior as a function of n is made more explicit. We see in Fig. 4(a) that the optimal α varies over several orders of magnitude with varying sample size n. The errors obtained when estimating this optimal α are shown versus n for the different approximators in Fig. 4(b).

It bears reminding that when $n = O(p)$ or smaller, then there is little or no computational advantage of these approximations compared to the full Hoffbeck-Landgrebe estimator. When $n \gg p$, then both of the approximate estimators are much cheaper to evaluate than the Hoffbeck-Landgrebe estimator. But for those larger values of n, the mean Mahalanobis estimator is much more accurate than the Monte-Carlo estimator.

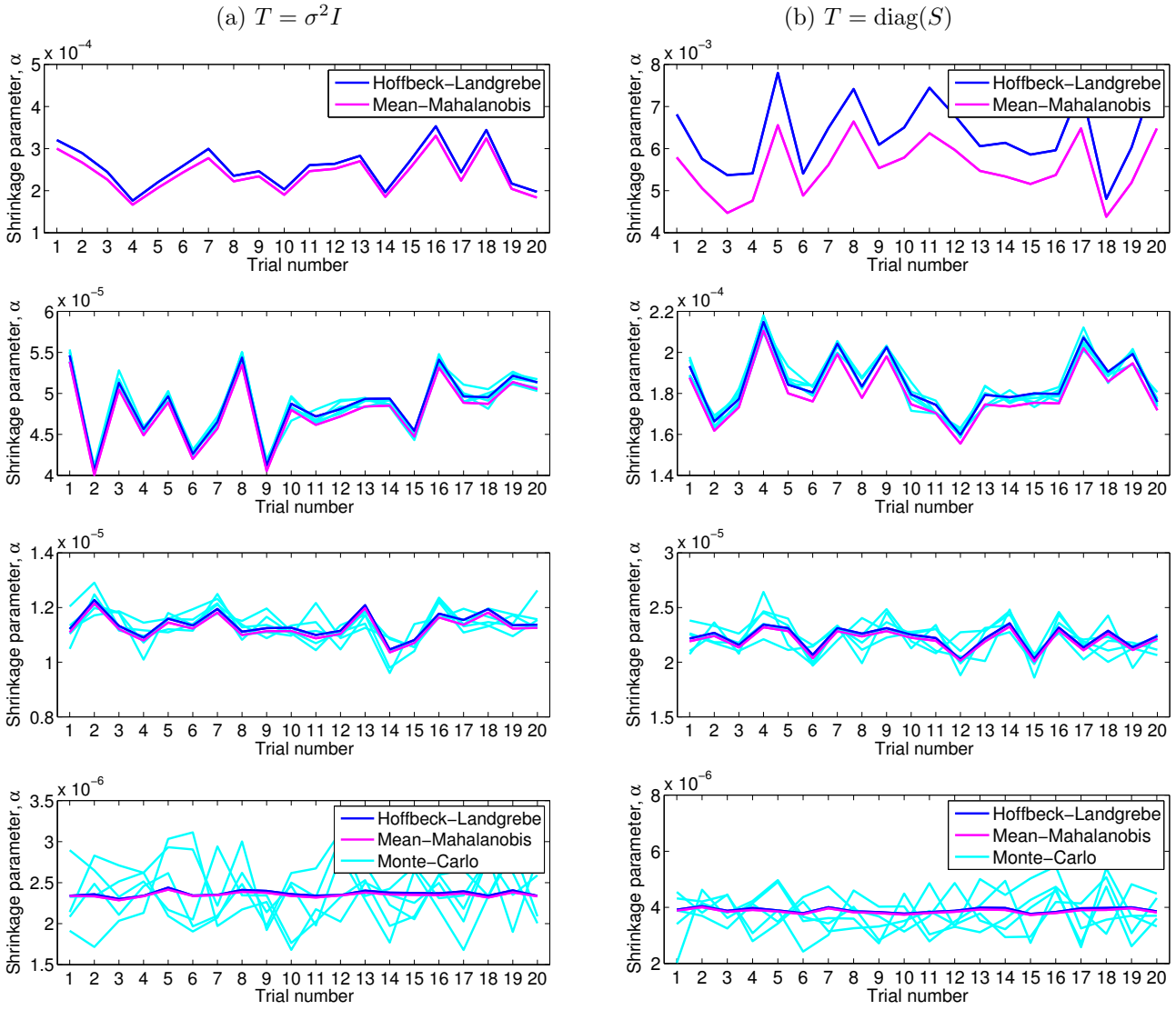

Figure 3. Twenty independent trials were performed, each corresponding to a different draw of n points (with, in order from top to bottom, $n = 45$, 224, 1120, and 5600) from the Gaussian distribution whose covariance R was derived from a hyperspectral dataset with $p = 224$ channels. In each trial, the optimal shinkage parameter α was estimated, using the different methods described in the text. For the Monte-Carlo method, five different resamplings are shown. We see that the Mean-Mahalanobis approximation closely follows the exact Hoffbeck-Landgrebe LOOC results; in fact, the approximation is smaller than the trial-to-trial variation. For $n > p/2$, we can use the Monte-Carlo approximation, which is done here with $n' = p/2$. For larger n, the Monte-Carlo variance grows much larger than the trial-to-trial variation. For the left column, the shinkage target was the ridge regularizer $T = \sigma^2 I$; for the right column, the shrinkage target was the diagonal sample covariance. In general, we observe that the shrinkage parameter α is larger for the diagonal target.

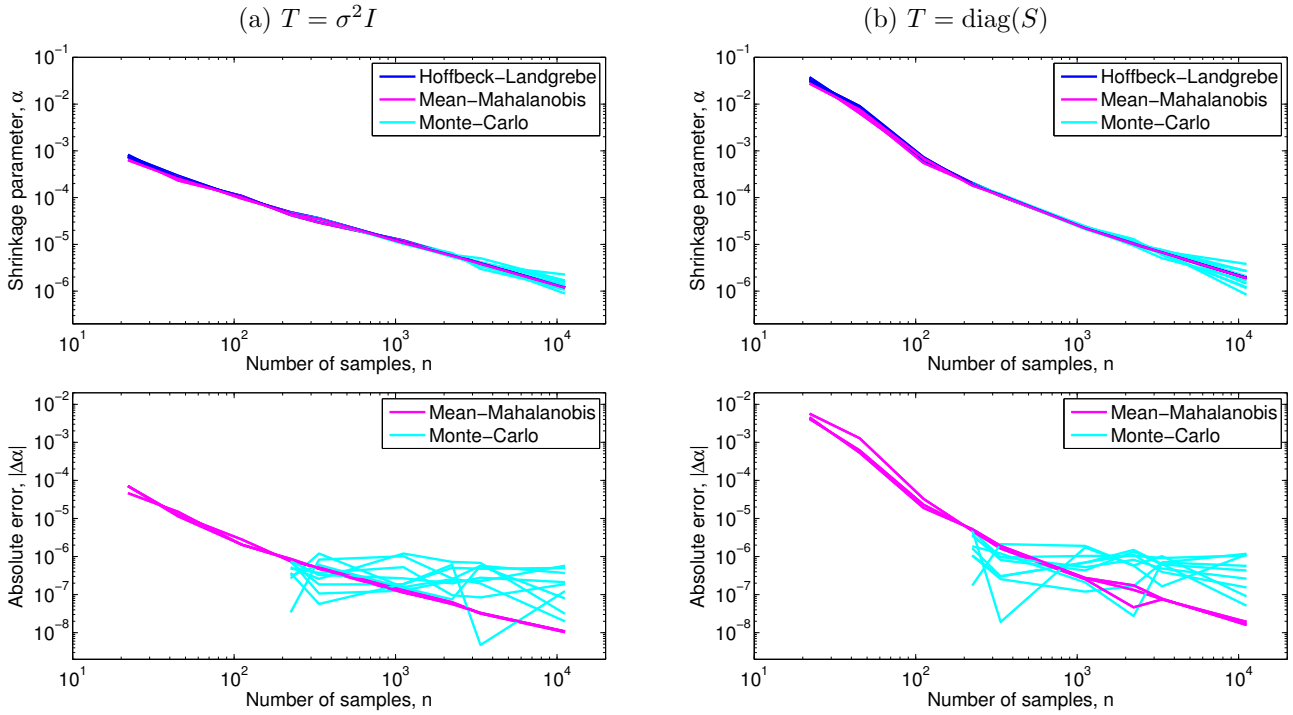

Figure 4. Top panel shows that the estimated shrinkage parameter α decreases with increasing number of samples n. Bottom panel shows absolute error of the estimator for shrinkage parameter as a function of n. Observe that the mean Mahalanobis estimator provides an increasingly better approximate as n increases, whereas the Monte-Carlo error is roughly independent of n for $n > p/2$ (for smaller n, there is not advantage to using Monte-Carlo instead of Hoffbeck-Landgrebe). For large n, which is the regime where these approximations are computationally favored over the Hoffbeck-Landgrebe estimator, the mean Mahalanobis estimator is better. For the left column, the shinkage target was the ridge regularizer $T = \sigma^2 I$; for the right column, the shrinkage target was the diagonal sample covariance. We observe that the shrinkage parameter α is larger for the diagonal target.

5. DISCUSSION

Two approximations to the leave-one-out cross-validation (LOOC) estimate of negative log likelihood are introduced. For large n, both of these estimators can achieve estimates with $O(p^3)$ effort, in contrast to the $O(np^3)$ effort that LOOC nominally appears to require, and in contrast to the $O(np^2)$ effort that the Hoffbeck-Landgrebe estimator achieves.

The Monte-Carlo approximation is unbiased, but has a variance that does not decrease as the number n of samples is increased. The mean Mahalanobis approximation *is* biased, but the bias is small (smaller than the trial-to-trial variation, according to numerical experiments), and tends to zero as $n \to \infty$. Furthermore, the mean Mahalanobis approximation is deterministic and therefore exhibits zero variance. For rapid and accurate estimation of the optimal shrinkage parameter, particularly in the $n \gg p$ regime, the mean Mahalnobis approximation is recommended.

In the experiments presented, the ridge and diagonal regularizers were considered. Future work will consider other regularizers, such as the sparse matrix transform.[18–20]

Acknowledgements

I have benefitted from valuable discussions with Leonardo Bachega, Guangzhi Cao, and Charlie Bouman. This work was supported by the Laboratory Directed Research and Development (LDRD) program at Los Alamos National Laboratory.

APPENDIX A. MEAN ESTIMATED FROM DATA

In the text, an expression for the leave-one-out sample covariance S_k was derived under the assumption of zero mean. In the more typical case, one estimates both the mean and the variance from the data; in this Appendix, we derive S_k for this more general situation.

Write for the full-sample mean and covariance:

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k; \quad \text{and} \quad S = \frac{1}{n^*}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T, \tag{39}$$

where $n^* = n$ for the maximum-likelihood estimator, and $n^* = n-1$ for the unbiased estimator. For the leave-one-out counterparts, write

$$\boldsymbol{\mu}_k = \frac{1}{n-1}\sum_{j\neq k}\mathbf{x}_j; \quad \text{and} \quad S_k = \frac{1}{n^*-1}\sum_{j\neq k}(\mathbf{x}_j - \boldsymbol{\mu}_k)(\mathbf{x}_j - \boldsymbol{\mu}_k)^T. \tag{40}$$

To simplify the manipulation, write

$$\tilde{\mathbf{x}}_k = \mathbf{x}_k - \boldsymbol{\mu} \tag{41}$$

$$\tilde{\boldsymbol{\mu}_k} = \boldsymbol{\mu}_k - \boldsymbol{\mu} = \frac{1}{n-1}\sum_{j\neq k}\tilde{\mathbf{x}}_j. \tag{42}$$

Observe that

$$\sum_{j=1}^{n}\tilde{\mathbf{x}}_j = 0; \quad \text{and} \quad \sum_{j=1}^{n}\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_j^T = n^*S. \tag{43}$$

It follows that

$$\tilde{\boldsymbol{\mu}_k} = \frac{1}{n-1}\sum_{j\neq k}\tilde{\mathbf{x}}_j = \frac{1}{n-1}\sum_{j=1}^{n}\tilde{\mathbf{x}}_j - \frac{1}{n-1}\tilde{\mathbf{x}}_k = \frac{-1}{n-1}\tilde{\mathbf{x}}_k, \tag{44}$$

and so

$$\tilde{\mathbf{x}}_k - \tilde{\boldsymbol{\mu}_k} = \frac{n}{n-1}\tilde{\mathbf{x}}_k. \tag{45}$$

Finally, we can write

$$S_k = \frac{1}{n^*-1}\sum_{j\neq k}(\tilde{\mathbf{x}}_j - \tilde{\boldsymbol{\mu}_k})(\tilde{\mathbf{x}}_j - \tilde{\boldsymbol{\mu}_k})^T \tag{46}$$

$$= \frac{1}{n^*-1}\sum_{j=1}^{n}(\tilde{\mathbf{x}}_j - \tilde{\boldsymbol{\mu}_k})(\tilde{\mathbf{x}}_j - \tilde{\boldsymbol{\mu}_k})^T - \frac{1}{n^*-1}(\tilde{\mathbf{x}}_k - \tilde{\boldsymbol{\mu}_k})(\tilde{\mathbf{x}}_k - \tilde{\boldsymbol{\mu}_k})^T \tag{47}$$

$$= \frac{1}{n^*-1}\sum_{j=1}^{n}(\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_j^T - \tilde{\boldsymbol{\mu}_k}\tilde{\mathbf{x}}_j^T - \tilde{\mathbf{x}}_j\tilde{\boldsymbol{\mu}_k}^T + \tilde{\boldsymbol{\mu}_k}\tilde{\boldsymbol{\mu}_k}^T) - \frac{1}{n^*-1}(\tilde{\mathbf{x}}_k - \tilde{\boldsymbol{\mu}_k})(\tilde{\mathbf{x}}_k - \tilde{\boldsymbol{\mu}_k})^T. \tag{48}$$

We can use the results in Eq. (43), Eq. (44), and Eq. (45) to simplify.

$$S_k = \frac{1}{n^*-1}\left[n^*S + \frac{n}{(n-1)^2}\tilde{\mathbf{x}}_k\tilde{\mathbf{x}}_k^T\right] - \frac{1}{n^*-1}\left[\frac{n^2}{(n-1)^2}\tilde{\mathbf{x}}_k\tilde{\mathbf{x}}_k^T\right] \tag{49}$$

$$= \frac{n^*}{n^*-1}S - \frac{n}{(n-1)(n^*-1)}\tilde{\mathbf{x}}_k\tilde{\mathbf{x}}_k^T \tag{50}$$

which has the same form as (and in the large n limit agrees with) Eq. (8). Also, when $n^* = n-1$, it agrees with the expression (see first equation in Section 3.3) in Hoffbeck and Landgrebe.[7]

REFERENCES

1. G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics* **21**, pp. 215–223, 1979.
2. J. H. Friedman, "Regularized discriminant analysis," *J. Am. Statistical Assoc.* **84**, pp. 165–175, 1989.
3. M. J. Daniels and R. E. Kass, "Shrinkage estimators for covariance matrices," *Biometrics* **57**, pp. 1173–1184, 2001.
4. D. I. Warton, "Penalized normal likelihood and ridge regularization of correlation and covariance matrices," *Journal of the American Statistical Association* **103**, pp. 340–349, 2008.
5. O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *J. Empirical Finance* **10**, pp. 603–621, 2003.
6. 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* **4**(32), 2005.
7. J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. Pattern Analysis and Machine Intelligence* **18**, pp. 763–767, 1996.
8. S. Tadjudin and D. A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Trans. Geoscience and Remote Sensing* **37**, pp. 2113–2118, 1999.
9. C. E. Caefer, J. Silverman, O. Orthal, D. Antonelli, Y. Sharoni, and S. R. Rotman, "Improved covariance matrices for point target detection in hyperspectral data," *Optical Engineering* **7**, p. 076402, 2008.
10. N. M. Nasrabadi, "Regularization for spectral matched filter and RX anomaly detector," *Proc. SPIE* **6966**, p. 696604, 2008.
11. H.-Y. Huang, B.-C. Kuo, J.-F. Liu, and N. Yang, "Localized shrinkage covariance estimation of hyperspectral image classification," *Proc. IGARSS* , pp. (IV)538–541, 2009.
12. C. Davidson and A. Ben-David, "Performance loss of multivariate detection algorithms due to covariance estimation," *Proc. SPIE* **7477**, p. 77470J, 2009.
13. S. Matteoli, M. Diani, and G. Corsini, "Improved estimation of local background covariance matrix for anomaly detection in hyperspectral images," *Optical Engineering* **49**, p. 046201, 2010.
14. A. Ben-David and C. E. Davidson, "Estimation of hyperspectral covariance matrices," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 4324 –4327, 2011.
15. A. Berge, A. C. Jensen, and A. H. S. Solberg, "Sparse inverse covariance estimates for hyperspectral image classification," *IEEE Trans. Geoscience and Remote Sensing* **45**, pp. 1399–1407, 2007.
16. A. C. Jensen, A. Berge, and A. H. S. Solberg, "Regression approaches to small sample inverse covariance matrix estimation for hyperspectral image classification," *IEEE Trans. Geoscience and Remote Sensing* **46**, pp. 2814–2822, 2008.
17. G. Cao and C. A. Bouman, "Covariance estimation for high dimensional data vectors using the sparse matrix transform," Tech. Rep. TR-ECE-08-05, School of Electrical and Computer Engineering, Purdue University, April 2008.
18. 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*, pp. 225–232, MIT Press, 2009.
19. 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)*, IEEE, 2009.
20. J. Theiler, G. Cao, L. R. Bachega, and C. A. Bouman, "Sparse matrix transform for hyperspectral image processing," *IEEE J. Selected Topics in Signal Processing* **5**, pp. 424–437, 2011.
21. N. A. Campbell, "Robust procedures in multivariate analysis I: Robust covariance estimation," *Applied Statistics* **29**, pp. 231–237, 1980.
22. J. Theiler and D. Hush, "Statistics for characterizing data on the periphery," *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* , pp. 4764–4767, 2010.
23. Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), Jet Propulsion Laboratory (JPL), National Aeronautics and Space Administration (NASA) `http://aviris.jpl.nasa.gov/`.
24. 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* **44**, pp. 127–143, 1993.