Journal of Multivariate Analysis 131 (2014) 149–162

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Estimating high dimensional covariance matrices: A new look at the Gaussian conjugate framework

Alexis Hannart [a,*], Philippe Naveau [b]

[a] CNRS, IFAECI, Argentina
[b] CNRS, LSCE, France

HIGHLIGHTS

- We (re)introduce a class of linear shrinkage estimators of the covariance matrix.
- We follow an empirical Bayesian approach to obtain shrinkage intensity and target.
- The method is generally applicable to any class of target matrices.
- Estimators are found to outperform those of the state-of-the-art Ledoit–Wolf class.
- The implementation is computationally light.

ARTICLE INFO

Article history:
Received 28 May 2013
Available online 11 June 2014

AMS subject classifications:
62H12
62C12
62J07

Keywords:
Covariance matrix estimation
Empirical Bayes
Shrinkage estimation

ABSTRACT

In this paper, we describe and study a class of linear shrinkage estimators of the covariance matrix that is well-suited for high dimensional matrices, has a rather wide domain of applicability, and is rooted into the Gaussian conjugate framework of Chen (1979). We propose here a new look at this framework. The linear shrinkage estimator is thereby obtained as the posterior mean of the covariance, using a Bayesian Gaussian model with conjugate inverse Wishart prior, and deriving the shrinkage intensity and target matrix by marginal likelihood maximization. We introduce some extensions to the seminal approach by deriving a closed-form expression of the marginal likelihood as well as computationally light schemes for its maximization. Further, these developments are implemented in a variety of situations and include a simulation-based performance comparison with a recent, widely used class of linear shrinkage estimators. The Gaussian conjugate estimators are found to outperform these estimators in every tested situation where the latter are available and to be more widely and directly applicable.

1. Context and motivations

Estimating the covariance matrix $\mathbf{\Sigma}$ of a p-dimensional Gaussian model is a common task in statistical analysis. Yet, it is also one which is generally recognized as particularly difficult and challenging (see, e.g., [26]). Recently, the availability of very large datasets from climate science, genomics, finance, marketing applications – among others – has exacerbated this problem with sample sizes n often much smaller than the matrix dimension p (see, e.g., [17,22,27,14]). In situations where $n < p$ the sample covariance matrix $\mathbf{S}$ performs poorly and is not positive definite, i.e. it is non invertible. When p/n has a fixed limit it is known that $\mathbf{S}$ is not consistent [8]. When $n > p$, its positive-definiteness is insured but its eigenvalues

* Corresponding author.
E-mail address: alexis.hannart@cima.fcen.uba.ar (A. Hannart).

http://dx.doi.org/10.1016/j.jmva.2014.06.001
0047-259X/

Table 1
Mapping of the nine illustrative target structures used in the article.

		Variance		
		Unit variance	Common variance	Unequal variances
Correlation	$\rho_{ij} = 0$	A1	A2	A3
	$\rho_{ij} = \rho$	B1	B2	B3
	$\rho_{ij} = \rho^{\lvert i-j \rvert}$	C1	C2	C3

tend to be distorted in such a way that $\mathbf{S}$ is ill-conditioned, implying that inverting it is possible but amplifies the estimation error. Alternative estimators of $\mathbf{\Sigma}$ have been proposed within both frequentist and Bayesian approaches, yielding substantial performance improvements compared to the sample covariance estimator $\mathbf{S}$ for small sample size n. Among these, linear shrinkage estimators are obtained as a weighted average of $\mathbf{S}$ and a covariance matrix $\mathbf{\Delta}$

$$\widehat{\mathbf{\Sigma}} = \alpha \mathbf{\Delta} + (1 - \alpha)\mathbf{S}, \tag{1}$$

where the so-called shrinkage target $\mathbf{\Delta}$ is assumed to have some degree of similarity with $\mathbf{\Sigma}$. The value of the target matrix $\mathbf{\Delta}$ is usually not assumed to be known; it is commonplace to assume instead that $\mathbf{\Delta}$ has a general structure, i.e. $\mathbf{\Delta}$ is assumed to belong to a given set $\mathcal{F} \subset \mathcal{S}^{+*}$ which reflects a structural constraint (where $\mathcal{S}^{+*}$ denotes the set of symmetric positive definite matrices). We thus refer to the set $\mathcal{F}$ going forward as the *target structure*. The choice of $\mathcal{F}$ is subjective and reflects an a priori belief about $\mathbf{\Sigma}$ that may be more or less precise. For instance, it is commonplace to assume that $\mathbf{\Delta}$ is equal to the identity matrix ($\mathcal{F} = \{\mathbf{I}\}$), is proportional to the identity matrix ($\mathcal{F} = \{\lambda \mathbf{I} \mid \lambda > 0\}$) or is diagonal ($\mathcal{F} = \{\mathbf{\Lambda} \mid \mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p),\ \lambda_i > 0\}$). Other more general structures are described in Appendix A and are summarized in Table 1. Beyond these general structures, the choice of $\mathcal{F}$ can also be specific, resulting from considerations that are ad-hoc to a particular context. For instance, an application to portfolio management motivated [16] to choose a structure derived from a stock return model ([24] and Appendix A).

No matter the choice of target structure $\mathcal{F}$, Eq. (1) is thus used to constrain the estimator $\widehat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$. The shrinkage allows this structural constraint to be flexible as $\mathbf{\Sigma}$ does not need to fully match the target structure—i.e. the assumption $\mathbf{\Sigma} \in \mathcal{F}$ is not required. Indeed, the introduction of the weight α – referred to as the shrinkage intensity – enables to adjust the level of structural constraint. If $\mathcal{F}$ is highly relevant, α should be chosen close to one and even equal to one if $\mathbf{\Sigma} \in \mathcal{F}$. Conversely if the relevance of $\mathcal{F}$ is poor, then α should be chosen close to zero. The shrinkage problem is thus to jointly determine an optimal value of $\mathbf{\Delta} \in \mathcal{F}$ together with an optimal value of α in [0, 1].

In a frequentist framework, the shrinkage estimation strategy has been often described as one of building an optimal tradeoff between the bias of the estimator and its variance [16]. Indeed, $\mathbf{S}$ is known to have no bias but has a high variance – especially for small n – whereas on the other hand, $\mathbf{\Delta}$ has a small variance due to the constraint imposed by its underlying structure, but does have a bias if this structure does not perfectly match with that of $\mathbf{\Sigma}$. It is hoped that a weighted average of these two extreme estimators may thus yield a new, improved estimator which would balance bias and variance in an optimal way, thus borrowing strength from both extremes. This intuitive idea is discussed extensively and formalized mathematically in the seminal work of Ledoit and Wolf [16,17]. In line with the intuitive idea of an optimal bias–variance tradeoff, the framework introduced by these authors, hereinafter referred to as the LW framework, consists in minimizing in α and $\mathbf{\Delta}$ over $[0, 1] \times \mathcal{F}$ the mean squared error (mse) $\mathbb{E}(\|\alpha\mathbf{\Delta} + (1-\alpha)\mathbf{S} - \mathbf{\Sigma}\|^2)$ where $\|.\|$ denotes the Frobenius norm defined by $\|\mathbf{A}\|^2 = \text{Tr}(\mathbf{A} \cdot \mathbf{A}')/p$ for any $p \times p$ matrix $\mathbf{A}$, and where $\mathbb{E}(.)$ denotes the expectation w.r.t. the random matrix $\mathbf{S}$. Under this formulation, the shrinkage estimator can be viewed geometrically as the orthogonal projection of $\mathbf{\Sigma}$ on the segment generated by $\mathbf{S}$ and $\mathbf{\Delta}$ (Fig. 1). The minimization yields:

$$\mathbf{\Delta}_o = \underset{\mathbf{\Delta} \in \mathcal{F}}{\text{argmax}} \frac{(\mathbb{E}(\text{Tr}((\mathbf{S} - \mathbf{\Delta})(\mathbf{S} - \mathbf{\Sigma}))))^2}{\mathbb{E}(\text{Tr}((\mathbf{S} - \mathbf{\Delta})^2))} \quad \text{and} \quad \alpha_o = \frac{\mathbb{E}(\text{Tr}((\mathbf{S} - \mathbf{\Delta}_o)(\mathbf{S} - \mathbf{\Sigma})))}{\mathbb{E}(\text{Tr}((\mathbf{S} - \mathbf{\Delta}_o)^2))}. \tag{2}$$

Of course, Eq. (2) cannot be applied straightforwardly because the expectations in $\mathbf{S}$ therein must be evaluated to approximate the so-called *oracle estimators* α_o and $\mathbf{\Delta}_o$. The latter quantities depend on $\mathbf{\Sigma}$ and are thus not known in practice (hence the term "oracle") and must be replaced by empirical estimates to obtain the final estimators α_{lw} and $\mathbf{\Delta}_{lw}$. In favorable situations where explicit calculations can be conducted, this approach yields estimators of α and $\mathbf{\Delta}$ that have a closed form and may also have some suitable asymptotic properties. Whether or not such explicit calculations are possible depends on the choice of the target structure $\mathcal{F}$. This approach was successfully applied for the first time to our knowledge in [16] to the aforementioned stock return target structure. In [17], the same authors then adapted this approach to the generally applicable case $\mathcal{F} = \{\lambda \mathbf{I} \mid \lambda > 0\}$ to obtain:

$$\mathbf{\Delta}_{lw} = \frac{\text{Tr}(\mathbf{S})}{p}\mathbf{I} \quad \text{and} \quad \alpha_{lw} = \min\left\{ \frac{\sum_{i=1}^{n} \|\mathbf{S} - x_i x_i'\|^2}{n^2\left(\text{Tr}(\mathbf{S}^2) - \text{Tr}^2(\mathbf{S})/p\right)}, 1 \right\}. \tag{3}$$

Then, [23] developed further adaptation and extension in the LW framework to cover four additional target structures (A1, A3, B1, B2). More recently, [3] have shown that the estimators of Eq. (3) can be improved substantially, especially for