Double shrinkage estimators for large sparse covariance matrices

S.-M. Chang

Department of Statistics, National Cheng Kung University, Tainan, Taiwan, Republic of China

(Received 29 August 2013; accepted 3 December 2013)

Covariance matrices play an important role in many multivariate techniques and hence a good covariance estimation is crucial in this kind of analysis. In many applications a sparse covariance matrix is expected due to the nature of the data or for simple interpretation. Hard thresholding, soft thresholding, and generalized thresholding were therefore developed to this end. However, these estimators do not always yield well-conditioned covariance estimates. To have sparse and well-conditioned estimates, we propose doubly shrinkage estimators: shrinking small covariances towards zero and then shrinking covariance matrix towards a diagonal matrix. Additionally, a richness index is defined to evaluate how rich a covariance matrix is. According to our simulations, the richness index serves as a good indicator to choose relevant covariance estimator.

Keywords: adaptive lasso; generalized thresholding; sparseness; positive definiteness

1. Introduction

Estimating variance–covariance matrices is the core of many statistical methods, including principal component analysis, graphical models, clustering, etc. These statistical methods are frequently applied to microarray data, social networks, functional images, data mining, climate change and many other multivariate problems with very high dimensions. Recently, sparse covariance estimation has become popular because the sparsity assumption not only eases the interpretation but also ensures the quality of covariance matrix estimation. The sparse covariance assumption is also essential in some applications. For example, in genetics, correlation is a measure of the linkage between two genetic loci and zero correlation can be the consequence of long-run evolution without interference. Therefore, the underlying correlation matrix should be very sparse and its non-zero off-diagonal elements represent interferences. Another example preferring sparse covariance matrix is association studies, which treat variables as nodes and non-zero covariances as edges to link these nodes. Social networks and metabolism pathways are examples because it is believed that an individual participates in only a few social groups and a protein directly interacts with only several other proteins. In other words, these nodes are linked sparsely.

Other than the sample covariance (SC), there are many other interesting covariance modelling. Chiu et al. [1] applied a matrix logarithmic covariance model; Boik [2] used a spectral decomposition method and Wong et al. [3] proposed a Bayesian model. To ensure positive definiteness,
Pourahmadi [4] formulated the covariance matrix in terms of Cholesky decomposition. Another branch of the covariance estimation focuses on squeezing all sample eigenvalues together because eigenvalues of SCs always over-spread, so the smallest eigenvalue is too small and the largest eigenvalue is too large. To this end, for instance, Haff [5], Yang and Berger [6], Daniels [7] and Daniels and Kass [8] propose closed form Bayesian covariance estimators with different priors on eigenvalues. These methods retain good large-sample properties when the dimension of the matrix is fixed. Ledoit and Wolf [9] suggests a strictly positive definite estimation by shrinking the SC matrix towards the identity matrix with little computational burden. This works well when \( p > n \) and when \( p \) increases with \( n \) and we will introduce this method in Section 2. Following this line, modifications were made by Schäfer and Strimmer [10] and Fisher and Sun [11].

Many solutions have also been proposed for the sparsity. For likelihood-based methods, sparsity is the result of adding a LASSO penalty [12] to the likelihood. Similar approaches are also provided e.g. [13–15]. These approaches are called shrinkage methods. But these methods all require sophisticated algorithms to derive the optimum. Alternatively, thresholding, setting small elements to be 0, is more computationally efficient than the likelihood-based methods, and properties of this method have been investigated extensively by El Karoui [16] and Bickel and Levina [17]. Following this route, Rothman et al. [18] proposed a generalized thresholding method which combines thresholding and shrinking. The resulting covariance estimates are consistent and sparse under a setup in which both the sample size \( n \) and the dimension \( p \) go to infinity. Following Rothman et al.,[18] Rothman [19] finds positive definite correlation estimates from a parameter space with a log-determinant barrier. Instead, Xiu et al. [20] and Liu et al. [21] developed algorithms utilizing alternating direction method of multipliers, [22], to find positive definite covariance/correlation estimates directly. It is worth to notice that their algorithms require eigenvalue decompositions if the current covariance estimate is not positive definite. So it costs significantly more computation time than generalized thresholding methods.

In this study, we combine the idea of shrinking towards the diagonal,[11] and the idea of generalized thresholding,[18] shrinking small covariances towards zero, and then propose a richness index to guide the choice of shrinkage estimators. The rest of this article is organized as follows. In Section 2, the shrinking towards the diagonal method and generalized thresholding method are reviewed. Section 3 addresses the proposed double shrinkage estimators, a richness index and its estimator, and cross validation (CV) approaches in order to select tuning parameter. In Section 4, we examine the performance of the random-splitting CV scheme, the richness index estimations, and the proposed covariance estimators. Conclusions are made in Section 5.

2. Preliminary

In this section, we review methods of shrinking towards the diagonal and generalized thresholding. Before moving forward, we define notations and several frequently applied functions. The \((i,j)\)th element of a matrix \( A \) is \( a_{ij} \), \( |A|_{ij} = a_{ij} \), and conversely, the collection of all \( a_{ij} \)'s form the matrix \( A, A = [a_{ij}] \). Suppose that \( x_1, \ldots, x_n \) are \( n \) i.i.d. samples drawn from a \( p \)-dimensional multivariate normal \( N(0, \Sigma) \). The sample covariance matrix is defined as \( \Sigma = (\sum_{i=1}^{n} x_i x_i^T) / (n - 1) \), where \( \bar{x} = \sum_{i=1}^{n} x_i / n \). Denote \( \Sigma = [\sigma_{ij}] \) and \( \bar{\Sigma} = [\bar{\sigma}_{ij}] \). For a particular matrix \( B \in \mathbb{R}^{p \times p} \), the Frobenius norm is defined as \( \| B \|_F = \sqrt{tr(B^TB)} \) and the operator norm of \( B \) is defined as \( \| B \| = \sqrt{\rho_1(B^TB)} \) where \( \rho_1(\cdot) \) denotes the largest eigenvalue of its argument matrix. In the context of sparse covariance estimation, Bickel and Levina [17] and Rothman et al. [18] discussed the convergence in terms of the operator norm while Ledoit and Wolf [9] discussed the convergence in terms of the relative Frobenius norm defined as \( \| B \|_{RF} = \| B \|_F / \sqrt{p} \). In this article, we adopt the relative Forbenius norm because it results in a simple form for inferences, e.g. \( \| \bar{\Sigma} - \Sigma \|_{RF}^2 \) = \( \| \bar{\Sigma} - \Sigma \|_F^2 / \sqrt{p} \).
\[ \sum_{i,j=1}^{p} (\tilde{\sigma}_{ij} - \sigma_{ij})^2 / p. \] Hereafter, we abbreviate \( \| \cdot \|_{RF} \) by \( \| \cdot \| \). Last, define an indicator function \( I(\omega) = 1 \) if \( \omega \) is true and \( I(\omega) = 0 \) otherwise, and \((a)_+ = \max(a, 0)\).

2.1. Shrinkage towards diagonal

A covariance estimator guaranteed to have a relevant condition number is desired if the inverse of a covariance matrix is intermediate to some estimation procedures, e.g. linear regressions and Gaussian graphical models. Ledoit and Wolf [9] proposed a well-conditioned estimator for high-dimensional covariance matrices. This is a convex linear combination of the SC matrix and a scaled identity matrix, \( \tilde{\Sigma}^{LW} = \alpha \mu \mathbf{I} + (1 - \alpha) \tilde{\Sigma} \) where \( \mathbf{I} \) is the identity matrix, \( \alpha \in [0, 1] \) and \( \mu = \sum_{j=1}^{p} \sigma_{jj} / p \in \mathbb{R}^+ \). Both \( \alpha \) and \( \mu \) can be estimated consistently. Heuristically, the estimator draws the SC \( \tilde{\Sigma} \) towards the target matrix \( \mu \mathbf{I} \) so that eigenvalues are squeezed towards the average \( \mu \). Along this line, rather than using consistent estimators of \( \mu \) and \( \alpha \), Schäfer and Strimmer [10] suggested using unbiased estimators. More recently, Fisher and Sun [11] extended this class of estimators to \( \tilde{\Sigma}^{FS} = (1 - \alpha) \tilde{\mathbf{D}} + \alpha \tilde{\Sigma} \), where \( \tilde{\mathbf{D}} \) is the diagonal matrix consisting of diagonal elements of \( \tilde{\Sigma} \). We name this class of approaches as ‘shrinking towards the diagonal’ approach. Approaches mentioned above are closely related to the empirical Bayesian method [10,23] and from another point of view, these approaches sacrifice the unbiasedness property to result in an estimator with smaller mean square error.

The performance of \( \tilde{\Sigma}^{LW} \) and \( \tilde{\Sigma}^{FS} \) are comparable under some simulation scenarios [11]. However, in the view of data analysis, we notice that the two matrices involved in \( \tilde{\Sigma}^{FS} \) have the same units element-wise whereas the two matrices involved in \( \tilde{\Sigma}^{LW} \) do not because \( \mu = \sum_{j=1}^{p} \sigma_{jj} / p \). Thus, except that all variables have the same unit or are normalized, we prefer \( \tilde{\Sigma}^{FS} \) for straightforward interpretation. In addition, by applying \( \tilde{\Sigma}^{FS} \), all estimators applied in Simulation Studies yield identical variance estimates so the interactions between specified correlation structures and applied covariance estimators can be fairly revealed.

2.2. Generalized thresholding

As a result of generalized thresholding,[18] the covariance estimator is the minimizer of the penalized least squares

\[ \sum_{i,j=1}^{p} (\sigma_{ij} - \tilde{\sigma}_{ij})^2 + 2 \sum_{i,j=1}^{p} p_\lambda(\sigma_{ij}) \]

for \( \lambda \geq 0 \) and variances \((\sigma_{ij})'s\) are estimated by their corresponding sample variances. There are many choices for the penalty term. Adopting the adaptive lasso penalty [24] \( p_{\lambda,\gamma}(\sigma_{ij}) = \lambda |\sigma_{ij}|^{1-\gamma} \) results in a class of generalized thresholding function/estimator

\[ g_{\lambda,\gamma}(\tilde{\sigma}_{ij}) = \tilde{\sigma}_{ij} \left( 1 - \left( \frac{\lambda}{|\tilde{\sigma}_{ij}|} \right)^{\gamma+1} \right)_+ \quad \text{and} \quad \hat{\Sigma} = [g_{\lambda,\gamma}(\tilde{\sigma}_{ij})] \]

for \( \gamma \geq 0 \). Note that \( g_{\lambda,0}(x) = x(1 - \lambda / |x|)_+ \) is referred to as soft thresholding in the literature; \( g_{\lambda,1}(x) \) is closely related to the non-negative garrote [25]; and \( \lim_{\gamma \to \infty} g_{\lambda,\gamma}(x) = x(\mathbb{I}(|x| > \lambda)) \) is called hard thresholding. Moreover, as shown in [18], the generalized thresholding estimator has three properties: (A.1) \( |g_{\lambda,\gamma}(x)| \leq |x| \), (A.2) \( g_{\lambda,\gamma}(x) = 0 \) if \( |x| < \lambda \), and (A.3) \( |x - g_{\lambda,\gamma}(x)| \leq \lambda \) for all \( x \in \mathbb{R}, \lambda \geq 0, \) and \( \gamma \geq 0 \). We further emphasize the property that, given \( \lambda \) and \( \gamma \in (0, \infty) \)
the curve $g_{\lambda, \gamma}(x)$ is bounded by $g_{\lambda, 0}(x)$ and $g_{\lambda, \infty}(x)$, and, for any $\gamma_3 < \gamma_2 < \gamma_1 \in [0, \infty]$,

$$|x - g_{\lambda, \gamma_3}(x)| \leq |x - g_{\lambda, \gamma_2}(x)| \leq |x - g_{\lambda, \gamma_1}(x)|,$$

(1)

where the two equalities hold simultaneously if and only if $|x| < \lambda$. The properties $(A.1)$–$(A.3)$ range the function space of $g$, whereas (1) provides the magnitude of $g_{\lambda, \gamma}(x)$ deviating from $x$ with respect to $\gamma$. This property will be applied later. Note that, the class of generalized thresholding operator defined in [18] is larger than the one defined above. We focus on this smaller class to ease interpretations.

3. Main results

3.1. Double shrinkage estimators

In general, neither positive definiteness nor well-conditioned-ness is guaranteed by generalized thresholding estimators. On the other hand, the estimators in Section 2.1 do not yield sparse estimates. A straightforward method to obtain a sparse and well-conditioned covariance estimator is to combine ideas in Sections 2.1 and 2.2. Here, we suggest hybridizing shrinking towards 0 (generalized thresholding [18]), and shrinking towards diagonal.[11] Define the doubly shrinkage function/estimator as $d_{\lambda, \gamma, \alpha}(\tilde{\sigma}_{ij}) = (1 - \alpha)g_{\lambda, \gamma}(\tilde{\sigma}_{ij})$, for $i \neq j$ and $d_{\lambda, \gamma, \alpha}(\tilde{\sigma}_{ij}) = \tilde{\sigma}_{ij}$, for $i = j$ or, equivalently, in a matrix representation

$$\tilde{\Sigma} = \alpha \tilde{D} + (1 - \alpha)\tilde{\Sigma}$$

where $\tilde{\Sigma} = [d_{\lambda, \gamma, \alpha}(\tilde{\sigma}_{ij})]$ and $\alpha \in [0, 1]$. Moreover, $d_{\lambda, \gamma, 0}(x) = g_{\lambda, \gamma}(x)$. Several generalized thresholding functions and doubly shrinkage functions are shown in the left panel and the right panel of Figure 1, respectively. Note that, since $\tilde{\sigma}_{ij}$ is unbiased, the area bounded by $g_{\lambda, \gamma}(\tilde{\sigma}_{ij})$ and $f(\tilde{\sigma}_{ij})$, the identity function, quantifies the bias due to generalized thresholding. The left panel of Figure 1 shows not only the relationship (1) but also that, among generalized thresholding functions defined by adopting adaptive lasso, the resulting absolute bias is a monotone decreasing function of $\gamma$. Consequently, the bias resulting from $g_{\lambda, 0}(\tilde{\sigma}_{ij})$ is the upper bound of this class of estimators,
The procedure is as follows. Denote the SCs. To choose the rest tuning parameters, CV approaches have become standard. We first choose the tuning parameters

$$d_{\lambda, \gamma, \alpha}(\bar{\sigma}_{ij}) - \bar{\sigma}_{ij} = g_{\lambda, \gamma}(\bar{\sigma}_{ij}) - \bar{\sigma}_{ij}$$

for very $\alpha > 0$ when $|\bar{\sigma}_{ij}| > \lambda$. This implies that, in order to control the bias, $\alpha$ should be sufficiently small especially around the dominant contribution of the distribution of $\bar{\sigma}_{ij}$.

Next, we offer the following theorem to address the convergence of the proposed estimator under the assumptions: (B.1) $x_i$'s follow $p$-dimensional multivariate normal; (B.2) $\log p/n = o(1)$ and $\lambda = O(\sqrt{\log p/n})$; and (B.3) $\alpha = O(\sqrt{\log p/n})$. The first two assumptions are also made in [17,18] and the last assumption quantifies the rate of choosing $\alpha$. Although the assumption (B.1) is too restrictive to apply to general cases, it is still meaningful to provide the magnitude of $\alpha$ and $\lambda$ in terms of $n$ and $p$ for research purpose. The proof is deferred to Appendix 1.

**Theorem 3.1** If (B.1)–(B.3) are satisfied then, asymptotically,

$$E[d_{\lambda, \gamma, \alpha}(\bar{\sigma}_{ij})] - \sigma_{ij} = o(p^{-1/2})I(\sigma_{ij} \neq 0)I(i \neq j)$$

and $E\|\tilde{\Sigma} - \Sigma\|^2 = E\|\tilde{\Sigma} - \Sigma\|^2 + o(1)$.

### 3.2. Choosing tuning parameters

For the proposed estimator, there are three tuning parameters: $\lambda$, $\gamma$, and $\alpha$. In the following, we fix $\gamma$ at 0, 1, and $\infty$ because their generalized thresholding counter parts are well studied in literatures. To choose the rest tuning parameters, CV approaches have become standard. We first demonstrate $k$-fold CV. For a data matrix $X \in \mathbb{R}^{n \times p}$, rows of the data matrix are independent, whereas the covariance matrix of elements in a row is $\Sigma$. We partition $X$ into $K$ sub-matrices with roughly equal numbers of rows, $X_{[1]}, \ldots, X_{[K]}$ such that $X^T = (X_{[1]}^T, \ldots, X_{[K]}^T)$. Moreover, define $X_{[-j]}$ as stacking $X_{[i]}$'s for $k = 1, \ldots, K$ but $k \neq j$. In order to emphasize the data matrix applied to covariance estimation, we denote the SC estimated using data matrix $X$ by $\bar{\Sigma}(X) = [\bar{\sigma}_{ij}(X)]$. The loss for generalized thresholding estimator is defined as

$$L(\lambda, \alpha) = E\{\sum_{i \neq j} [d_{\lambda, \gamma, \alpha}(\bar{\sigma}_{ij}(X)) - \sigma_{ij}]^2\}.$$

The $(\lambda, \alpha)$ that minimize $L(\lambda, \alpha)$ is called optimal. However, $\Sigma$ is unknown and thus we can not estimate $L(\lambda, \alpha)$ directly. Instead,

$$\frac{1}{K} \sum_{k=1}^{K} \sum_{i \neq j} [d_{\lambda, \gamma, \alpha}(\bar{\sigma}_{ij}(X_{[-k]})) - \bar{\sigma}_{ij}(X_{[k]})]^2$$

is applied to estimate the loss. Note that, in $k$-fold CV, we estimate $\sigma_{ij}$ twice by $d_{\lambda, \gamma, \alpha}(\bar{\sigma}_{ij}(X_{[-k]}))$ and $\bar{\sigma}_{ij}(X_{[k]})$. According to our simulation (not shown in this work), five-fold CV severely over-estimates the optimal $\lambda$. This may because only one-fifth of the data are applied to estimate $\bar{\sigma}_{ij}$, so the estimation precision is relatively low and hence the best $\lambda$ chosen by CV is not close enough to the optimal. Other than $k$-fold CV, Rothman et al. [18] suggested reserving some data for the testing data set and using the rest as the training data set. The training data set is applied to $d_{\lambda, \gamma, \alpha}(\sigma_{ij})$ and the testing data set is applied to calculate the SCs. This approach performs well if the sample size is large enough. On the other hand, Bickel and Levina [17] suggested the random splitting CV. The procedure is as follows. Denote $X^b$ as the matrix whose rows are the result of the $b$th random permutation of the rows of $X$. Randomly split the data matrix $X^b$ into two submatrices $X_{[1]}^b$ and
with roughly equal numbers of rows, and then the random splitting CV estimator of \( L(\lambda, \alpha) \) is

\[
CV(\lambda, \alpha; \mathbf{X}) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i \neq j} \left[ d_{\lambda, \gamma, \alpha}(\hat{\sigma}_{ij}(\mathbf{X}^b_{[1]})) - \tilde{\sigma}_{ij}(\mathbf{X}^b_{[2]}) \right]^2
\]

\[
= \frac{1}{B} \sum_{b=1}^{B} \sum_{i \neq j} \left[ (1 - \alpha) \times g_{\lambda, \gamma}(\hat{\sigma}_{ij}(\mathbf{X}^b_{[1]})) - \tilde{\sigma}_{ij}(\mathbf{X}^b_{[2]}) \right]^2.
\]

In Theorem 3.2, we show that \( \sum_{i \neq j} \left[ d_{\lambda, \gamma, \alpha}(\hat{\sigma}_{ij}(\mathbf{X}^b_{[1]})) - \tilde{\sigma}_{ij}(\mathbf{X}^b_{[2]}) \right]^2 \) is unbiased to the loss \( L(\lambda, \alpha) \) up to a constant, for all \( b \). The key point is that \( d_{\lambda, \gamma, \alpha}(\hat{\sigma}_{ij}(\mathbf{X}^b_{[1]})) \) and \( \tilde{\sigma}_{ij}(\mathbf{X}^b_{[2]}) \) are independent. The proof is deferred to Appendix 2.

**Theorem 3.2** \( E[CV(\lambda, \alpha; \mathbf{X})] = L(\lambda, \alpha) + O(E||\hat{\Sigma} - \Sigma||^2) \).

In addition, since CV is a quadratic function in \( \alpha \) the unique minimizer of CV is

\[
\hat{\alpha}_{\lambda, \gamma} = 1 - \frac{\sum_{b=1}^{B} \sum_{i \neq j} \hat{\sigma}_{ij}(\mathbf{X}^b_{[2]}) g_{\lambda, \gamma}(\hat{\sigma}_{ij}(\mathbf{X}^b_{[1]}))}{\sum_{b=1}^{B} \sum_{i \neq j} \left[ g_{\lambda, \gamma}(\hat{\sigma}_{ij}(\mathbf{X}^b_{[1]})) \right]^2}.
\]

Moreover, we have to enforce the \( \alpha \) estimate falling in \([0, 1]\) and thus define the estimator \( \tilde{\alpha}_{\lambda, \gamma} = \min\{1, \max\{0, \hat{\alpha}_{\lambda, \gamma}\}\} \) for \( \alpha \). Then \( \lambda \) is chosen by grid search so that \( CV(\lambda, \tilde{\alpha}_{\lambda, \gamma}, \alpha; \mathbf{X}) \) is minimized. In practice, \( \lambda \) should be no greater than the largest covariances \( |\tilde{\sigma}_{ij}| \). However, if the data is normalized (estimating correlation rather than estimating covariance) then it is proper to search \( \lambda \) in the interval \((0, 1)\). Hereafter, we shorthand \( L(\lambda, \alpha) \) by \( L(\lambda) \) because the best \( \alpha \) is also a function of \( \lambda \).

### 3.3. Richness index

As emphasized, the second purpose of this work is choosing relevant estimators among those introduced in Section 2 and proposed in Section 3. We conjecture that the ‘richness’ of a covariance matrix can be an index in choosing estimators because the possible improvement of aforementioned estimators is due to the sparsity assumption.\[10\] Without loss of generality, we consider \( \Sigma \) as a correlation matrix, i.e. \( \sigma_{11} = \cdots = \sigma_{pp} = 1 \). For a particular correlation matrix \( \Sigma \), let \( p\|\Sigma_D\|^2 = \sum_{i \neq j} \sigma_{ij}^2 \), the sum of correlation squares, and define the richness index as

\[
RI = \|\Sigma_D\|^2 / \|\Sigma\|^2 = \|\Sigma_D\|^2 / (1 + \|\Sigma_D\|^2).
\]

So, for a sparse covariance model, \( \|\Sigma_D\|^2 \) is relatively small and so is \( RI \), whereas for a rich covariance model, \( \|\Sigma_D\|^2 \) is relatively large and so is \( RI \). For instance, consider correlation models defined at the beginning of Section 4. As \( p \to \infty \), the ZR model results in \( \|\Sigma_D\|^2 = 2r^2/p \to 0 \); the MA model results in \( \|\Sigma_D\|^2 = 2r^2(p - 1)/p \to 2r^2 \); and the the compound symmetry (CS) model results in \( \|\Sigma_D\|^2 = r^2(p - 1) \to \infty \). In this sequel, if \( p \to \infty \) and \( p/n \to c < \infty \), then, for the ZR model, \( RI \to 0 \); for the MA model, \( RI \to 2r^2/(2r^2 + 1) \); and, for the CS model, \( RI \to 1 \). It appears that \( RI \) can be a classifier to distinguish the aforementioned covariance models into three classes if the parameter \( r \) is not extreme, e.g. \( r \approx 0 \) or \( |r| \approx 1 \).
In order to estimate RI, we utilize the fact that, under normality,

\[ E \| \tilde{\Sigma} \|^2 = \frac{1}{(n - 1)p} \sum_{i,j=1}^{p} \sigma_{ij}^2 \sigma_{ji}^2 + \frac{n}{(n - 1)p} \sum_{i=1}^{p} \sigma_{ii}^2 + \frac{n}{(n - 1)p} \sum_{i \neq j} \sigma_{ij}^2 \approx \frac{p}{n} + 1 + \| \Sigma_D \|^2. \]

Consequently, an estimator of the richness index RI can be

\[ \hat{RI} = \frac{\| \tilde{\Sigma} \|^2 - 1 - p/n}{\| \tilde{\Sigma} \|^2 - p/n}. \]

The performance of \( \hat{RI} \) is shown in Section 4. Note that in practice we substitute the \( \tilde{\Sigma} \) in \( \hat{RI} \) by sample correlation matrix.

### 4. Simulation studies

The simulation plan was as follows. For a given covariance matrix \( A \), we factorized it by LU decomposition, say \( A = LL^T \). Next, a \( p \)-vector independent sample, \( z \), were sampled from the standard normal distribution or from the exponential distribution with mean 1. Then each data set consists of \( n \) independent transformed vectors, \( x = Lz \). The sample size was fixed at \( n = 50 \) with dimension \( p = 50f \), where \( f = 0.6, 1, 2, 3, 4 \). Five covariance models were applied with parameter \( r = 0.5 \). They are

1. **Model ZR**: \( \Sigma_{ij} = r \times I(i + j = 3) + I(i = j) \)
2. **Model MA**: \( \Sigma_{ij} = r \times I(|i - j| = 1) + I(i = j) \)
3. **Model AR**: \( \Sigma_{ij} = \rho^{|i-j|} \)
4. **Model BL**: \( \Sigma_{ij} = (1 - |i - j|/d)_+ \) and \( d \) is the largest integer which is smaller than \( p \times r \).
5. **Model CS**: \( \Sigma_{ij} = r \times I(i \neq j) + I(i = j) \)

In addition to the ZR and CS models, other models were also investigated by Bickel and Levina [17] and Rothman et al. [18]. Model ZR has two non-zero correlations, (1,2) and (2,1). The MA covariance matrix is very sparse as only 2 elements are non-zero over \( p^2 - p \) elements. Model AR is only approximately sparse,[18] entries farther away from the diagonal tend to be closer to 0 at an exponential rate but not exactly equal to 0. Model BL is proposed by Bickel and Levina [17]. Correlations decrease linearly from the diagonal to those far from the diagonal. Finally, the CS model has no zero elements unless \( r = 0 \).

In the first simulation study, we explored the performance of random-splitting CVs in terms of choosing \( \lambda \). The sample size was 50 with dimension 100, and the number of random splitting was \( B = 10 \). To estimate \( L(\lambda) = E \| \tilde{\Sigma} - \Sigma \|^2 \), 200 independent data sets were generated and the expectation was approximated by Monte Carlo integration. That is, \( L(\lambda) \) is approximated by

\[ \text{MCL}(\lambda) = \frac{1}{B_1} \sum_{j=1}^{B_1} \| \tilde{\Sigma}(X_j) - \Sigma \|^2, \]

where \( B_1 = 200 \) and \( X_j \)'s are independent data matrices generated from the same data generating process. Moreover, define \( \lambda^0 = \arg\min_{\lambda \geq 0} \text{MCL}(\lambda) \) and \( \hat{\lambda} = \arg\min_{\lambda \geq 0} \text{CV}(\lambda) \). Three models, AR, BL, and CS, and three estimators, hard thresholding (GT\( \infty \)), soft thresholding (GT\( 0 \)), and a double shrinkage estimator DS\( \infty \) were applied. Other estimators performed similarly to them, and so were omitted. The results for normal data and for exponential data are similar so we only show the former, in Figure 2. The MCL(\( \lambda \)) is in black and additional 20 CV(\( \lambda \)) curves in gray are
Figure 2. Loss estimations with sample size 50 and dimension 100. From top to bottom, rows represent covariance models: AR, BL, and CS; whereas, from left to right, columns represent estimators: hard thresholding, soft thresholding, and DS$_{\infty}$. For each box, the x-axis is the value of the tuning parameter $\lambda$ and the y-axis is the estimates of loss. In each box, the black line is the MCL curve, whereas 20 grey curved are CV’s. The ‘O’ locates $\lambda^0$ and the ‘X’ locates $\hat{\lambda}$.

It appears that $\hat{\lambda}$’s are much closer to $\lambda^0$’s under the BL and CS covariance models than under the AR model. However, as shown in the next simulation study, the discrepancy does not result in abrupt covariance estimates in general.

Next, we examined the performance of the richness index estimator. Values of $\hat{RI}$ under above covariance models with various dimensions $p$ are summarized in Figure 3. By Figure 3, for both Gaussian data (left panel) and exponential data (right panel), the richness index estimate successfully classifies covariance models into three groups: ZR ($RI \to 0$) has very low richness index estimates; MA ($RI \to 1/3$) and AR have mild richness index estimates; and BL and CS ($RI \to 1$) have richness index estimates approaching 1. In the next simulation, we found that this grouping coincides with choices of covariance estimators under these covariance models.

In the third simulation, we focus on comparing several covariance estimators: the SC, the FS estimator,[11] generalized thresholding estimators (GT$_{\gamma}$, $\gamma = 0, 1, \infty$), the proposed double shrinkage methods (DS$_{\gamma}$, $\gamma = 0, 1, \infty$), and the positive definite estimator proposed by Xiu et al. [20], named XMZ estimator. Three measures were applied to illustrate the performance of these estimators: the sample loss (SL) $\sum_{i \neq j} [d_{\lambda, \gamma, \alpha}(\hat{\sigma}_{ij}) - \sigma_{ij}]^2$, the true positive rate (TPR)

$$TPR = \frac{\# \text{ true non-zero entries claimed to be non-zero}}{\# \text{ true non-zero entries}}$$
and the false positive rate (FPR)

\[ \text{FPR} = \frac{\# \text{ true zero entries claimed to be non-zero}}{\# \text{ true zero entries}}. \]

Moreover, we monitored the condition number (CN, the absolute value of the largest eigenvalue divided by the smallest eigenvalue) and the positive definite rate (PDR, the rate of being positive definite). We reported these measures by averaging 100 replicates, and simulation results are summarized in Tables D1–D10 in Appendix D. Note that Tables D1–D10 and Figures C1–C3 are shown in Supplemental Data.

First, we are interested in the effect of shrinking towards diagonal. So we show the relative loss, the loss of DS_\gamma divided by the loss of GT_\gamma, in Figure 4 for Gaussian data and in Figure C1 for exponential data. They show that, in most cases, DS_\gamma outperforms GT_\gamma except under BL and CS models with Gaussian data. Second, we are interested in choosing \( \gamma \) (0, 1, or \( \infty \)). For both Gaussian data (Figure 5) and exponential data (Figure C2), the losses due to DS_0 and DS_1 are similar than the losses due to DS_\infty. Under MA and AR model, this tendency is strong when the sample size is small and is vanished when the sample size is large. So, under ZR, MA and AR model (RI < \( \infty \)), using DS_0 may result in estimates with reasonably small loss in contrast to DS_\gamma's and GT_\gamma's. Last, we contrast the performance of estimators FS, DS_0, and XMZ. From Figures 6 and C3, we have the following conclusion. In terms of having smallest loss, under ZR model, three methods are roughly equivalent; under MA and AR model, DS_0 is slightly better.
than XMZ and is much better than FS, especially when the dimension is larger than the sample size; under BL and CS model, FS is the best.

In general, DS$_0$ has higher PDR than DS$_γ$, $γ = 1$ and $∞$, under all situation. For a fixed $γ$, most of PDR’s of DS$_γ$’s are higher than or equal to PDR’s of GT$_γ$’s. Note that the PDR of DS$_0$ can be as low as 0.71 when the true covariance model is BL or CS (models with $IR \rightarrow 1$). As suggested, DS$_γ$’s are preferred when the covariance model has $0 \leq RI < 1$ and under ZR, MA and AR model, all observed PDR’s are 1. We also consider the median condition number (MCN) over 100 replicates. We observed that in most cases, the MCN of DS$_0$ is smaller than the MCN’s of XMZ and FS. Last, we consider the TPRs and the FPRs of those estimators. These rates vary among covariance models. Under most situations, TPR and FPR are increased when $γ$ becomes smaller. We observe that DS$_0$ has slightly higher TPR than XMZ does, up to 10%, and DS$_0$ has much higher TPR than XMZ does, up to 37%. These extreme cases happen when the BL model is applied. Finally, we notice that under ZR model (having only two non-zero entries), the TPRs are extremely low, ranging from 0.00 to 0.30. This phenomenon reveals that DS$_γ$’s, GT$_γ$’s and XMZ result in too many zero’s. Therefore, we do not suspect the observation that both TPR and FPR are 0 which merely says that the corresponding covariance estimate is a diagonal matrix.
Figure 5. Estimated loss of DS, under different covariance models, different dimensions and different γ’s where data sets were generated from normal distribution. Each losses was estimated by 100 independent replicates.

5. Concluding remarks

In this work, we have proposed a class of double shrinkage estimators for large sparse covariance matrices. The SC is first shrunk element-wise towards zero and then shrunk towards a diagonal matrix. The performance of the doubly shrinkage estimator was examined via simulation under various covariance models. The results showed that for sparse (ZR and MA) or approximately sparse (AR) covariance models, double shrinkage estimators outperform the corresponding generalized thresholding estimators in terms of having smaller losses. On the contrary, for rich (BL and CS) covariance models, the generalized thresholding estimators with Gaussian data are better. In addition, we observed that the performance of estimators depends on how rich the target covariance matrix is but not merely how sparse the target covariance matrix is. This motivated us to define an index to measure the richness of a matrix. The proposed richness index estimator defined in Section 3.2 and examined in Section 4 appears to be useful in choosing better estimators. Combining the information resulting from our simulation, we conjecture that, for a given data set, if the richness index estimate is greater than 0.8 then the XMZ estimators (for sparse estimation) or the FS estimator (for non-sparse estimation) may perform well and, otherwise, the proposed doubly shrinkage estimator with γ = 0 may perform better.

We advocate using RI as an indicator of choosing covariance estimator. The suggestion rule is summarized in Table 1. Note that since AR is approximately sparse and BL is less sparse then
Figure 6. Estimated loss under different estimators, FS, DS₀, and XMZ, different covariance models, and different dimensions where data sets were generated from normal distribution. Each losses was estimated by 100 independent replicates.

Table 1. Richness, sparseness and recommended estimator of the covariance models.

|       | ZR   | MA   | AR   | BL   | CS   |
|-------|------|------|------|------|------|
| Richness RI → 0 0 < RI < 1 0 < RI < 1 RI → 1 RI → 1 |
| Sparseness sparse sparse non-sparse sparse non-sparse |
| Recom. Est. DS₀ DS₀ DS₀ XMZ/FS XMZ/FS |

ZR and MA, the sparseness is also a good indicator for estimation selection. However, we can not estimate the sparseness of a sample covariance before choosing a relevant estimator but we can estimate the proposed richness index prior to choosing an estimator. Note that the proposed richness index is a function of the relative Frobenius norm of a sample correlation. Consequently, a single huge covariance hardly affect the RI estimate when the dimension is large.

In some association studies, covariance matrix is the final product where zero covariances indicate no association directly and hence the positive definiteness is not desired. On the other hand, in some discrimination analyses, covariance estimate is part of the Gaussian density function and therefore the covariance matrix has to be positive definite. When the positive definiteness is required, a mild modification can be imposed on the proposed shrinkage estimators. Because the
eigenvalues of $\tilde{\Sigma}$ is a monotone increasing function of $\alpha$, we can simply tune $\alpha$ upward until the estimator becomes positive definite although, by doing so, the loss increases in general. Last, DS$_0$ is suggested when RI $< 1$ and fortunately, our simulations show that the PDR of DS$_0$ is 1 when the underlying covariance models have RI $< 1$ and hence above modification is barely required.

Supplemental data

Supplemental data for this article can be accessed from doi:10.1080/00949655.2013.873042

Acknowledgements

The author thanks the editor and anonymous referees for their helpful comments and suggests. This work is supported in part by the grant from the National Science Council of Taiwan with grant No. NSC99-2511-S-006-005

References

[1] Chiu YM, Leonard T, Tsui K-W. The matrix-logrithmic covariance model. J Amer Statist Assoc. 1996;91:198–209.
[2] Boik RJ. Spectral models for covariance matrices. Biometrika. 2002;89:159–182.
[3] Wong F, Carter CK, Kohn R. Efficient estimation of covariance selection models. Biometrika. 2003;90:809–830.
[4] Pourahmadi, M. Joint mean-covariance models with application to longitudinal data: unconstrained parameterization. Biometrika. 1999;86:677–690.
[5] Haff LR. The variational form of certain Bayes estimators. Ann Statist. 1991;19:1163–1190.
[6] Yang R, Berger JO. Estimation of a covariance matrix using the reference prior. Ann Statist. 1994;22:1195–1211.
[7] Daniels MJ. A prior for the variance in hierarchical models. Canad J Statist. 1999;27:569–580.
[8] Daniels MJ, Kass RE. Shrinkage estimators for covariance matrices. Biometrics. 2001;57:1173–1184.
[9] Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. J Multivar Anal 2004;88:365–441.
[10] Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol. 2005;4:1175–1189.
[11] Fisher TJ, Sun X. Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. Comput Statist Data Anal. 2011;55:1909–1918.
[12] Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc B. 1996;58:267–288.
[13] Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika. 2007;94:19–35.
[14] d’Apremont A, Banerjee O, El Chaoui L. First-order methods for sparse covariance selection. SIAM J Matrix Anal Appl. 2008;30:56–66.
[15] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9:432–441.
[16] El Karoui N. Operator norm consistent estimation of large dimensional sparse covariance matrices. Ann Statist. 2008;36:2717–2756.
[17] Bickel PJ, Levina E. Covariance regularization by thresholding. Ann Statist. 2008;36:2577–2604.
[18] Rothman AJ, Levina E, Zhu J. Generalized thresholding of large covariance matrices. J Amer Statist Assoc. 2009;104:177–186.
[19] Rothman AJ. Positive definite estimators of large covariance matrices. Biometrika. 2012;99:733–740.
[20] Xiu L, Ma S, Zou H. Positive-definite $l_1$-penalized estimation of large covariance matrices. J Amer Statist Assoc. 2012;107:1480–1491.
[21] Liu H, Wang L, Zhao T. Sparse covariance matrix estimation with eigenvalue constraints. J Comput Graph Statist. Available from: http://www.tandfonline.com/doi/abs/10.1080/10618600.2013.782818#.UrVwX2elqVg
[22] Bien J, Tibshirani R. Sparse estimation of a covariance matrix. Biometrica. 2011;98:807–820.
[23] Efron B. Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. IMS monographs. Cambridge: Cambridge University Press; 2010.
[24] Zou H. The adaptive lasso and its oracle properties. J Amer Statist Assoc. 2006;101:1418–1429.
[25] Breiman L. Better subset regression using the nonnegative garrote. Technometrics. 1995;37:373–384.
[26] Imhof JP. Computing the distribution of quadratic forms in normal variables. Biometrika. 1961;48:417–426.

Appendix 1. Proof of Theorem 3.1

Proof For all $i \neq j$, according to Imhof [26], the SC $\tilde{\sigma}_{ij}$ follows the distribution of a linear combination of two non-central chi-squares. Moreover, for $i, j = 1, \ldots, p$, $\tilde{\sigma}_{ij}$ has mean $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$ and variance $(1 + \rho_{ij}^2)\sigma_i^2\sigma_j^2/(n-1)$. Note that, here,
we use $\sigma_i^2 = \sigma^2$ by convention. This follows that, as $n$ is sufficiently large,

$$\sqrt{\frac{n-1}{(1+\gamma^2)\sigma_i^2}}(\hat{\sigma}_{ij} - \sigma_{ij}) \sim N(0,1)$$

by central limit theorem. Let $W \sim N(\mu, \sigma^2)$ and denote its p.d.f. and c.d.f. as $\phi(w; \mu, \sigma^2)$ and $\Phi(w; \mu, \sigma^2)$, respectively. The SC $\hat{\sigma}_{ij}$ is unbiased whereas the generalized thresholding estimator $g_{2\gamma}(\hat{\sigma}_{ij})$ is biased, for all $\gamma \in [0, \infty]$. Moreover, because of (1) and $E(g_{2\gamma}(\hat{\sigma}_{ij}) - \sigma_{ij}) = E(g_{2\gamma}(\tilde{\sigma}_{ij}) - \tilde{\sigma}_{ij})$, we conclude that the absolute bias of generalized thresholding estimators has a lower bound, the absolute bias due to the hard thresholding $\gamma = \infty$, and an upper bound, the absolute bias due to the soft thresholding $\gamma = 0$. Therefore, below we firstly investigate the biases due to the hard thresholding and the soft thresholding and then extend the result to the proposed double shrinkage estimators.

For convenience, denote $\hat{\sigma}_{ij}^H = g_{2\gamma}(\tilde{\sigma}_{ij})$ and $\hat{\sigma}_{ij}^S = g_{2\gamma}(\tilde{\sigma}_{ij})$. Let $\mathbb{S} = [-\lambda, \lambda]$ and $\hat{\sigma}_{ij} = W \sim N(\sigma_{ij}, \tau_{ij}/n)$ where $\tau_{ij} = (1 + \rho_{ij}^2)\sigma_i^2\sigma_j^2/n(n-1)$. Then

$$E(\hat{\sigma}_{ij}^H) = \int_{\mathbb{S}} wI(|w| > \lambda)\phi(w; \sigma_{ij}, \tau_{ij}/n) \, dw = \sigma_{ij} - \int_{\mathbb{S}} w\phi(w; \sigma_{ij}, \tau_{ij}/n) \, dw.$$

The bias, $b_{ij}^H = E(\hat{\sigma}_{ij}^H) - \sigma_{ij}$, is therefore

$$b_{ij}^H = -\int_{\mathbb{S}} w\phi(w; \sigma_{ij}, \tau_{ij}/n) \, dw$$

$$= (\tau_{ij}/n)\int_{\mathbb{S}} w - \sigma_{ij}/\tau_{ij}/n \phi(w; \sigma_{ij}, \tau_{ij}/n) \, dw - \sigma_{ij}\int_{\mathbb{S}} \phi(w; \sigma_{ij}, \tau_{ij}/n) \, dw$$

$$= \tau_{ij}/n[\phi(-\lambda; \sigma_{ij}, \tau_{ij}/n) - \phi(-\lambda; \sigma_{ij}, \tau_{ij}/n) - \sigma_{ij}[\Phi(\lambda; \sigma_{ij}, \tau_{ij}/n) - \Phi(-\lambda; \sigma_{ij}, \tau_{ij}/n)]$$

$$= \tau_{ij}/n[\phi(-\lambda; \sigma_{ij}, \tau_{ij}/n) - \phi(-\lambda; \sigma_{ij}, \tau_{ij}/n) - \sigma_{ij}[2\lambda\phi(-\lambda; \sigma_{ij}, \tau_{ij}/n)],$$

where $\varepsilon \in \mathbb{S}$ by mean value theorem. Note that if $\sigma_{ij} = 0$ then $b_{ij}^H = 0$ and if $\sigma_{ij} \neq 0$ then, by assumption (B.2),

$$pb_{ij}^H \propto \exp[\log \lambda + 0.5 \log n + \log p - C_1n] \to 0,$$

where $C_1 = (\sigma_{ij} + \varepsilon)^2 / \tau_{ij} > 0$. Thus, $E(\hat{\sigma}_{ij}^H) = \sigma_{ij} + o(1/p)$. Similarly,

$$E(\hat{\sigma}_{ij}^S) = \int_{\mathbb{S}} w\left(1 - \frac{\lambda}{|w|}\right) + \phi(w; \sigma_{ij}, \tau_{ij}/n) \, dw$$

$$= \int_{\mathbb{S}} w\phi(w; \sigma_{ij}, \tau_{ij}/n) \, dw - \lambda \int_{\mathbb{S}} \phi(w; \sigma_{ij}, \tau_{ij}/n) \, dw + \lambda \int_{\mathbb{S}} \phi(w; \sigma_{ij}, \tau_{ij}/n) \, dw$$

$$= E(\hat{\sigma}_{ij}^H) - \lambda[1 - \Phi(\lambda; \sigma_{ij}, \tau_{ij}/n) - \Phi(-\lambda; \sigma_{ij}, \tau_{ij}/n)] \equiv E(\hat{\sigma}_{ij}^H) - \lambda c_{ij}^S.$$}

Note that, if $\sigma_{ij} = 0$ then $c_{ij}^S = 0$ and if $\sigma_{ij} \neq 0$ then

$$c_{ij}^S = \Phi(-\lambda; \sigma_{ij}, \tau_{ij}/n) - \Phi(-\lambda; \sigma_{ij}, \tau_{ij}/n) - \Phi(-\lambda; \sigma_{ij}, \tau_{ij}/n) + \phi(\varepsilon_2; \sigma_{ij}, \tau_{ij}/n)(\sigma_{ij} + \lambda),$$

where $\varepsilon_1$ is a constant between $\sigma_{ij}$ and $\lambda$ and $\varepsilon_2$ is a constant between $\sigma_{ij}$ and $-\lambda$. Therefore, the bias due to the soft thresholding is $b_{ij}^S = E(\hat{\sigma}_{ij}^S) - \sigma_{ij} = o(1/p) - \lambda c_{ij}^S$. Additionally,

$$pb_{ij}^S \propto \exp[\log \lambda + 0.5 \log n + \log p - C_2n] \to 0,$$

where $C_2 = \min\{(\sigma_{ij} + \varepsilon_1)^2 / \tau_{ij}, (\sigma_{ij} + \varepsilon_2)^2 / \tau_{ij}\}$. So $b_{ij}^S = o(1/p)$. Consequently, when $\sigma_{ij} = 0, E(g_{2\gamma}(\tilde{\sigma}_{ij})) = \sigma_{ij}$, and when $\sigma_{ij} \neq 0, E(g_{2\gamma}(\tilde{\sigma}_{ij})) = \sigma_{ij} + o(1/p)$, for $\gamma \in [0, \infty]$.

For the double shrinkage estimator, we have that, for $i \neq j$,

$$b_{ij}^D = E(d_{\alpha, \gamma, \alpha}(\tilde{\sigma}_{ij})) - \sigma_{ij} = -\alpha \sigma_{ij} + (1 - \alpha) E(d_{\alpha, \gamma, \alpha}(\tilde{\sigma}_{ij})) - \sigma_{ij}$$

and thus $b_{ij}^D = -\alpha \sigma_{ij} + (1 - \alpha) o(1/p) = o(1/\sqrt{p})$ because of assumptions (B.2) and (B.3). Consequently, $(b_{ij}^D)^2 = o(1/p)$ and hence $E\|\hat{\Sigma} - \Sigma\|^2 = E\|\hat{\Sigma} - \Sigma\|^2 + o(1)$. So the proof is complete. \qed
Appendix 2. Proof of Theorem 3.2

Proof Suppose there are two independent data matrix $X \in \mathbb{R}^{n \times p}$ and $X_a \in \mathbb{R}^{n_a \times p}$. Rows of both of them are independently generated from identical data generating process, where the covariance matrix of elements in each row is $\Sigma_1$. Denote the SC estimator using $X_a$ and $X$ as $\bar{\Sigma}(X_a) = [\bar{\sigma}_{ij}^a]$ and $\bar{\Sigma}(X) = [\bar{\sigma}_{ij}]$, respectively, and denote another covariance estimator using $X$ as $\hat{\Sigma}_D(X) = [\hat{\sigma}_{ij}^D]$ where $D$ denotes a double shrinkage estimator given tuning parameters $\lambda, \gamma,$ and $\alpha$. Now, $E(\bar{\sigma}_{ij}^a) = E(\bar{\sigma}_{ij}) = \sigma_{ij}$, and $E(\hat{\sigma}_{ij}^D) \neq \sigma_{ij}$, in general. Consequently,

$$
p \times \left( \| \hat{\Sigma}_D(X) - \Sigma \|^2 - \| \hat{\Sigma}_D(X) - \bar{\Sigma}(X_a) \|^2 \right)
= \sum_{i,j=1}^{p} (\hat{\sigma}_{ij}^D - \sigma_{ij})^2 - (\hat{\sigma}_{ij}^D - \bar{\sigma}_{ij}^a)^2
= \sum_{i,j} (2\hat{\sigma}_{ij}^D - \sigma_{ij} - \bar{\sigma}_{ij}^a)(\bar{\sigma}_{ij}^a - \sigma_{ij})
= 2 \sum_{i,j} (\hat{\sigma}_{ij}^D - \bar{\sigma}_{ij}^a)(\bar{\sigma}_{ij}^a - \sigma_{ij}) - \sum_{i,j} (\bar{\sigma}_{ij}^a - \sigma_{ij})(\sigma_{ij} - \bar{\sigma}_{ij}^a) - \sum_{i,j} (\bar{\sigma}_{ij}^a - \bar{\sigma}_{ij}^a)(\sigma_{ij} - \bar{\sigma}_{ij}^a).
$$

Taking expectation yields

$$
E\| \hat{\Sigma}_D(X) - \Sigma \|^2 - E\| \hat{\Sigma}_D(X) - \bar{\Sigma}(X_a) \|^2 = 0 + 0 - E\| \bar{\Sigma}(X_a) - \Sigma \|^2
$$

and thus $E\| \hat{\Sigma}_D(X) - \bar{\Sigma}(X_a) \|^2 = E\| \hat{\Sigma}_D(X) - \Sigma \|^2 + E\| \bar{\Sigma}(X_a) - \Sigma \|^2$. Note that $E\| \bar{\Sigma}(X_a) - \Sigma \|^2$ is free of the tuning parameters involved in the double shrinkage estimator so minimizing $E\| \hat{\Sigma}_D(X) - \Sigma \|^2$ with respect to $\lambda$ is equivalent to minimizing $E\| \hat{\Sigma}_D(X) - \bar{\Sigma}(X_a) \|^2$. Last, since $E\| \bar{\Sigma}(X) - \Sigma \|^2 \propto E\| \bar{\Sigma}(X_a) - \Sigma \|^2$, the proof is complete. □