

Ledoit, Olivier; Wolf, Michael

**Working Paper**

## The power of (non-)linear shrinking: A review and guide to covariance matrix estimation

Working Paper, No. 323

**Provided in Cooperation with:**

Department of Economics, University of Zurich

*Suggested Citation:* Ledoit, Olivier; Wolf, Michael (2019) : The power of (non-)linear shrinking: A review and guide to covariance matrix estimation, Working Paper, No. 323, University of Zurich, Department of Economics, Zurich,  
<https://doi.org/10.5167/uzh-170642>

This Version is available at:

<https://hdl.handle.net/10419/200186>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



**University of  
Zurich**<sup>UZH</sup>

University of Zurich  
Department of Economics

Working Paper Series  
ISSN 1664-7041 (print)  
ISSN 1664-705X (online)

---

Working Paper No. 323

# **The Power of (Non-)Linear Shrinking: A Review and Guide to Covariance Matrix Estimation**

Olivier Ledoit and Michael Wolf

May 2019

---

# The Power of (Non-)Linear Shrinking: A Review and Guide to Covariance Matrix Estimation

Olivier Ledoit

Department of Economics

University of Zurich

CH-8032 Zurich, Switzerland

[olivier.ledoit@econ.uzh.ch](mailto:olivier.ledoit@econ.uzh.ch)

Michael Wolf

Department of Economics

University of Zurich

CH-8032 Zurich, Switzerland

[michael.wolf@econ.uzh.ch](mailto:michael.wolf@econ.uzh.ch)

May 2019

## Abstract

Many econometric and data-science applications require a reliable estimate of the covariance matrix, such as Markowitz portfolio selection. When the number of variables is of the same magnitude as the number of observations, this constitutes a difficult estimation problem; the sample covariance matrix certainly will not do. In this paper, we review our work in this area going back 15+ years. We have promoted various shrinkage estimators, which can be classified into linear and nonlinear. Linear shrinkage is simpler to understand, to derive, and to implement. But nonlinear shrinkage can deliver another level of performance improvement, especially if overlaid with stylized facts such as time-varying co-volatility or factor models.

KEY WORDS: Dynamic conditional correlations, factor models, large-dimensional asymptotics, Markowitz portfolio selection, rotation equivariance.

JEL CLASSIFICATION NOS: C13, C58, G11.

# 1 Introduction

The covariance matrix is, arguably, the second most important object in all of statistics. In practical applications, such as in Markowitz portfolio selection, the true covariance matrix is typically unknown and must be estimated from data instead. It has long been known — by academic researchers and practitioners alike — that the textbook estimator, the sample covariance matrix, suffers from the curse of dimensionality. This curse is most obvious when the matrix dimension exceeds the sample size — in which case the sample covariance matrix is singular — but it is pervasive also otherwise, unless the matrix dimension is *negligible* with respect to the sample size.

We have devoted more than fifteen years of our academic careers to shrinkage estimation of large-dimensional covariance matrices; here, the term “large dimensional” indicates a scenario where the matrix dimension is not negligible with respect to the sample size or, in other words, where the matrix dimension and the sample size are of the same magnitude.<sup>1</sup> After three papers on linear shrinkage in the first decade of this century, and a creative break, there have been an additional ten papers on nonlinear shrinkage in the second decade. So perhaps the time has come to look back in order to give an overview of our work, which we hope will serve as a useful starting point to anyone who is new to the area, whether academic researcher or practitioner.

In a large(ish) collection of individual papers, there are bound to exist differences in notation. In particular, we have used two conventions for denoting (matrix dimension, sample size) over the years, namely  $(p, n)$  for papers in statistics journals and  $(N, T)$  for papers in finance journals. Since this review paper is for a finance journal, we shall use the convention  $(N, T)$ . Needless to say, there are other differences in notation, but we cannot point them out all one by one.

Another compromise that we will have to make is in terms of mathematical rigor. Unlike in individual papers, there is no space to provide proofs. And to go one step beyond, it would also go too far to spell out all the assumptions in detail for every method, respectively approach. So we will take the liberty to be purposefully vague about some assumptions here and there and refer the reader to the corresponding paper(s) for the details instead.

Last but not least, this review is necessarily restricted to our own papers. Estimation of large-dimensional covariance matrices has become a very active research field and we simply do not have the space here to provide a comprehensive review, which would take an entire book, such as the work of [Pourahmadi \(2013\)](#).

---

<sup>1</sup>Our research does not address the scenario where the matrix dimension is vastly larger than the sample size.

The remainder of the paper is organized as follows. Section 2 presents linear shrinkage to the identity matrix, which is the starting point of the journey. Section 3 presents the first extension, namely, linear shrinking to a custom-tailored target. Section 4 presents the second and more recent extension, namely, nonlinear shrinkage. Section 5 presents an extension to dynamic models. Section 6 presents an extension to factor models. Section 7 discusses computational aspects. Finally, Section 8 concludes.

## 2 Linear Shrinkage to the Identity Matrix

In this section, and in the sections to come until further notice, the data are independent and identically distributed (i.i.d.), collected in a  $T \times N$  matrix  $X_T$ , so that the rows of the matrix correspond to observations and the columns correspond to variables. The true (or population) covariance matrix is denoted by  $\Sigma_T$  and assumed to be positive definite. To simplify the notation, we assume that all variables have mean zero. In this way, the sample covariance matrix is given by

$$S_T := \frac{1}{T} X_T' X_T .$$

**Remark 2.1** (Demeaning the Data). In many applications, variables do not have mean zero, or at least it is not known whether they have mean zero. In such a setting, it is more common to base the sample covariance matrix on the demeaned data instead; see Section 7. ■

### 2.1 Finite-Sample Analysis

The sample covariance matrix  $S_T$  is unbiased and the maximum likelihood estimator under normality. There was a time when it was thought that such an estimator would be ideal, or at least desirable. This line of thinking was changed by the seminal work of Stein (1956) and James and Stein (1961) in the related context of the estimation of a multivariate mean: In dimensions  $N > 3$ , a better estimator than the sample mean can be constructed by shrinking the sample mean to a target vector, that is, by using a linear combination of the sample mean and the target vector; the original proposal by Stein was to use the zero vector for the target. Better in which sense? In the ‘usual’ sense, that is, in the sense of the mean squared error (MSE), which is the most widely used generic risk function in statistics. This is a classic example of a bias-variance tradeoff. On the one hand, shrinking to a fixed (or structured) target introduces bias; on the other hand, shrinking reduces variance. Stein’s genius was to recognize that using the optimal

shrinkage intensity reduces the MSE compared to the sample covariance matrix, which is unbiased but exhibits high(er) variance.

Note that the optimal shrinkage intensity (that is, the weight given to the target vector) depends on population quantities, but those can be estimated from the data in practice. The work of [Stein \(1956\)](#) and [James and Stein \(1961\)](#) was so revolutionary that it took a while for it to be digested and embraced by the academic community; this process was aided by [Efron and Morris \(1973, 1975, 1977\)](#) who provided a more in-depth analysis of Stein's shrinkage method that included the suggestion of alternative shrinkage targets and some empirical examples to real data. In particular, they suggested as an alternative shrinkage target a multiple of the identity vector (of dimension  $N$ ) rather than the zero vector, where the multiplier was given as the mean of the  $N$  individual sample means (which is equal to the grand mean of all the individual observations). This alternative shrinkage target received a warmer welcome from applied researchers because the effect of shrinkage was a more intuitive one: move the small sample means up and the large sample means down (whereas shrinkage to the zero vector moved all sample means towards zero, which means moving all sample means down in case they are all positive).

The motivation of [Ledoit and Wolf \(2004b\)](#) was very simple: extend Stein's shrinkage estimation the of mean vector to the estimation of the covariance matrix, keeping the MSE risk function. To this end, denote by  $\|\cdot\|_F$  the (scaled) Frobenius norm of a square matrix; more specifically, for a  $N \times N$  matrix  $A$ , this norm is given by

$$\|A\|_F := \sqrt{\langle A, A \rangle} := \sqrt{\text{Tr}(A'A)/N} = \sqrt{\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N a_{ij}^2}, \quad (2.1)$$

where  $a_{ij}$  denotes the  $(i, j)$  entry of  $A$  and  $\text{Tr}(\cdot)$  denotes the trace of a square matrix. Note that the division by the dimension  $N$  inside the square root is not standard in the definition of the Frobenius norm (which is just the generalization of the Euclidian norm from a vector to a matrix), but we will use it for the purpose of the asymptotic analysis below. The (scaled)Frobenius loss function is given by

$$\mathcal{L}_F(\hat{\Sigma}_T, \Sigma_T) := \|\hat{\Sigma}_T - \Sigma_T\|_F^2, \quad (2.2)$$

where  $\hat{\Sigma}_T$  denotes a generic estimator of  $\Sigma_T$ . Finally, the (scaled) MSE is given by

$$\mathbb{E}[\mathcal{L}_F(\hat{\Sigma}_T, \Sigma_T)] = \mathbb{E}[\|\hat{\Sigma}_T - \Sigma_T\|_F^2]. \quad (2.3)$$

The class of estimators we considered was that of linear combinations of the identity matrix, denoted by  $I_T$ , and the sample covariance matrix, so that the optimization problem became:

$$\min_{\rho_1, \rho_2} \mathbb{E}[\|\hat{\Sigma}_T - \Sigma_T\|_F^2] \quad (2.4)$$

$$\text{s.t. } \hat{\Sigma}_T = \rho_1 I_T + \rho_2 S_T \quad (2.5)$$

The solution to this problem turns out to be

$$\Sigma_T^* := \frac{\beta_T^2}{\delta_T^2} \mu_T I_T + \frac{\alpha_T^2}{\delta_T^2} S_T, \quad (2.6)$$

with  $\mu_T := \langle \Sigma_T, I_T \rangle$ ,  $\alpha_T^2 := \|\Sigma_T - \mu_T I_T\|_F^2$ ,  $\beta_T^2 := \mathbb{E}[\|S_T - \Sigma_T\|_F^2]$ , and  $\delta_T^2 := \|S_T - \mu_T I_T\|_F^2$ ; see Equation (2.1) for the definition of the operator  $\langle \cdot, \cdot \rangle$ . Since it can be shown that  $\alpha_T^2 + \beta_T^2 = \delta_T^2$ , the solution (2.6) can also be written as a *convex* linear combination of a multiple of the identity matrix,  $\mu_T I_T$ , and  $S_T$ , namely,

$$\Sigma_T^* = \gamma_T^* \mu_T I_T + (1 - \gamma_T^*) S_T \quad \text{with} \quad \gamma_T^* := \frac{\beta_T^2}{\delta_T^2}. \quad (2.7)$$

In this way, the solution  $\Sigma_T^*$  can be interpreted as shrinking the sample covariance matrix  $S_T$  towards the *shrinkage target*  $\mu_T I_T$  with (*shrinkage*) *intensity*  $\gamma_T^* \in [0, 1]$ . The multiplier  $\mu_T$  in the shrinkage target makes intuitive sense: Since  $\mu_T$  is equal to the average diagonal element of  $\Sigma_T$  — that is, equal to the average of the individual variances — it puts the shrinkage target on the right ‘scale’ for a convex linear combination with the sample covariance matrix. For example, if all the variables are multiplied by two, the shrinkage targets gets multiplied by four, just like the sample covariance matrix and also the true covariance matrix.

An important feature of  $\Sigma_T^*$  is that it is positive definite, and thus invertible, even in the case  $N > T$ .<sup>2</sup> This is because  $\Sigma_T^*$  is a convex linear combination of a matrix that is positive definite,  $\mu_T I_T$ , and another matrix that is positive semi-definite,  $S_T$ , where the weight given to the shrinkage target  $\mu_T I_T$  is positive (in all cases of practical relevance).

## 2.2 Asymptotic Analysis

The optimal linear combination  $\Sigma_T^*$  depends, as to be expected, on unknown population quantities and must therefore be thought of as an ideal (or ‘oracle’) but infeasible estimator. But knowing its formula, it is not difficult to derive a feasible estimator that, asymptotically, is just as good. When we say “asymptotically”, we must specify what we mean by that. In most of our work, we have used *large-dimensional* (or Kolmogorov) asymptotics, where the dimension,  $N$ , is allowed to go to infinity together with the sample size,  $T$ . Since we are only interested in situations where  $N$  is of the same magnitude as  $T$ , and not situations where  $N$  is vastly larger than  $T$ , we may assume without loss of generality that  $N/T \rightarrow c \in [0, \infty)$ , where  $c$  is called the *limiting concentration (ratio)*.<sup>3</sup>

<sup>2</sup>In this case the sample covariance is rank-deficient and thus not invertible.

<sup>3</sup>The paper assumes the weaker condition that there exist a finite constant  $K_1$  such that  $N/T < K_1$  always, but there does not appear to be any practical relevance in having this weaker condition.

We further may assume without loss of generality that  $N$  is an implicit function of  $T$ , so that we can index all quantities simply by  $T$ , as before.

Based on Equation (2.7), we only need to estimate the three parameters  $\mu_T$ ,  $\delta_T^2$ , and  $\beta_T^2$ . The respective solutions are as follows. First,

$$\hat{\mu}_T := \langle S_T, I_T \rangle = \frac{1}{N} \sum_{i=1}^N s_{ii}^T, \quad (2.8)$$

where  $s_{ij}^T$  denotes the  $(i, j)$  entry of  $S_T$ . Second,

$$\hat{\delta}_T^2 := \|S_T - \hat{\mu}_T I_T\|_F^2. \quad (2.9)$$

Third, denote by  $x_t^T$  the  $t^{\text{th}}$  row of the  $T \times N$  data matrix  $X_T$  ('converted' to a proper  $N \times 1$  vector), so that, in particular, it holds that

$$S_T := \frac{1}{T} X_T' X_T = \frac{1}{T} \sum_{t=1}^T x_t^T (x_t^T)'. \quad (2.10)$$

The estimator of  $\beta_T^2$  is then given by

$$\hat{\beta}_T^2 := \min\{\tilde{\beta}_T^2, \hat{\delta}_T^2\} \quad \text{with} \quad \tilde{\beta}_T^2 := \frac{1}{T} \sum_{t=1}^T \|x_t^T (x_t^T)' - S_T\|_F^2, \quad (2.11)$$

where the truncation of  $\tilde{\beta}_T^2$  is used to ensure a proper convex linear combination in the feasible shrinkage estimator

$$S_T^* := \hat{\gamma}_T^* \hat{\mu}_T I_T + (1 - \hat{\gamma}_T^*) S_T \quad \text{with} \quad \hat{\gamma}_T^* := \frac{\hat{\beta}_T^2}{\hat{\delta}_T^2}. \quad (2.12)$$

(Note that in practice this truncation rarely kicks in.)

Under a set of regularity conditions one can show that the three estimators (2.8)–(2.10) are consistent in quadratic mean, which implies that  $S_T^*$  is also a consistent estimator of  $\Sigma_T^*$  in quadratic mean, namely that

$$\mathbb{E}[\|S_T^* - \Sigma_T^*\|_F^2] \rightarrow 0. \quad (2.13)$$

(The corresponding regularity conditions consist of moment conditions and certain distributional assumptions that are weaker than assuming an elliptical distribution, let alone a multivariate normal distribution.)

An implication of (2.12) is that the feasible estimator  $S_T^*$  has asymptotically the same risk as the infeasible optimal linear combination  $\Sigma_T^*$  for estimating  $\Sigma_T$  in the sense that

$$\mathbb{E}[\|S_T^* - \Sigma_T\|_F^2] - \mathbb{E}[\|\Sigma_T^* - \Sigma_T\|_F^2] \rightarrow 0. \quad (2.14)$$



**Remark 2.2.** It is important not to mistake consistent estimation of  $\Sigma_T^*$  — that is, consistent estimation of the optimal linear combination (2.7) — for consistent estimation of  $\Sigma_T$  itself — that is, for consistent estimation of the true covariance matrix. The latter, stronger result obtains in the special case when the limiting concentration ratio  $c$  is zero, that is, when  $N/T \rightarrow 0$ ; this case includes traditional asymptotics where  $N$  is fixed whereas  $T$  alone tends to infinity. In such a case, already the sample covariance matrix  $S_T$  is a consistent estimator, so nothing is gained by using a shrinkage estimator instead, at least not asymptotically. On the other hand, consistency does not obtain in general for the case  $c > 0$ , which is the relevant case in situations where  $N$  is not negligible with respect to  $T$ . This should not be surprising, since one cannot expect to estimate  $N(N+1)/2$  parameters (namely, the distinct entries of the symmetric matrix  $\Sigma_T$ ) from  $N \times T$  random univariate realizations if these two numbers are of the same magnitude, at least not in the absence of restrictive assumptions.

There is a different strand of literature, going back to at least [Bickel and Levina \(2008b,a\)](#), that makes the strong (and unverifiable in practice) assumption that the true covariance matrix  $\Sigma_T$  is sparse. In such a restrictive setting, consistent estimation of  $\Sigma_T$  itself is possible even when  $c > 0$ . ■

The feasible estimator  $S_T^*$  shares with the optimal linear combination  $\Sigma_T^*$  the important property of being positive definite and thus invertible (with probability one) even in the case when  $N > T$ . Indeed, as can be seen from Equation (2.11),  $S_T^*$  is also a convex linear combination of a positive definite (with probability one) matrix,  $\hat{\mu}_T I_T$ , and positive semi-definite matrix,  $S_T$ , where the weight given to the target  $\hat{\mu}_T I_T$  is positive (with probability one).

## 2.3 Simulation Evidence

We compared the finite-sample performance of the feasible estimator  $S_T^*$  to that of three other estimators from the literature via Monte Carlo simulations:

- The empirical Bayes estimator of [Haff \(1980\)](#)
- The better performing estimator, for any given simulated data set, of [Stein \(1975\)](#) and [Haff \(1982\)](#)
- The minimax estimator derived independently by [Stein \(1982\)](#) and [Dey and Srinivasan \(1985\)](#)

(The latter two estimators were derived under a different loss function, namely Stein’s loss rather than Frobenius loss; see Section 4.5. But at the beginning of the century, estimation of large-dimensional covariance matrices was not an active field yet and so there were not

many estimators to choose from; in particular, the empirical Bayes estimator of [Haff \(1980\)](#) was the only other estimator that we were aware of that was also derived under Frobenius loss.)

Over a wide range of scenarios, all four estimators improved over the sample covariance matrix in terms of empirical MSE, and the shrinkage estimator  $S_T^*$  was overall the best; in particular our estimator improved over the sample covariance matrix in every single scenario. Moreover, as long as both  $N, T \geq 20$ , the finite-sample performance was already well approximated by asymptotic results.

## 2.4 Applications

The demand for a well-conditioned estimator of large-dimensional covariance matrices in applied research is great, and far greater than we had originally imagined. Indeed, our estimator has been used in a variety of different fields for a wide range of applications. To list some representative examples only:

**Acoustics** Optimally removing noise from signals captured from an array of hydrophones ([Zhang et al., 2009](#)).

**Cancer Research** Mapping out the influence of the Human Papillomavirus (HPV) on gene expression ([Pyeon et al., 2007](#)).

**Chemistry** Estimating the temporal autocorrelation function (TACF) for fluorescence correlation spectroscopy ([Guo et al., 2012](#)).

**Civil Engineering** Detecting and identifying vibration-based bridge damage through Random Coefficient Pooled (RCP) models ([Michaelides et al., 2011](#)).

**Climatology** Detecting trends in average global temperature through the optimal fingerprinting method ([Ribes et al., 2013](#)).

**Electromagnetics** Studying correlation between reverberation chamber measurements collected at different stirrer positions ([Pirkl et al., 2012](#)).

**Entertainment Technology** Designing a video game controlled by performing tricks on a skateboard ([Anlauff et al., 2010](#)).

**Genetics** Improving the accuracy of breeding value (GEBV) predictions with low-density markers ([Endelman and Jannink, 2012](#)).

**Geology** Modeling multiphase flow in subsurface petroleum reservoirs with the iterative stochastic ensemble method (ISEM) on inverse problems ([Elsheikh et al., 2013](#)).

**Image Recognition** Detecting anomalous pixels in hyperspectral imagery ([Bachega et al., 2011](#)).

**Macroeconomics** Improved GLS regressions of stochastic discount factor models

(Korniotis, 2008).

**Neuroscience** Calibrating brain-computer interfaces (Lotte and Guan, 2009).

**Psychology** Modeling co-morbidity patterns among mental disorders (Markon, 2010).

**Road Safety Research** Developing an emergency braking assistance system (Haufe et al., 2011).

**Signal Processing** Adaptive ‘Capon’ beamforming to recover signals of interest (Abrahamsson et al. (2007)).

**Speech Recognition** Automatic transcription of phone conversation records (Bell and King, 2009).

### 3 Linear Shrinkage to a Custom-Tailored Target

Our own motivation in deriving reliable estimators of large-dimensional covariance matrices was mainly in the area of finance, namely for the application of Markowitz-type portfolio selection. So why is finance missing in the previous list of fields? This is because we devised some alternative estimators, at the same time already, that tend to work even better for finance applications. These estimators recognize that financial covariance matrices typically have some stylized features that can be exploited in devising an improved shrinkage target compared to using (a multiple of) the identity matrix, which is the natural choice for a generic target.

To build some motivation, what is a good shrinkage target? It should come as close as possible to the true covariance matrix with as few parameters as possible. So it is a ‘balancing act’ between accuracy and parsimony. A good shrinkage target will benefit from application-specific knowledge and thus involves some custom tailoring. If our goal is to estimate the covariance matrix of a universe of stock returns, then we can exploit one of the following known features: first, stock returns have a factor-model structure, at least to some extent; second, the average correlation is positive; third, the average covariance is positive also. Depending on which feature we want to use, the resulting shrinkage target will differ accordingly.

In Ledoit and Wolf (2003), we used the first feature, namely the capital asset pricing model (CAPM) dating back to Sharpe (1964) and Lintner (1965). According to his model,

$$x_{ti} = \alpha_i + \beta_i x_{t0} + u_{ti} \quad i = 1, \dots, N, \quad t = 1, \dots, T, \quad (3.1)$$

where  $x_{ti}$  is the return on stock  $i$  in period  $t$ ,  $\alpha_i$  and  $\beta_i$  are parameters specific to stock  $i$ ,  $x_{t0}$  is the market return in period  $t$ , and  $u_{ti}$  is an error term that satisfies  $\mathbb{E}(u_{ti}|x_{t0}) = 0$  and  $\mathbb{E}(u_{ti}u_{tj}) = 0$  for  $i \neq j$ . Model (3.1) implies that  $\Sigma_T = \Phi_T$  with

$$\Phi_T := \sigma_{00}\beta_T\beta_T' + \Delta_T, \quad (3.2)$$

where  $\sigma_{00} := \text{Var}(x_{t0})$ ,  $\beta_T := (\beta_1, \dots, \beta_N)'$  and  $\Delta_T$  is a diagonal matrix with typical entry  $\delta_{ii} := \text{Var}(u_{ti})$ . The matrix  $\Phi_t$  is unknown in practice but it can be estimated as

$$F_T := \hat{\sigma}_{00}\hat{\beta}_T\hat{\beta}_T' + \hat{\Delta}_T, \quad (3.3)$$

where  $\hat{\sigma}_{00}$  is the sample variance of the  $\{x_{t0}\}$ ,  $\hat{\beta}_T$  is obtained through estimating model (3.1) by ordinary least squares (OLS), one stock at a time, and a typical element of  $\hat{\Delta}_T$  is given by  $\hat{\delta}_{ii}$ , taken to be the sample variance of the OLS residuals  $\{\hat{u}_{ti}\}$ .

If one believes in the CAPM,  $F_t$  is the ‘perfect’ estimator of  $\Sigma_t$ : Just like the sample covariance matrix  $S_T$  it is (asymptotically) unbiased, but it only has  $2N + 1$ , rather than  $N(N + 1)/2$ , parameters and therefore contains much less estimation error (that is, variance). The problem is that no one believes in the CAPM anymore, at least not as a ‘perfect’ model. Indeed, the CAPM serves as useful approximation but its assumptions do not hold exactly true; in particular, for many pairs of stocks  $(i, j)$ , there is evidence that  $\mathbb{E}(u_{ti}u_{tj}) \neq 0$ ; for example, for pairs of stocks belonging to the same industry. The idea of Ledoit and Wolf (2003) was, therefore, to use  $F_T$  as a shrinkage target ‘only’, that is, to use a convex linear combination  $\gamma_T F_T + (1 - \gamma_T)S_T$  as the estimator of  $\Sigma_T$ . Intuitively, the closer  $F_T$  is to the true  $\Sigma_T$ , the larger should be the shrinkage intensity  $\gamma_T$ .

Clearly, this approach can also be used with other (feasible) shrinkage targets. In Ledoit and Wolf (2004a), we used the second feature mentioned above, namely the fact that the average correlation of stock returns is positive. The infeasible (or population) target is given by  $\Phi_T$  with typical element  $\phi_{ij}^T = \sqrt{\sigma_{ii}\sigma_{jj}}\rho_T$ , where  $\rho_T$  is a common correlation, that is,  $\text{Cor}(x_{ti}, x_{tj}) \equiv \rho_T$ . The feasible shrinkage target is then given by

$$F_T \quad \text{with typical element} \quad f_{ij}^T := \sqrt{\hat{\sigma}_{ii}^T \hat{\sigma}_{jj}^T} \hat{\rho}_T, \quad (3.4)$$

where  $\hat{\sigma}_{ii}^T$  is the sample variance of the  $\{x_{ti}\}$  and  $\hat{\rho}_T$  is the average of the  $N(N - 1)$  sample correlations between  $\{x_{ti}\}$  and  $\{x_{tj}\}$ , for  $1 \leq i \neq j \leq N$ . Note that for this shrinkage target  $N + 1$  parameters need to be estimated.

Using the third feature mentioned above, namely the fact that the average covariance of stock returns is also positive, suggests a shrinkage target  $\Phi_T$  that has a common variance  $\sigma^2$  on the diagonal and a common covariance  $\eta$  on the off-diagonal. The feasible shrinkage target is then given by

$$F_T \quad \text{with} \quad f_{ii}^T := \hat{\sigma}_T^2 \quad \text{and} \quad f_{ij}^T := \hat{\eta}_T, \quad (3.5)$$

where  $\hat{\sigma}_T^2$  is the average of the  $N$  sample variances of the  $\{x_{ti}\}$ , for  $i = 1, \dots, N$ , and  $\hat{\eta}_T$  is the average of the  $N(N - 1)$  sample covariances between  $\{x_{ti}\}$  and  $\{x_{tj}\}$ , for  $1 \leq i \neq j \leq N$ . Note that for this shrinkage target only two parameters need to be

estimated. Obviously, the two-parameter shrinkage target (3.5) is a generalization of the generic target used in Equation (2.11), which has the same value  $\hat{\sigma}_T^2$  on the diagonal but sets the off-diagonal elements to  $\hat{\eta}_T := 0$ . Therefore, the generic shrinkage target seems less appropriate for a covariance matrix of a universe of stock returns. The target (3.5) was first suggested by Schäfer and Strimmer (2005) in a different context related to genomics and has been used by Wolf and Wunderli (2012) to estimate the covariance matrix of a universe of hedge fund returns.

### 3.1 Finite-Sample Analysis

Again, we base the optimal solution on the (scaled) Frobenius loss function, so that the minimization problem becomes

$$\min_{\gamma_T} \mathbb{E}[\|\hat{\Sigma}_T - \Sigma_T\|_F^2] \quad (3.6)$$

$$\text{s.t. } \hat{\Sigma}_T = \gamma_T F_T + (1 - \gamma_T) S_T \quad (3.7)$$

Assuming that  $\mathbb{E}(F_T) = \Phi_T$ , the solution to this problem turns out to be

$$\gamma_T^* := \frac{\sum_{i=1}^N \sum_{j=1}^N [\text{Var}(s_{ij}^T) - \text{Cov}(f_{ij}^T, s_{ij}^T)]}{\sum_{i=1}^N \sum_{j=1}^N [\text{Var}(f_{ij} - s_{ij}^T) + (\phi_{ij}^T - \sigma_{ij}^T)]} , \quad (3.8)$$

resulting in the optimal linear combination

$$\Sigma_T^* := \gamma_T^* F_T + (1 - \gamma_T^*) S_T . \quad (3.9)$$

(More generally, one needs to replace  $\phi_{ij}^T$  with  $\mathbb{E}(f_{ij}^T)$  in (3.8) if  $\mathbb{E}(F_T) \neq \Phi_T$ , but doing so does not affect our asymptotic analysis below, as long as  $\mathbb{E}(F_T) = \Phi_T$  holds true asymptotically.)

This optimal linear combination  $\Sigma_T^*$  can be interpreted as an oracle (and thus infeasible) empirical Bayes estimator, in the sense that the shrinkage target  $F_T$  can be thought of as a *data-dependent* prior. Importantly, this prior is computed from the same set of data as the sample covariance matrix  $S_T$  itself, and this fact should be reflected in the optimal shrinkage intensity  $\gamma_T^*$ ; this is indeed the case as will be explained below.

### 3.2 Asymptotic Analysis

Again, the optimal linear combination  $\Sigma_T^*$  depends on unknown population quantities and must be estimated in practice to obtain a feasible estimator. Unfortunately, in this more general setting, we have not been able to derive a consistent estimator under large-dimensional asymptotics and, therefore, had to settle for standard (or traditional)

asymptotics where  $N$  remains fixed and only  $T$  tends to infinity. Such a setting is not entirely satisfactory on theoretical grounds but it can (and does) still yield feasible estimators that perform well in practice.

Under traditional asymptotics, and the mild assumption that  $\Phi_T \neq \Sigma_T$ , it can be shown that

$$\gamma_T^* = \frac{1}{T} \frac{\pi_T - \rho_T}{\nu_T} + O\left(\frac{1}{T^2}\right) \quad (3.10)$$

$$\text{with } \pi_T := \sum_{i=1}^N \sum_{j=1}^N \text{AsyVar}(\sqrt{T} s_{ij}^T), \quad (3.11)$$

$$\rho_T := \sum_{i=1}^N \sum_{j=1}^N \text{AsyCov}(\sqrt{T} f_{ij}^T, \sqrt{T} s_{ij}^T), \quad (3.12)$$

$$\text{and } \nu_T := \sum_{i=1}^N \sum_{j=1}^N (\phi_{ij}^T - \sigma_{ij}^T)^2, \quad (3.13)$$

where  $\text{AsyVar}(\cdot)$  stands for asymptotic variance and  $\text{AsyCov}(\cdot)$  stands for asymptotic covariance, where we tacitly assume a set of regularity conditions to ensure that these quantities exist.<sup>4</sup>

It is instructive to briefly study the influence of the three terms that co-determine the optimal shrinkage intensity  $\gamma_T^*$ , up to higher-order terms.

First, the term  $\pi_T$  measures the estimation uncertainty in the sample covariance matrix  $S_T$ ; *ceteris paribus*, the larger this estimation uncertainty, the larger should be the shrinkage intensity, which is indeed the case. Second, the term  $\rho_T$  measures the ('combined') covariance between the data-dependent shrinkage target  $F_T$  and  $S_T$ . The inclusion of this term explicitly accounts for the fact that the shrinkage target is computed from the same set of data as  $S_T$  and was ignored in related work by [Frost and Savarino \(1986\)](#), for example; *ceteris paribus*, the larger the covariance between  $F_T$  and  $S_T$ , the smaller should be the shrinkage intensity,<sup>5</sup> which is indeed the case. Third, the term  $\nu_T$  measures how close the population version of the shrinkage target,  $\Phi_T$ , is to the population covariance matrix,  $\Sigma_T$ ; *ceteris paribus*, the closer the two are to each other, the larger should be the shrinkage intensity, which is indeed the case as well.

Given formula (3.10), the estimation of the optimal shrinkage intensity  $\gamma_T^*$  is straightforward: estimate the three terms  $\pi_T$ ,  $\rho_T$ , and  $\nu_T$  separately, and ignore higher-order terms.

---

<sup>4</sup>Consider an estimator  $\hat{\theta}_T$  of a parameter  $\theta$  that satisfies  $\sqrt{T}(\hat{\theta}_T - \theta) \xrightarrow{d} N(0, \sigma^2)$ , where  $\xrightarrow{d}$  denotes convergence in distribution. Then, in our parlance,  $\text{AsyVar}(\sqrt{T}\hat{\theta}_T) = \sigma^2$ ; and analogous for the definition of the operator  $\text{AsyCov}(\cdot, \cdot)$ .

<sup>5</sup>Intuitively, this is because a large covariance implies that  $F_T$  provides little independent information about  $\Sigma_T$  relative to  $S_T$ .

First, a consistent estimator of  $\pi_T$  is standard and given by

$$\hat{\pi}_T := \sum_{i=1}^N \sum_{j=1}^N \hat{\pi}_{ij}^T \quad \text{with} \quad \hat{\pi}_{ij}^T := \frac{1}{T} \sum_{t=1}^T \left[ x_{ti}^T x_{tj}^T - s_{ij}^T \right]^2.$$

(Note that this estimator is numerically equal to the estimator  $\tilde{\beta}_T$  of Equation (2.10).)

Second, a consistent estimator of  $\rho_T$  depends on the choice of shrinkage target  $F_T$  and thus requires a case-by-case analysis. We provide the corresponding details for the choices (3.3) and (3.4) in Ledoit and Wolf (2003, Lemma 2) and Ledoit and Wolf (2004a, Appendix B), respectively. Some details for the choice (3.5) can be found in Schäfer and Strimmer (2005, Appendix A).

Third, a consistent estimator of  $\nu_T$  is simply given by

$$\hat{\nu}_T := \sum_{i=1}^N \sum_{j=1}^N (f_{ij}^T - s_{ij}^T)^2,$$

and thus does not require any extra work.

In this way, we get an estimator of the optimal shrinkage intensity as

$$\hat{\gamma}_T^* := \min\{\max\{\tilde{\gamma}_T^*, 0\}, 1\} \quad \text{with} \quad \tilde{\gamma}_T^* := \frac{1}{T} \frac{\hat{\pi}_T - \hat{\rho}_T}{\hat{\nu}_T}, \quad (3.14)$$

where the truncation of  $\tilde{\gamma}_T^*$  is used to ensure a proper convex linear combination in the feasible shrinkage estimator

$$S_T^* := \hat{\gamma}_T^* F_T + (1 - \hat{\gamma}_T^*) S_T. \quad (3.15)$$

(Note that in practice this truncation rarely kicks in.)

The methodology for estimating  $\hat{\gamma}_T^*$  is completely generic and can be easily adapted to other shrinkage targets as well. The only ‘hard’ work to be done is to find a consistent estimator of the term  $\rho_T$ , which requires a case-by-case analysis. For example, consider a setting where we want to estimate the covariance matrix of a universe of assets that belong to two different asset classes, such as stocks and bonds. Then a sensible population shrinkage target  $\Phi_T$  would be one having five parameters: two common variances (one for each asset class), two common within-covariances (one for each asset class), and one common across-covariance; the feasible shrinkage target  $F_T$  can be estimated in the obvious way and the estimator  $\hat{\gamma}_T^*$  of corresponding optimal shrinkage intensity  $\gamma_T$  can be derived with the methodology described above.

### 3.3 Simulation Evidence

In Ledoit and Wolf (2004a), we carried out a ‘hybrid’ empirical analysis that is part real-life back-testing and part simulation study. The idea was to emulate a portfolio manager

that has a certain amount of skill in predicting future stock returns. The setting is one of a manager trying to outperform a given index subject to a constraint on the tracking error, which can be formulated as a Markowitz-type portfolio selection problem requiring a good estimator of the covariance matrix of the stock returns in practice. For the forecasts of the manager we took the actual stock returns (in the upcoming, ‘future’ period) and added a certain amount of noise to them in order to match a desired level of skill. The performance measure was the information ratio of the excess returns of the portfolio (that is, the returns in excess of the index returns).

We compared four estimators of the covariance matrix: the sample covariance matrix, linear shrinkage to the single-factor matrix (3.3), linear shrinkage to the constant-correlation matrix (3.4), and a matrix based on a five-factor model where the factors were the first five principal components of the sample covariance matrix; this last estimator is in the spirit of Connor and Korajczyk (1988, 1993). We also considered five different portfolio sizes  $N \in \{30, 50, 100, 225, 500\}$ .

All three ‘improved’ estimators of the covariance matrix dominated the sample covariance matrix, but there was no clear winner among the three. If anything, shrinkage to the constant-correlation matrix was best for portfolio sizes  $N \leq 100$ , whereas shrinkage to the single-factor matrix was best for portfolio sizes  $N \geq 225$ .

### 3.4 Applications

There are plenty of applications of linear shrinkage to the single-factor matrix or the constant-correlation matrix, going back to the empirical analysis we provided already in Ledoit and Wolf (2003). Most applications are in the context of Markowitz-type portfolio selection and are too numerous to list them all. We think it is fair to say that, for quite some time, no one was able to outperform linear shrinkage by alternative methods. For example, DeMiguel et al. (2009) proposed the methodology of norm-constraining (that is, using an upper bound on the norm of a portfolio weight vector, such as on the gross exposure) to construct improved portfolios, as an alternative to Markowitz-type portfolio selection based on an improved estimator of the covariance matrix of asset returns. In a backtest exercise involving a universe of 500 randomized stocks, the various portfolios they proposed all (weakly) underperformed the Markowitz-type portfolio based on covariance matrix derived by shrinkage to the single-factor matrix.<sup>6</sup>

An application outside of Markowitz-type portfolio selection involves the famous

---

<sup>6</sup>Backtest exercises involving (smaller) universes of portfolios as assets, such as ten industry portfolios or 25 Fama-French portfolios yielded inconclusive results, with neither methodology dominating the other.



Hansen and Jagannathan (1997) distance test for the evaluation of stochastic discount factors. Unfortunately, the test in its originally suggested form is quite liberal in finite samples, meaning that the probability of rejecting a true null hypothesis can be far above the nominal significance level. The main reason for this fact is that the test statistic needs an estimator of the inverse of a certain covariance matrix and using the standard estimator based on the sample covariance matrix does not work well. Ren and Shimotsu (2009) show that if instead one uses an estimator based on shrinkage to the single-factor model, the size-distortion problems of the test are greatly alleviated.

Yet another application that uses shrinkage to the single-factor matrix for improved estimation of a covariance matrix in order to whiten and demean data as one of many ‘wheels’ in the estimation of a multivariate GARCH model; see Broda and Paoletta (2009).

## 4 Nonlinear Shrinkage

Linear shrinkage to a custom-tailored target is one way of generalizing and improving the generic method of linear shrinkage to (a multiple of) the identity matrix. But it requires a judicious choice of the shrinkage target, which must be based on known features of the true covariance matrix for the application at hand.

Is it possible to generalize and improve linear shrinkage to (a multiple of) the identity matrix in the absence of such knowledge? In other words, can we be totally ignorant about the true covariance matrix and still do better than generic linear shrinkage?

The intuitive way of interpreting the optimal linear combination (2.7) is the one of moving each entry of the sample covariance matrix  $S_T$  to the shrinkage target  $\mu_T I_T$  with *common* intensity  $\gamma_T^*$ . A generalization that comes immediately to mind would be to use different intensities for different entries; for example, entries of  $S_T$  that have relatively more (less) sampling error should be moved more (less) to the corresponding entries of  $\mu_T I_T$ . But there are two problems with this idea. First, the number of distinct entries of  $S_T$  is of the order  $N^2$  and so one would have to consider a very large and rapidly growing number of different shrinkage intensities. Second, and more importantly, if different shrinkage intensities are used on the different entries of  $S_T$ , there is no (easy) way of ensuring that the resulting shrinkage estimator will be positive semi-definite, let alone positive definite. Therefore, we needed another starting point in order to generalize linear shrinkage to the identity matrix.

The proper starting point the *spectral decomposition* of the sample covariance matrix, which is given by  $S_T = U_T \Lambda_T U_T'$ . Here  $\Lambda_T := \text{Diag}(\lambda_{T,1}, \dots, \lambda_{T,N})$  is a diagonal matrix<sup>7</sup>

---

<sup>7</sup>In slight abuse of notation, we will use the operator  $\text{Diag}(\cdot)$  for two different purpose. On the one hand, for a  $N \times 1$  vector  $a := (a_1, \dots, a_N)'$ ,  $\text{Diag}(a_1, \dots, a_N)$  denotes a  $N \times N$  diagonal matrix

whose diagonal elements are the sample eigenvalues  $\lambda_{T,i}$  and  $U_T := [u_{T,1}, \dots, u_{T,N}]$  is a (full) orthogonal matrix whose columns are the sample eigenvectors  $u_{T,i}$ . This starting point provides an alternative interpretation of the optimal linear combination (2.7) which we already pointed out in Ledoit and Wolf (2004b): One can also express this linear combination as

$$\Sigma_T^* := U_T \Delta_T^* U_T' \text{ with } \Delta_T^* := \text{Diag}(\delta_{T,1}^*, \dots, \delta_{T,N}^*) \text{ and } \delta_{T,i}^* := \gamma_T^* \mu_T + (1 - \gamma_T^*) \lambda_{T,i} . \quad (4.1)$$

Therefore,  $\Sigma_T^*$  has the same eigenvectors as  $S_T$  but replaces the sample eigenvalues  $\lambda_{T,i}$  with convex linear combinations  $\gamma_T^* \mu_T + (1 - \gamma_T^*) \lambda_{T,i}$ . This means that shrinking the elements of  $S_T$  to the elements of the target  $\mu_T I_T$  with common intensity  $\gamma_T^*$  is equivalent to keeping the eigenvectors of  $S_T$  and shrinking its eigenvalues to the target  $\mu_T$  with *the same* common intensity  $\gamma_T^*$ . The generalization is now obvious: use different shrinkage intensities for different sample eigenvalues; or, equivalently, move a given sample eigenvalue by an ‘individual’ amount, up or down! (This approach allows for individual shrinkage intensities to be negative, that is, to move sample eigenvalues away from the target  $\mu_T$ .) Clearly, this more general approach will be at least as good as using a common shrinkage intensity, at least as long as the distinct intensities are chosen ‘suitably’. What is more, this approach will lead to a positive-definite estimator as long as all the transformed eigenvalues are positive.

Interestingly, this approach was already introduced by Stein (1975, 1986) as a class of *rotation-equivariant* estimators. Let  $W_T$  be a rotation matrix, that is, an orthogonal matrix with determinant one. Also, let  $\hat{\Sigma}_T(\cdot)$  denote a generic estimator of the covariance matrix  $\Sigma_T$ . An estimator is said to be rotation-equivariant if it satisfies

$$\hat{\Sigma}_T(W_T X_T) = W_T \hat{\Sigma}_T(X_T) W_T' ,$$

that is, rotating the data results in an according rotation of the estimator.<sup>8</sup> In the absence of any a priori knowledge on the structure of  $\Sigma_T$ , rotation-equivariance is a natural and desirable property of a covariance matrix estimator.

Stein (1975, 1986) proposed the following class:

$$\hat{\Sigma}_T := U_T \Delta_T U_T' \quad \text{with} \quad \Delta_T := \text{Diag}(\delta_{T,1}, \dots, \delta_{T,N}) . \quad (4.2)$$

Here,  $\Delta_T$  is an unrestricted diagonal matrix, apart from assuming that  $\min\{\delta_{T,i}\} \geq 0$  so that the estimator  $\hat{\Sigma}_T$  will be positive semi-definite. Imposing the stronger assumption  $\min\{\delta_{T,i}\} > 0$  ensures that the estimator  $\hat{\Sigma}_T$  will be positive definite.

---

whose diagonal is the vector  $a$ . On the other hand, for a  $N \times N$  matrix  $A$  with typical element  $a_{ij}$ ,  $\text{Diag}(A)$  denotes the  $N \times 1$  vector  $(a_{11}, \dots, a_{NN})'$ .

<sup>8</sup>Stein actually used the term “rotation-invariant” instead, which gives the wrong idea, strictly speaking: rotating the data leaves the estimator unchanged.

## 4.1 Finite-Sample Analysis

It had been our original plan to do as before with linear shrinkage, that is, to minimize the Frobenius risk in the class of considered estimators to find the optimal one in finite samples. However, it turned out that, in this setting, we could do even better, namely, minimize the actual Frobenius loss in the class of considered estimators. Needless to say, an estimator that minimizes the loss also minimizes the risk.

The optimization problem thus became:

$$\min_{\Delta_T} \|U_T \Delta_T U_T' - \Sigma_T\|_F^2 \quad (4.3)$$

$$\text{s.t. } \Delta_T = \text{Diag}(\delta_{T,1}, \dots, \delta_{T,N}) \quad (4.4)$$

The solution to this problem turns out to be

$$\Delta_T^* := \text{Diag}(\delta_{T,1}^*, \dots, \delta_{T,N}^*) \quad \text{with} \quad \delta_{T,i}^* := u_{T,i}' \Sigma_T u_{T,i} . \quad (4.5)$$

The optimal diagonal elements  $\delta_{T,i}^*$ , unsurprisingly, are not equal to the sample eigenvalues, since then we would recuperate the sample covariance matrix as the estimator. But, less surprisingly perhaps, the optimal diagonal elements are also not equal to the population eigenvalues; this is because the population eigenvalues are a perfect match for the population eigenvectors — which combination would recuperate the true covariance matrix — but not for the sample eigenvectors. The problem is that in practice we only get to observe the sample eigenvectors but not the population eigenvectors. The optimal diagonal elements  $\delta_{T,i}^*$  combine sample and population information, namely how the  $i^{\text{th}}$  sample eigenvector  $u_{T,i}$  relates to the population covariance matrix  $\Sigma_T$ .

**Remark 4.1** (Relation to PCA). Another way to look at the optimal element  $\delta_{T,i}^*$  is that it is equal to the variance of the  $i^{\text{th}}$  principal component, where the principal components are derived from the sample covariance matrix, as is customary in practice. By being able to consistently estimate the quantities  $\delta_{T,i}^*$ , one can carry out improved principal component analysis (PCA) in situations where  $N$  is of the same magnitude as  $T$ . Because of space constraints, we will not address this topic any further here; the interested reader is referred to [Ledoit and Wolf \(2015, Section 4\)](#) for the corresponding details. ■

The finite-sample optimal ‘estimator’ is then given by

$$\Sigma_T^* := U_T \Delta_T^* U_T' \quad \text{with} \quad \Delta_T^* \text{ as defined in (4.5)} . \quad (4.6)$$

An important feature of  $\Sigma_T^*$  is that it is positive definite, and thus invertible, even in the case when  $N > T$ . This is because  $\delta_{T,i}^* > 0$ , for  $i = 1, \dots, N$ .

We actually already mentioned the elements  $\delta_{T,i}^*$  in [Ledoit and Wolf \(2004b, p. 374\)](#), but without formally proving their finite-sample optimality. Also, at that time we had no idea how to carry out asymptotic analysis in order to find a related feasible estimator; we had to discover, and extend, a whole new machinery to this end, as outlined below.

## 4.2 Asymptotic Analysis

The ‘estimator’ (4.6) is not feasible in practice (hence the use of the single quotation marks), since the optimal diagonal elements  $\delta_{T,i}^*$  depend on the very object that we want to estimate, namely the true covariance matrix  $\Sigma_T$ .

What complicates matters compared to linear shrinkage is that the number of parameters is not fixed but is equal to  $N$  and thus tends to infinity. Therefore, it is useful to rephrase the estimation problem in order to only have a single, dimension-free ‘object’ to estimate. The proper object to think about is a *function* that transforms the sample eigenvalues to the diagonal elements of the matrix  $\Delta_T$ ; clearly, such a function does not depend on the number of diagonal elements,  $N$ , and is thus dimension free. For full flexibility, we do not want to impose any restriction on such a function, other than that the output must be positive in order to obtain a positive-definite and thus invertible matrix. For the purpose of asymptotic analysis, the function may depend on the sample size  $T$  but converges to a (non-stochastic) limit as  $T$  tends to infinity. The goal is then to find the optimal limiting function, and estimate it consistently from the observed data. Since any estimator depends on the data, we must also allow for the functions used in practice to be data-dependent (or stochastic).

The class of estimators we consider is therefore given by

$$\hat{\Sigma}_T := U_T \hat{\Delta}_T U_T \quad \text{with} \quad \hat{\Delta}_T := \text{Diag}(\hat{\delta}_{T,1}, \dots, \hat{\delta}_{T,N}) , \quad (4.7)$$

where  $\hat{\delta}_{T,i} := \hat{\phi}_T(\lambda_{N,i})$  and  $\hat{\phi}_T : \mathbb{R} \rightarrow \mathbb{R}_+$  is a real univariate function, called the *shrinkage function*, allowed to depend on the observed data through  $S_T$ . We further assume that, as  $T$  tends to infinity,  $\hat{\phi}_T$  converges to a non-stochastic limiting shrinkage function  $\phi$ . The goal is to find the ‘optimal’  $\phi$ .

In order to accomplish this goal, we had to invoke a heavy-duty machinery from a research field called *random matrix theory* (RMT), which dates back to the seminal works of [Wigner \(1955\)](#) and [Marčenko and Pastur \(1967\)](#). This field studies large-sample, or asymptotic, properties of various features of sample covariance matrices, with a major focus on the sample eigenvalues. Under a rather lengthy set of regularity conditions, it can be shown that the sample eigenvalues are non-stochastic in the limit. More particularly, as the dimension  $N$  and the sample size  $T$  tend to infinity together, with

their concentration ratio  $N/T$  converging to a limit  $c \in (0, 1) \cup (1, \infty)$ ,<sup>9</sup> the empirical distribution of the sample eigenvalues converges almost surely to non-stochastic limit distribution  $F$ , which is guaranteed to be continuous. This limit distribution is completely characterized by two inputs only: first, the limiting concentration ratio  $c$  and, second, the limiting distribution of the population eigenvalues, commonly called  $H$ , whose existence (and certain properties thereof) are part of the assumed set of regularity conditions; for example, see [Ledoit and Wolf \(2015, Section 2.1\)](#) for a detailed listing of this set. This characterization is known as the *fundamental equation* of random matrix theory, originally due to [Marčenko and Pastur \(1967\)](#) and restated in various alternative expressions over the years by other authors, such as [Silverstein and Bai \(1995\)](#) and [Silverstein \(1995\)](#).

A key step in finding the optimal limiting shrinkage function  $\phi$  was the realization that the loss function

$$\mathcal{L}_F(\hat{\Sigma}_T, \Sigma_T) := \|\hat{\Sigma}_T - \Sigma_T\|_F^2$$

is non-stochastic in the limit for estimators  $\hat{\Sigma}_T$  in the class (4.7). That is, under the assumed set of regularity conditions,  $\mathcal{L}_F(\hat{\Sigma}_T, \Sigma_T)$  converges almost surely to a non-stochastic limiting expression, as  $T$  goes to infinity.<sup>10</sup> This limiting expression, as to be expected, depends on the limiting shrinkage function  $\phi$  and can be minimized with respect to it. The corresponding minimizer is then the optimal  $\phi$ , denoted by  $\phi^\circ$ , where the subscript o stands for “oracle”. Over the years, we have established different formulas for  $\phi^\circ$ ; they are all equivalent and just look different, since they are based on different mathematical tools, respectively, ingredients. For the sake of space, we shall limit ourselves here to the formula given in [Ledoit and Wolf \(2018a\)](#):

$$\phi^\circ(x) := \begin{cases} \frac{1}{\pi(c-1)\mathcal{H}_{\underline{f}}(0)} & \text{if } x = 0 \text{ and } c > 1 \\ \frac{x}{[\pi c x f(x)]^2 + [1 - c - \pi c x \mathcal{H}_f(x)]^2} & \text{otherwise} \end{cases} \quad (4.8)$$

There are three ingredients in this formula that need to be explained:  $f$ ,  $\underline{f}$ , and the operator  $\mathcal{H}$ . First,  $f$  denotes the density of  $F$ , the (continuous) limiting distribution of the sample eigenvalues; second,  $\underline{f}$  is the density of  $\underline{F} := (1 - c) \mathbb{1}_{[0, +\infty)} + cF$ ; third, for a

---

<sup>9</sup>For certain technical reasons, the value  $c = 1$  is ruled out in many relevant results from RMT. But Monte Carlo studies show that our nonlinear shrinkage estimator also works well in practice for scenarios with  $N/T = 1$ .

<sup>10</sup>In finite-sample analysis, it is common to go from the loss to the risk function, by taking expectations, and thereby to move from a stochastic expression to a non-stochastic expression that is to be minimized. But under large-dimensional RMT asymptotics this is not necessary, since the loss is already non-stochastic in the limit; in other words, in the limit, the loss is equal to the risk.

real function  $g$ ,  $\mathcal{H}_g$  denotes its *Hilbert transform*, defined as

$$\forall x \in \mathbb{R} \quad \mathcal{H}_g(x) := \frac{1}{\pi} PV \int_{-\infty}^{+\infty} g(t) \frac{dt}{t-x} . \quad (4.9)$$

Here,  $PV$  denotes the *Cauchy Principal Value*, which is used to evaluate the singular integral in the following way:

$$PV \int_{-\infty}^{+\infty} g(t) \frac{dt}{t-x} := \lim_{\varepsilon \rightarrow 0^+} \left[ \int_{-\infty}^{x-\varepsilon} g(t) \frac{dt}{t-x} + \int_{x+\varepsilon}^{+\infty} g(t) \frac{dt}{t-x} \right] . \quad (4.10)$$

Recourse to the Cauchy Principal Value is needed because the Cauchy kernel is singular, as a consequence of which the integral does not converge in the usual sense.

Equation (4.8), in a different mathematical expression, was first discovered by [Ledoit and P  ch   \(2011, Theorem 3\)](#), based on a generalization of the fundamental equation of random matrix theory. The formula here is the first one expressed without any reference to complex numbers; previous (mathematically equivalent) formulas used the complex-valued Stieltjes transform instead of the Hilbert transform; for example, see [Ledoit and Wolf \(2015, Equation \(3.6\)\)](#).

The corresponding oracle ‘estimator’ of  $\Sigma_T$  is then given by

$$\Sigma_T^o := U_T \Delta_T^o U_T' \quad \text{with} \quad \Delta_T^o := \text{Diag}(\phi^o(\lambda_{T,1}), \dots, \phi^o(\lambda_{T,N})) , \quad (4.11)$$

where the quantities  $(\phi^o(\lambda_{T,1}), \dots, \phi^o(\lambda_{T,N}))$  represent large-dimensional asymptotic counterparts to the finite-sample optimal quantities  $(u_{T,1}' \Sigma_T u_{T,1}, \dots, u_{T,N}' \Sigma_T u_{T,N})$  of Equation (4.5).

What have we gained by moving from the finite-sample optimal estimator  $\hat{\Sigma}_T^*$  to the oracle estimator  $\hat{\Sigma}_T^o$ ? Nothing in the sense of feasibility, since the oracle estimator is also infeasible in practice, as it also depends on unknown population quantities, such as the density  $f$ . But unlike  $\hat{\Sigma}_T^*$ , the oracle estimator serves as useful starting point for deriving a feasible estimator; this is because it is possible to consistently estimate the oracle shrinkage function  $\phi^o$  of Equation (4.8).

Indeed, a consistent estimator of  $c$ , by definition, is given by  $\hat{c}_T := N/T$ . This leaves us with the task of finding consistent<sup>11</sup> estimators of  $f$ ,  $\underline{f}$ , and their two respective Hilbert transforms. In [Ledoit and Wolf \(2018a\)](#), we detail how this task can be accomplished by means of kernel estimation.<sup>12</sup> The feasible nonlinear shrinkage estimator of  $\Sigma_T$  is then given by

$$S_T^o := U_T \hat{\Delta}_T^o U_T' \quad \text{with} \quad \hat{\Delta}_T^o := \text{Diag}(\hat{\phi}_T^o(\lambda_{T,1}), \dots, \hat{\phi}_T^o(\lambda_{T,N})) , \quad (4.12)$$

<sup>11</sup>As a technical detail, these estimators need to be *uniformly* consistent.

<sup>12</sup> This idea goes back to [Jing et al. \(2010\)](#), although they only considered estimation of  $f$ .

where

$$\hat{\phi}_T^o(x) := \begin{cases} \frac{1}{\pi(\hat{c}_T - 1)\hat{\mathcal{H}}_{T,f}(0)} & \text{if } x = 0 \text{ and } \hat{c}_T > 1 \\ \frac{x}{[\pi\hat{c}_Tx\hat{f}_T(x)]^2 + [1 - \hat{c}_T - \pi\hat{c}_Tx\hat{\mathcal{H}}_{T,f}(x)]^2} & \text{otherwise} \end{cases} \quad (4.13)$$

Here,  $\hat{\mathcal{H}}_{T,f}$  denotes the kernel estimator of  $\mathcal{H}_f$  and analogously for  $\hat{\mathcal{H}}_{T,f}$ .

Crucially, the feasible estimator  $S_T^o$  is asymptotically just as good as the infeasible oracle  $\Sigma_T^o$ , in the sense that it also minimizes the non-stochastic limit of the loss function  $\mathcal{L}_F(\hat{\Sigma}_T, \Sigma_T)$  with respect to  $\hat{\Sigma}_T$  (in the class of rotation-equivariant estimators considered), namely,

$$\mathcal{L}_F(S_T^o, \Sigma_T) - \mathcal{L}_F(\Sigma_T^o, \Sigma_T) \xrightarrow{P} 0.$$

**Remark 4.2** (Other Estimation Strategies). The title of our paper [Ledoit and Wolf \(2018a\)](#) stems from the fact that the outlined method is the first to ‘directly’ estimate the oracle shrinkage function  $\phi^o$  based on the analytical formula (4.8).

In the earlier works [Ledoit and Wolf \(2012, 2015\)](#), we had proposed alternative, ‘indirect’ estimation strategies that work as follows: first, estimate  $H$ , the limiting distribution of the population eigenvalues; second, use the resulting estimator  $\hat{H}_T$  together with the estimator  $\hat{c}_T := N/T$  to estimate  $F$  via the previously mentioned fundamental equation of RMT; third, use the resulting estimator  $\hat{F}_T$  to back out estimators of the various features of  $F$  that appear in the alternative formulas for  $\phi^o$  used in those previous papers. The most demanding step in practice is the first one, the estimation of  $H$ , since it involves a large-dimensional optimization problem that has no analytical solution and must be solved in practice by large-scale optimization software. Therefore, these alternative strategies can also be characterized as numerical strategies.

Such numerical strategies work well in practice but they are cumbersome to implement<sup>13</sup> and take a lot of computational time. On the other hand, the analytical strategy is very easy to implement (in 20+ lines of Matlab code) and, basically, as fast as linear shrinkage. In addition, and importantly in the age of Big Data, the analytical strategy can easily handle dimensions of  $N = 10,000$  and more, whereas the numerical strategies cannot handle dimensions much larger than  $N = 1,000$ .

A completely different strategy to consistently estimate the oracle, based on repeated sample splitting, has been suggested by [Abadir et al. \(2014\)](#) and [Lam \(2016\)](#); this strategy is of completely numerical nature and bypasses estimation of the oracle shrinkage function  $\phi^o$  entirely. Like our analytical strategy, it is easy to implement, but like our numerical strategies, it takes a lot of computational time and cannot handle dimensions much larger than  $N = 1,000$ . ■

<sup>13</sup>In fact, we had to write a separate paper to detail the implementation; see [Ledoit and Wolf \(2017b\)](#).



It will be useful to compare nonlinear shrinkage to generic linear shrinkage, that is, linear shrinkage to (a multiple of) the identity matrix. The essential distinction is that linear shrinkage is a global operator: all sample eigenvalues are moved *towards* their grand mean, with common intensity; on the other hand, nonlinear shrinkage is a local operator: some of the sample eigenvalues might be moved *away* from their grand mean, towards local ‘centers of attraction’.<sup>14</sup> Figure 1 provides a graphical illustration for these contrasting behaviors; in order to eliminate ‘noise differences’ and to focus only on ‘systematic differences’, both  $N$  and  $T$  have been chosen rather large.

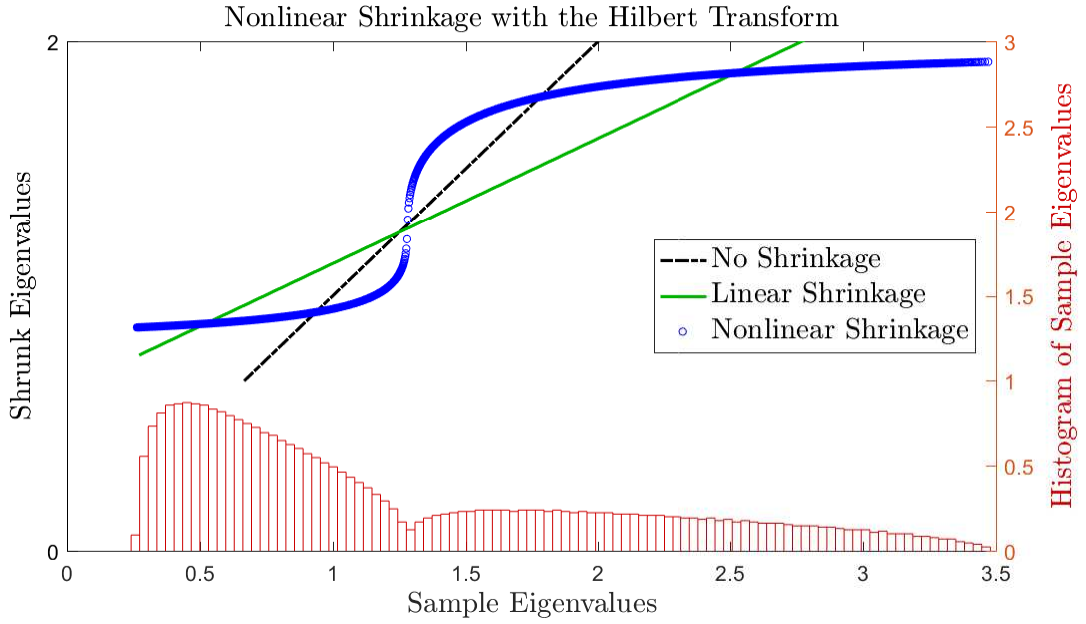


Figure 1: Local Attraction Effect. 2,500 population eigenvalues are equal to 0.8, and 1,500 are equal to 2, so that  $N = 4,000$ . The sample size is  $T = 18,000$ . At the bottom of the figure is a histogram displaying the location of the sample eigenvalues.

In this example, the average eigenvalue is equal to 1.25. Sample eigenvalues below the average but above 1 need to be ‘shrunk’ downwards because they are attracted by the cluster to their immediate left; this is because these sample eigenvalues generally correspond to a population eigenvalue of 1. Similarly, sample eigenvalues above the average but below 1.75 need to be ‘shrunk’ upwards because they are attracted by the cluster to their immediate right; this is because these sample eigenvalues generally correspond to a population eigenvalue of 2. Linear shrinkage, being a global operator, is not equipped to sense a disturbance in the force: It applies the same shrinkage intensity across the board and shrinks all sample eigenvalues towards the grand mean of 1.25.

<sup>14</sup>Nevertheless, nonlinear shrinkage still reduces the overall spread compared to the sample eigenvalues.



### 4.3 Simulation Evidence

Over the years, we have proposed three different strategies to carry out nonlinear shrinkage: the two numerical (or indirect) strategies of [Ledoit and Wolf \(2012, 2015\)](#) and the analytical (or direct) strategy of [Ledoit and Wolf \(2018a\)](#). Of the two numerical strategies, the first one should by now be considered obsolete for two reasons: first, it can only handle the case  $N < T$  and, second, it is somewhat less stable than the second strategy, which we coined QuEST (acronym for *quantized eigenvalues sampling transform*).

Extensive Monte Carlo studies in [Ledoit and Wolf \(2012, 2015\)](#) revealed that nonlinear shrinkage generally outperforms linear shrinkage, and often by a large margin. There are two exceptions of different nature. The first exception is the case when linear shrinkage is already optimal, that is, when the oracle shrinkage function  $\phi^\circ$  is (nearly) a linear function.<sup>15</sup> This is as to be expected: ‘fitting’ a nonlinear function to a linear relationship cannot perform as well as fitting a linear function, which is what linear shrinkage does. The good news is that the performance difference is generally negligible, unless also the second exception holds true. This second exception is the case when  $N$  and  $T$  are ‘not large’. This is as to be expected as well: successfully fitting a nonlinear function requires a certain amount of data. It is impossible to give a perfect rule in this regard, since the performance difference also depends on the distribution of the population eigenvalues; but as a rule of thumb, one generally needs both  $N$  and  $T$  to be above 100 for nonlinear shrinkage to outperform linear shrinkage when  $\phi^\circ$  is ‘markedly’ nonlinear, on the one hand, and for nonlinear shrinkage to be, basically, as good as linear shrinkage when  $\phi^\circ$  is (nearly) linear. In other words, when both  $N$  and  $T$  are larger than 100, then, as a rule of thumb, there is basically nothing to lose but potentially a lot to gain by upgrading from linear to nonlinear shrinkage. Therefore, for large data sets, nonlinear shrinkage should become the new ‘generic’ estimator of a covariance matrix.

Nevertheless, until recently, there were two reasons why applied researchers might have shied away from using nonlinear shrinkage. First, QuEST is a highly complex strategy and the corresponding code far from easy to understand and digest; for most people, it is like using a black box, which might make them uncomfortable. Second, the method is slow to run and cannot handle dimensions much above  $N = 1,000$ , which is a limitation in the age of Big Data. We fixed both problems with the analytical strategy of [Ledoit and Wolf \(2018a\)](#): first, the code is short, 20+ lines of Matlab code, and easy to understand and verify; second, the strategy is, basically, as fast as linear shrinkage. In terms of accuracy, extensive Monte Carlo studies in [Ledoit and Wolf \(2018a\)](#) showed

---

<sup>15</sup>For example, this is the case if the population eigenvalues are (nearly) identical.

that the analytical strategy is, for all practical purposes, as accurate as the numerical QuEST strategy. So, finally, there is nonlinear shrinkage for the (educated) masses!

## 4.4 Applications

Nonlinear shrinkage has not been around as long as generic linear shrinkage and thus, unsurprisingly, it has not been applied as widely yet. Nevertheless, we can list some examples already and trust that many more will follow in the future, especially given the recent addition of the analytical strategy to our toolbox, as discussed above:

**Biomedicine** Searching for genetic effects on the brain (Chi et al., 2013).

**Computer Vision** 3D Human pose tracking (Simo-Serra et al., 2017) and multimodal emotion recognition (Fu et al., 2017).

**Cosmology** Extraction of cosmological information from clustering measurements of galaxy surveys (Vakili et al., 2017).

**Econometrics** Improved generalized estimating equations (Kwon et al., 2017).

**Image Recognition** Person re-identification (An et al., 2015) and kinship verification (Chen et al., 2017).

**Neural Networks** Mitigating catastrophic forgetting (Lancewicki et al., 2015).

**Operations Research** Continuous optimization via covariance matrix adaptation in evolution strategies (Meyer-Nieberg and Kropat, 2015).

**Signal Processing** Effective numerical algorithms for distributed information processing (Grant et al., 2014).

A major application of interest to us was, again, Markowitz portfolio selection. Needless to say, one can argue that the generic (scaled) Frobenius loss function  $\mathcal{L}_F$  of (2.2) may not be ideally suited for this specific problem. The good news is that another loss function that is custom-tailored to the problem of portfolio selection yields *the same* oracle shrinkage function  $\phi^\circ$  of (4.8) and, therefore, also *the same* feasible nonlinear shrinkage estimator  $S_T^\circ$  of (4.12).

This loss function, proposed by Engle et al. (2019, Definition 1) and called the *minimum variance* (MV) loss (function) is given by

$$\mathcal{L}_{\text{MV}}(\hat{\Sigma}_T, \Sigma_T) := \frac{\text{Tr}(\hat{\Sigma}_T^{-1} \Sigma_T \hat{\Sigma}_T^{-1})/N}{[\text{Tr}(\hat{\Sigma}_T^{-1})/N]^2} - \frac{1}{\text{Tr}(\Sigma_T^{-1})/N} . \quad (4.14)$$

Roughly speaking,  $\mathcal{L}_{\text{MV}}$  represents the *true* variance of the linear combination of the original variables that has the minimum *estimated* variance, under a generic linear constraint, after suitable normalization. Further justification for the minimum variance loss function is given in Engle and Colacito (2006) and Ledoit and Wolf (2017a).

**Remark 4.3** (Related Loss Function). In Ledoit and Wolf (2017a, Definition 1) we used a related loss function that actually depended on a given signal  $m_T$ , that is, a given estimator of the vector of expected returns of the assets in the underlying investment universe. For the mathematical analysis, we then needed to make certain distributional assumptions on the signal  $m_T$ . But, again, we recovered *the same* oracle shrinkage function  $\phi^\circ$  of (4.8) and, therefore, also *the same* feasible nonlinear shrinkage estimator  $S_T^\circ$  of (4.12). ■

In Ledoit and Wolf (2017a), we compared the nonlinear shrinkage estimator  $S_T^\circ$  to a variety of other covariance matrix estimators from the literature for the purpose of Markowitz portfolio selection, using backtest exercises with real-life stock return data; we used both daily and monthly return data. We also included some other strategies that do not rely on a (sophisticated) estimator of the covariance matrix, such as the equal-weighted portfolio or a norm-constrained portfolio. Just like the nonlinear shrinkage estimator, also all the other strategies are based on the assumption that return data are i.i.d. over time. In our analysis, nonlinear shrinkage was the clear overall winner and, in particular, dominated generic linear shrinkage.

## 4.5 Different Oracle Shrinkage Formulas

We have seen that the oracle shrinkage formula  $\phi^\circ$  of (4.8) holds for two different loss functions: the Frobenius loss  $\mathcal{L}_F$  of (2.2) and the minimum variance loss  $\mathcal{L}_{MV}$  of (4.14). To allow for a pun, this coincidence is somewhat of a coincidence: We certainly did not expect that these two different loss functions would lead to the same oracle shrinkage formula.

Another loss function, which is important for historic reasons but perhaps less relevant from a perspective of applications, is *Stein's* loss defined as

$$\mathcal{L}_{St}(\hat{\Sigma}_T, \Sigma_T) := \frac{1}{N} \text{Tr}(\Sigma_T^{-1} \hat{\Sigma}_T) - \frac{1}{N} \log[\det(\Sigma_T^{-1} \hat{\Sigma}_T)] - 1 . \quad (4.15)$$

Yet another loss function one can consider is *symmetrized* Stein's loss, defined as

$$\mathcal{L}_{S-St}(\hat{\Sigma}_T, \Sigma_T) := \frac{1}{2N} \text{Tr}(\Sigma_T^{-1} \hat{\Sigma}_T + \Sigma_T \hat{\Sigma}_T^{-1}) - 1 . \quad (4.16)$$

These two loss functions yield oracle shrinkage formulas different from (4.8), and also different from each other; for the interest of space, the interested reader is referred to Ledoit and Wolf (2018b) for the details.

## 5 Extension to Dynamic Models

So far, we have used the assumption that the  $T$  observations (that is, the rows of the matrix  $X_T$ ) are i.i.d. Of course, such an assumption does not necessarily hold for financial return data, at least at shorter frequencies, such as at the daily frequency. It is, therefore, of interest to (try to) use a model that allows for a time-changing nature of the conditional covariance matrix. In other words, it is of interest to use a *dynamic* model instead of a *static* model. Arguably, the most popular class of such dynamic models are multivariate GARCH models; for example, see [Bauwens et al. \(2006\)](#).

Unfortunately, these models suffer from the curse of dimensionality: Since they can be quite complex and contain a large number of parameters, they do not work well, or cannot even be estimated to begin with, when the number of assets is large, such as  $N = 1,000$ ; indeed, in basically all prior applications, the number of assets was at most  $N = 100$  and often even in the single digits. Consequently, we wondered whether we could use (non)linear shrinkage in a suitable way to ‘robustify’ a multivariate GARCH model against large dimensions and make it work well for, at least,  $N = 1,000$  assets. The challenge was clear: All our shrinkage methods had been designed for a static setting, where the conditional covariance matrix does not change over time; so how could we possibly use them gainfully in a dynamic setting? The key was to find a multivariate GARCH model that has as one of its components a large-dimensional covariance matrix that is static (or time-invariant). A suitable model turned out to be the *dynamic conditional correlation* (DCC) model of [Engle \(2002\)](#); more precisely, a certain version of this model based on *correlation targeting*. Some notation is needed first:

- $r_{t,i}$ : observed return for asset  $i$  at date  $t$ , stacked into  $r_t := (r_{t,1}, \dots, r_{t,N})'$
- $d_{t,i}^2 := \text{Var}(r_{t,i} | \mathcal{F}_{t-1})$ : conditional variance of the  $i^{\text{th}}$  return at date  $t$
- $D_t$  is the  $N$ -dimensional diagonal matrix whose  $i^{\text{th}}$  diagonal element is  $d_{t,i}$
- $H_t := \text{Cov}(r_t | \mathcal{F}_{t-1})$ : conditional covariance matrix at date  $t$ ; thus  $\text{Diag}(H_t) = D_t^2$
- $s_{t,i} := r_{t,i}/d_{t,i}$ : devolatilized return, stacked into  $s_t := (s_{t,1}, \dots, s_{t,N})'$
- $R_t := \text{Cor}(r_t | \mathcal{F}_{t-1}) = \text{Cov}(s_t | \mathcal{F}_{t-1})$ : conditional correlation matrix at date  $t$
- $C := \mathbb{E}(R_t) = \text{Cor}(r_t) = \text{Cov}(s_t)$ : unconditional correlation matrix

Here,  $\mathcal{F}_{t-1}$  denotes the information set available at time  $t - 1$ ; also, it is tacitly assumed that the model yields a stationary return series  $\{r_t\}$ , so that unconditional moments, such as  $\mathbb{E}(R_t)$ , are time-invariant.

The key part of this model is the evolution of the conditional pseudo correlation matrix of the asset returns  $r_t$  over time:

$$Q_t = (1 - \alpha - \beta)C + \alpha s_{t-1} s_{t-1}' + \beta Q_{t-1} . \quad (5.1)$$

Here,  $(\alpha, \beta)$  are the DCC parameters, which are analogous to related parameters in a univariate GARCH(1,1) model, and assumed to satisfy  $0 \leq \alpha + \beta < 1$ , which is needed for a stationary model. The key feature of model (5.1) is that the implied unconditional covariance matrix of  $Q_t$  is guaranteed to be equal to  $C$ , no matter what the choices of  $(\alpha, \beta)$  are. This is just the property of correlation targeting: The model is set up in such a way that it targets the true quantity (in terms of the unconditional covariance matrix).

The matrix  $Q_t$  can be interpreted as a conditional *pseudo* correlation matrix, or a conditional covariance matrix of devolatilized residuals. It cannot be used directly because its diagonal elements, although close to one, are not exactly equal to one. From this representation, one obtains the conditional correlation matrix and the conditional covariance matrix as

$$R_t := \text{Diag}(Q_t)^{-1/2} Q_t \text{Diag}(Q_t)^{-1/2} , \quad (5.2)$$

$$H_t := D_t R_t D_t , \quad (5.3)$$

and the data-generating process is driven by the multivariate normal law<sup>16</sup>

$$r_t | \mathcal{F}_{t-1} \sim \mathcal{N}(0, H_t) . \quad (5.4)$$

How can model (5.1) be estimated in practice? In the first step, one needs to estimate the vector  $s_{t-1}$ , which is done by dividing the individual asset returns at time  $t-1$  by their *estimated* conditional standard deviations,<sup>17</sup> resulting in a vector  $\hat{s}_{t-1}$ . In the second step, one needs to estimate the *correlation targeting* matrix  $C$ . In the third step, one estimates  $(\alpha, \beta)$  based on the feasible relation

$$Q_t = (1 - \alpha - \beta) \hat{C} + \alpha \hat{s}_{t-1} \hat{s}_{t-1}' + \beta Q_{t-1} , \quad (5.5)$$

using maximum likelihood. The curse of dimensionality arises in both in step 2 and in step 3.

Shrinkage estimation helps with step 2. The problem here is that  $C$  is a  $N \times N$  matrix, which can be large-dimensional. The original proposal of Engle (2002) was to use the sample correlation matrix of the ‘feasible’ devolatilized return series  $\{\hat{s}_t\}$ ; this approach works well for dimensions  $N \leq 100$ , but not for dimensions  $N = 1,000$  and above. A superior approach in large dimensions is to use a ‘generic’ shrinkage estimator on the return series  $\{\hat{s}_t\}$ ; linear shrinkage to (a multiple of) the identity already works well but

---

<sup>16</sup>One can also try to model the time-varying conditional mean of  $r_t$  instead of setting it equal to zero, but doing so makes virtually no difference in practice in terms of estimating  $(\alpha, \beta)$ .

<sup>17</sup>The methodology is flexible in this regard, the most popular approach being to use individual GARCH(1,1) models, one model for each asset.

nonlinear shrinkage works even better. Note that the resulting estimator needs to be post-processed along the lines of (5.2) to produce a proper correlation matrix  $\hat{C}$ .

Having estimators  $\hat{s}_{t-1}$  and  $\hat{C}$ , one can now estimate the DCC parameters  $(\alpha, \beta)$  based on the feasible relation (5.5). The natural approach is to use maximum likelihood, based on assumption (5.4). This becomes a computational problem in large dimensions, since full (or exact) maximum likelihood cannot be carried out with current computational power for dimensions much above  $N = 100$ . The solution of Pakel et al. (2017) is to use a *composite likelihood* instead which, in a nutshell, combines likelihoods over many small subsets of the data (such as neighboring pairs of assets) in one ‘joint’ likelihood; doing so results in a likelihood function that can be maximized with respect to  $(\alpha, \beta)$  in a manner that is computationally feasible.

The resulting DCC-NL estimation strategy of Engle et al. (2019) unfolds in a three-stage process:

1. For each asset, fit a univariate GARCH(1,1) model and use the fitted models to devolatilize the return series  $\{r_t\}$  to obtain the series  $\{\hat{s}_t\}$ .
2. Estimate the unconditional correlation matrix  $C$  by applying nonlinear shrinkage (with post-processing) to the series  $\{\hat{s}_t\}$  and use the resulting estimator  $\hat{C}$  for correlation targeting.
3. Maximize the composite likelihood (over all neighboring pairs of assets) to estimate the correlation-dynamics parameters  $(\alpha, \beta)$ .

In Engle et al. (2019), we studied the properties of the DCC-NL estimator when used for the purpose of Markowitz portfolio selection, using backtest exercises with real-life stock return data; we used daily return data only. We could not compare DCC-NL to other multivariate GARCH models from the literature, since none of them (currently) work for dimensions well above  $N = 100$ . But we did compare DCC-NL to DCC-S, which uses the sample covariance matrix of the series  $\{\hat{s}_t\}$  to estimate  $C$ , and to DCC-Lin, which uses linear shrinkage applied to the series  $\{\hat{s}_t\}$ . We found that for  $N = 100$ , all three methods performed about equally well but that for  $N = 500$  and  $N = 1000$ , DCC-Lin and DCC-NL outperformed DCC-S by a considerable margin, with DCC-NL being the clear winner; more specifically, the improvement of DCC-NL over DCC-Lin was of the same magnitude as the improvement of DCC-Lin over DCC-S.<sup>18</sup>

As a further application, in Ledoit et al. (2019), we showed how to use the DCC-NL estimator to construct more powerful tests for cross-sectional anomalies, that is, more

---

<sup>18</sup>We also included the (second-generation) RiskMetrics 2006 model in the study, which also estimates time-varying covariance matrices and is computationally feasible for dimensions up to  $N = 1,000$ ; see Zumbach (2007). Unfortunately, this model performed poorly and cannot be recommended for the purpose of portfolio selection.

powerful tests to establish the validity of a so-called return anomaly (also called factor or return-predictive signal) whose goal it is to explain the cross-section of expected stock returns. Traditional tests construct dollar-neutral long-short portfolios that load on the return anomaly under study by sorting the stocks into quantiles according to their anomaly scores; if such a zero-cost portfolio can be shown to deliver a positive expected return with statistical significance, the anomaly under study is established as ‘successful’ or ‘for real’. The problem is that such quantile-based sorting portfolios completely ignore the covariance matrix of the stock returns. More efficient dollar-neutral long-short portfolio can be constructed by incorporating an accurate estimator of this covariance matrix in the spirit of Markowitz-type portfolio selection; in practice, we propose the use of the DCC-NL estimator to this end. In an empirical analysis using 60+ suggested return anomalies from the literature, we showed that using such “efficient sorting” portfolios yields much more powerful tests compared to the *status quo* of portfolios based on sorting into quantiles.

## 6 Extension to Factor Models

Factor models have a long history in finance, with wide range of applications both in theory and in practice. Examples of theoretical applications are asset pricing models, such as the Capital Asset Pricing Model (CAPM) and the Arbitrage Pricing Theory (APT) of [Ross \(1976\)](#), and various fund-separation theorems. In practice, factor models are used, among others, to evaluate the performance of portfolio managers, to assess return anomalies, to predict returns, and to construct portfolios; for example, see [Meucci \(2005\)](#) and [Chincarini and Kim \(2006\)](#).

It is, therefore, also of interest to (try to) use factor models to estimate the covariance matrix of a large universe of asset returns, in particular stock returns. We have already described one such approach, namely linear shrinkage to the single-factor model; see [Section 3](#). But this approach would be hard to extend to using more than one factor or to a dynamic setting where the conditional covariance matrix varies over time. A more flexible approach, explored in [De Nard et al. \(2019\)](#), is to use shrinkage estimation for the residual covariance matrix of a general factor model.

There are many different ‘versions’ of factor models: factors can be observed or latent; factors loadings can be fixed or vary over time; factor models can be exact or approximate; and the conditional covariance matrix of the residuals can be fixed or vary over time. We do not have the space here to describe all versions in detail, which was done in [De Nard et al. \(2019\)](#), and so we shall focus on the two models that were the most promising in the end.



The basic model assumption is that, for every asset  $i = 1, \dots, N$ ,

$$r_{t,i} = \alpha_i + \beta_i' f_t + u_{t,i} , \quad (6.1)$$

where  $f_t := (f_{t,1}, \dots, f_{t,K})'$  is a vector of returns on  $K$  observed (or explicit) factors,  $\beta_i := (\beta_{i,1}, \dots, \beta_{i,K})'$  is a vector of time-invariant factor loadings, and  $u_{t,i}$  is an error term that satisfies  $\mathbb{E}(u_{t,i}|f_t) = 0$ .

Let  $\Sigma_{f,t} := \text{Cov}(f_t|\mathcal{F}_{t-1})$  and  $\Sigma_{u,t} := \text{Cov}(u_t|\mathcal{F}_{t-1})$ , with  $u_t := (u_{t,1}, \dots, u_{t,N})'$ . Our models assume  $\Sigma_{f,t} \equiv \Sigma_f$ , that is, the conditional covariance matrix of the factor returns is time-invariant; on the other hand, regarding  $\Sigma_{u,t}$  we allow for both the time-invariant (or static) and the time-varying (or dynamic) case.<sup>19</sup>

Using again the notation  $H_t := \text{Cov}(r_t|\mathcal{F}_{t-1})$ , and denoting by  $B$  the  $K \times N$  matrix whose  $i^{\text{th}}$  column is the vector  $\beta_i$ , model (6.1) in conjunction with our stated assumption implies that

$$H_t = B' \Sigma_f B + \Sigma_{u,t} , \quad (6.2)$$

which specializes to

$$H_t \equiv H := B' \Sigma_f B + \Sigma_u \quad (6.3)$$

under the additional assumption  $\Sigma_{u,t} \equiv \Sigma_u$ .

In practice, we need to estimate three ingredients:  $B$ ,  $\Sigma_f$ , and  $\Sigma_{u,t}$ , respectively,  $\Sigma_u$ .

For every asset  $i = 1, \dots, N$ , we estimate model (6.1) by a time-series regression using OLS, resulting in estimators  $\hat{\alpha}_i$  and  $\hat{\beta}_i$  and in residuals  $\{\hat{u}_{t,i}\}$ . The estimator of  $B$ , denoted by  $\hat{B}$ , is then the  $K \times N$  matrix whose  $i^{\text{th}}$  column is the vector  $\hat{\beta}_i$ .

The estimator of  $\Sigma_f$ , denoted by  $\hat{\Sigma}_f$ , is the sample covariance matrix of the factor returns  $\{f_t\}$ . We only consider factor models where  $K$  is in the single digits; therefore, there is nothing to gain here by using a more sophisticated shrinkage estimator instead.

Let  $\hat{u}_t := (\hat{u}_{t,1}, \dots, \hat{u}_{t,N})'$ . In general, we estimate  $\Sigma_{u,t}$  by applying DCC-NL to the residuals  $\{\hat{u}_t\}$  and denote the estimator by  $\hat{\Sigma}_{u,t}$ ; in the specialized case  $\Sigma_{u,t} \equiv \Sigma_u$ , we estimate  $\Sigma_u$  by applying nonlinear shrinkage to the residuals  $\{\hat{u}_t\}$  and denote the estimator by  $\hat{\Sigma}_u$ .

Doing so results in the estimator

$$\hat{H}_t := \hat{B}' \hat{\Sigma}_f \hat{B} + \hat{\Sigma}_{u,t} , \quad (6.4)$$

which specializes to

$$\hat{H}_t \equiv \hat{H} := \hat{B}' \hat{\Sigma}_f \hat{B} + \hat{\Sigma}_u \quad (6.5)$$

under the additional assumption  $\Sigma_{u,t} \equiv \Sigma_u$ .

---

<sup>19</sup>We also tried models with  $\Sigma_{f,t}$  time-varying but the performance was not better.



**Remark 6.1** (Comparison with More Traditional Estimators). For the special case (6.3), we can compare our estimator (6.5) with more traditional approaches. The earliest approach in the literature was to assume an *exact* factor model (EFM), which corresponds to assuming that  $\Sigma_u$  is a diagonal matrix. In this setting, it is customary to take  $\hat{\Sigma}_u$  as the diagonal matrix based on the sample covariance matrix of the residuals  $\{\hat{u}_t\}$ , that is, start with sample covariance matrix and then set all off-diagonal element to zero. The problem with this approach is that the assumption of an exact factor model is often violated in practice and thus the resulting estimator  $\hat{\Sigma}_f$  can suffer from severe biases, resulting in unsatisfactory performance.

In this day and age, the assumption of an *approximate* factor model (AFM) is more common, which corresponds to assuming that  $\Sigma_u$  is a sparse matrix. In this setting, one obtains an estimator of  $\Sigma_u$  by applying some sort of thresholding scheme to the sample covariance matrix of the residuals  $\{\hat{u}_t\}$ ; for example, this is the underlying idea of the popular POET covariance matrix estimator of Fan et al. (2013).<sup>20</sup> Whether such an approach works better in practice than using nonlinear shrinkage on the residuals  $\{\hat{u}_t\}$  is really an empirical question in the end. ■

In De Nard et al. (2019), we studied the properties of a variety of covariance matrix estimators when used for the purpose of Markowitz portfolio selection, using backtest exercises with real-life stock return data; we used daily return data only. Broadly speaking, these estimators can be categorized into a two-by-two table: static versus dynamic estimators<sup>21</sup> and structure-free versus factor-model-based estimators. Most empirical studies only compare estimators from one of the four categories; to the best of our knowledge, our paper is the first one that includes (sophisticated) estimators from all four categories. We found that dynamic estimators performed better than static estimators; in particular, the best structure-free estimator was DCC-NL. In terms of factor models, approximate factor models unsurprisingly performed better than exact factor models. The overall best estimator, which we call AFM-DCC-NL, was estimator (6.4). A recurring problem with factor models is how many (and which) factors to include, that is, the proper choice of  $K$ . We restricted attention to the five factors of Fama and French (2015), so the nature of the factors was given. As to be expected, an exact factor model worked better when using all five factors as opposed to using the first factor (that is, the market factor) only. An important advantage of the AFM-DCC-NL estimator is that it worked just well when using the first factor only as when using all five factors. Apparently,

---

<sup>20</sup>Note that the POET estimator is based on unobserved (or latent) factors, which are estimated by PCA in practice, rather than on observed factors, such as Fama-French factors.

<sup>21</sup>A static estimator assumes the conditional covariance matrix is time-invariant whereas a dynamic estimator assumes that it is time-varying.

DCC-NL is able to recover ‘left-over’ factor structure in the residual covariance matrix in an automated way, and only the (dominant) market factor needs to be accounted for explicitly.<sup>22</sup> This is good news, in particular for managers who want to invest outside of the US, where the extra four Fama-French factors are not always available and the market factor can be easily constructed if need be.

Concerning the final sentence of Remark 6.1, both POET and estimator (6.5), which we call AFM-NL, are static estimators. Using  $K = 5$  factors, where the factors for POET are the first five principal components of the sample covariance matrix of the stock returns and the factors for AFM-NL are the five Fama-French factors, the performance of AFM-NL is somewhat better. The interesting find is that, similar to AFM-DCC-NL, the performance of AFM-NL is just as good when using the single market factor only, whereas the performance of POET is known to suffer if not enough factors are included.<sup>23</sup>

Therefore, one can say that both AFM-NL and AFM-DCC-NL are robust to the number of factors chosen, as long as the market factor is included, which is an unusual property of factor models.

## 7 Computational Aspects & Code

To simplify the notation, we have assumed so that all variables have mean zero. In this way, the sample covariance matrix is given by

$$S_T := \frac{1}{T} X_T' X_T .$$

In many applications, variables do not have mean zero, or at least it is not known whether they have mean zero. In such a setting, it is more common to base the sample covariance matrix on the demeaned data instead. Denote by  $x_{\cdot i}^T$  the  $i^{\text{th}}$  column of the matrix  $X_T$  and by  $\bar{x}_{\cdot i}^T := T^{-1} \sum_t x_{ti}^T$  its mean. Further denote by  $Y_T$  the  $T \times N$  matrix with typical element  $y_{ti}^T := x_{ti}^T - \bar{x}_{\cdot i}^T$ . The sample covariance matrix based on the demeaned data is given by

$$\tilde{S}_T := \frac{1}{T-1} Y_T' Y_T .$$

One then simply replaces  $S_T$  with  $\tilde{S}_T$  and  $T$  with  $T-1$  in all the previous descriptions and computations in practice.

Another way to look at this issue is from a coding perspective. In any code, there needs to be a variable for “sample covariance matrix”; in the case of demeaning, this

---

<sup>22</sup>Indeed, AFM-DCC-NL performs better than the completely structure-free DCC-NL.

<sup>23</sup>We only tried using  $K = 5$  principals components for POET in our empirical analysis; following the recommendations of Fan et al. (2013), we did not try using (the first)  $K = 1$  principal component only.

variable needs to be assigned the value  $\tilde{S}_T$  and otherwise the value  $S_T$ . Also, there needs to be variable for “sample size”; in the case of demeaning, this variable needs to be assigned the value  $T - 1$  and otherwise the value  $T$ . The importance of this latter adjustment of the ‘effective’ sample size in the case of demeaning becomes especially clear when  $N > T$ . In this scenario, the number of zero sample eigenvalues is equal to  $N - T$  without demeaning but equal to  $N - (T - 1) = N - T + 1$  with demeaning; correctly keeping track of this number is important for certain aspects of coding.

Speaking of code, our various estimators range from trivial-to-code, such as generic linear shrinkage, to super-hard-to-code, such as the QuEST version of nonlinear shrinkage. But free programming code in Matlab for most of the estimators reviewed in this paper can be downloaded at [www.econ.uzh.ch/en/people/faculty/wolf/publications.html](http://www.econ.uzh.ch/en/people/faculty/wolf/publications.html) under the header “Programming Code”. There also exists an R package `nlshrink`<sup>24</sup> which implements the QuEST version of nonlinear shrinkage; however, this package was not written by us and so we cannot necessarily vouch for it.

## 8 Conclusion

Estimation of large-dimensional covariance matrices is an important problem with applications in many applied fields, one of them being finance. With the amount of data ever increasing in the age of Big Data, this problem will only become more important over time. In this paper, we have reviewed our own work in this area stretching back 15+ years. In various shapes and forms, what we have been promoting is *shrinkage* estimation of covariance matrices.

In early work, which can be classified as *linear* shrinkage, this amounts to taking a convex linear combination of the sample covariance matrix and a target matrix; this target matrix can either be completely generic and taken to be (a multiple of) the identity matrix or it can incorporate application-specific structure, such as a factor model. At any rate, the target matrix always contains a (relatively) small number of parameters and thus little estimation error, albeit generally a bias, and thus constitutes a ‘counterpart’ to the sample covariance matrix which is unbiased but contains a large number of free parameters. Linear shrinkage works by providing a bias-variance tradeoff: optimally combining two ‘extremes’ works better than either one of them. This insight goes back to the genius of Charles Stein, who proposed such linear shrinkage estimation for the mean vector; we just adapted his line of thinking to the covariance matrix instead.

Later work considered an extension to *nonlinear* shrinkage, which does not operate on the sample covariance matrix as a whole but on its eigenvalues instead, while keeping

---

<sup>24</sup>The package can be found at [cran.r-project.org/web/packages/](http://cran.r-project.org/web/packages/).

its eigenvectors. By allowing an arbitrary, or nonlinear transformation of the sample eigenvalues, one can do much better than linear shrinkage to (a multiple of) the identity matrix. The idea of nonlinear shrinkage also goes back to Charles Stein but he lacked the mathematical machinery to solve the problem in a satisfactory fashion. This machinery is called *random matrix theory* and has been developed by a number of probability theorists and statisticians over the last 60+ years. Their collective work has enabled us to bring nonlinear shrinkage to its fruition.

In its basic form, nonlinear shrinkage is also a generic estimator that does not incorporate any particular structure. In many finance applications, certain structure is ‘known’, such as time-varying co-volatility or a factor model. In our most recent work, we have shown how to overlay such structure on nonlinear shrinkage in order to improve performance even further.

We hope that we have assembled, over time, a large and useful toolbox that will help applied researches in many fields, particularly in finance, to solve real-life problems. Let them, and time, be our judge.

## References

- Abadir, K., Distaso, W., and Žikesš, F. (2014). Design-free estimation of variance matrices. *Journal of Econometrics*, 181:165–180.
- Abrahamsson, R., Selen, Y., and Stoica, P. (2007). Enhanced covariance matrix estimators in adaptive beamforming. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP 2007*, volume II, pages 969–972.
- An, L., Yang, S., and Bhanu, B. (2015). Person re-identification by robust canonical correlation analysis. *IEEE Signal Processing Letters*, 22(8):1103–1107.
- Anlauff, J., Weitnauer, E., Lehnhardt, A., Schirmer, S., Zehe, S., and Tonekaboni, K. (2010). A method for outdoor skateboarding video games. In *Proceedings of the 7th International Conference on Advances in Computer Entertainment Technology*, pages 40–44. ACM.
- Bachega, L. R., Theiler, J., Bouman, C., et al. (2011). Evaluating and improving local hyperspectral anomaly detectors. In *Applied Imagery Pattern Recognition Workshop (AIPR), 2011 IEEE*, pages 1–8. IEEE.
- Bauwens, L., Laurent, S., and Rombouts, J. V. (2006). Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, 21(1):79–109.
- Bell, P. and King, S. (2009). Diagonal priors for full covariance speech recognition. In *IEEE Workshop on Automatic Speech Recognition & Understanding, 2009. ASRU 2009*, pages 113–117. IEEE.
- Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604.
- Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227.
- Broda, S. A. and Paoletta, M. S. (2009). CHICAGO: A fast and accurate method for portfolio risk calculation. *Journal of Financial Econometrics*, 7(4):412–436.
- Chen, X., An, L., Yang, S., and Wu, W. (2017). Kinship verification in multi-linear coherent spaces. *Multimedia Tools and Applications*, 76(3):4105–4122.
- Chi, E. C., Allen, G. I., Zhou, H., Kohannim, O., Lange, K., and Thompson, P. M. (2013). Imaging genetics via sparse canonical correlaton analysis. In *2013 IEEE 10th*

- International Symposium on Biomedical Imaging: From Nano to Macro*, pages 740–743. IEEE.
- Chincarini, L. B. and Kim, D. (2006). *Quantitative Equity Portfolio Management: An Active Approach to Portfolio Construction and Management*. McGraw-Hill, New York.
- Connor, G. and Korajczyk, R. A. (1988). Risk and return in an equilibrium APT: Application of a new test methodology. *Journal of Financial Economics*, 21(2):255–289.
- Connor, G. and Korajczyk, R. A. (1993). A test for the number of factors in an approximate factor model. *Journal of Finance*, 48:1263–1291.
- De Nard, G., Ledoit, O., and Wolf, M. (2019). Factor models for portfolio selection in large dimensions: The good, the better and the ugly. *Journal of Financial Econometrics*. Forthcoming.
- DeMiguel, V., Garlappi, L., Nogales, F. J., and Uppal, R. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812.
- Dey, D. K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein’s loss. *Annals of Statistics*, 13(4):1581–1591.
- Efron, B. and Morris, C. (1973). Stein’s estimation rule and its competitors—an Empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130.
- Efron, B. and Morris, C. (1975). Using Stein’s estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319.
- Efron, B. and Morris, C. (1977). Stein’s paradox in statistics. *Scientific American*, 236(5):119–127.
- Elsheikh, A. H., Wheeler, M. F., and Hoteit, I. (2013). An iterative stochastic ensemble method for parameter estimation of subsurface flow models. *Journal of Computational Physics*, 242:696–714.
- Endelman, J. B. and Jannink, J.-L. (2012). Shrinkage estimation of the realized relationship matrix. *G3: Genes, Genomes, Genetics*, 2(11):1405–1413.
- Engle, R. F. (2002). Dynamic conditional correlation — a simple class of multivariate GARCH models. *Journal of Business & Economics Statistics*, 20:339–350.

- Engle, R. F. and Colacito, R. (2006). Testing and valuing dynamic correlations for asset allocation. *Journal of Business & Economic Statistics*, 24(2):238–253.
- Engle, R. F., Ledoit, O., and Wolf, M. (2019). Large dynamic covariance matrices. *Journal of Business & Economic Statistics*. Forthcoming.
- Fama, E. F. and French, K. R. (2015). A five factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.
- Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Series B*, 75(4):603–680.
- Frost, P. A. and Savarino, J. E. (1986). An empirical Bayes approach to portfolio selection. *Journal of Financial and Quantitative Analysis*, 21:293–305.
- Fu, J., Mao, Q., Tu, J., and Zhan, Y. (2017). Multimodal shared features learning for emotion recognition by enhanced sparse local discriminative canonical correlation analysis. *Multimedia Systems*.
- Grant, A., Torokhti, A., and Miklavcic, S. (2014). Efficient compression of distributed information in estimation fusion. *Electronic Notes in Discrete Mathematics*, 46:297–304. Jornadas de Matemtica Discreta y Algortmica.
- Guo, S.-M., He, J., Monnier, N., Sun, G., Wohland, T., and Bathe, M. (2012). Bayesian approach to the analysis of fluorescence correlation spectroscopy data II: application to simulated and in vitro data. *Analytical Chemistry*, 84(9):3880–3888.
- Haff, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Annals of Statistics*, 8:586–597.
- Haff, L. R. (1982). Solutions of the Euler-Lagrange equations for certain multivariate normal estimation problems. Unpublished manuscript.
- Hansen, L. P. and Jagannathan, R. (1997). Assessing specification errors in stochastic discount factor models. *The Journal of Finance*, 52(2):557–590.
- Haufe, S., Treder, M. S., Gugler, M. F., Sagebaum, M., Curio, G., and Blankertz, B. (2011). EEG potentials predict upcoming emergency brakings during simulated driving. *Journal of Neural Engineering*, 8(5):056001.

- James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1**, pages 361–380.
- Jing, B.-Y., Pan, G., Shao, Q.-M., and Zhou, W. (2010). Nonparametric estimate of spectral density functions of sample covariance matrices: A first step. *Annals of Statistics*, 38(6):3724–3750.
- Korniotis, G. M. (2008). Habit formation, incomplete markets, and the significance of regional risk for expected returns. *The Review of Financial Studies*, 21(5):2139–2172.
- Kwon, Y., Choi, Y.-G., Park, T., Ziegler, A., and Paik, M. C. (2017). Generalized estimating equations with stabilized working correlation structure. *Computational Statistics & Data Analysis*, 106:1 – 11.
- Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Annals of Statistics*, 44(3):928–953.
- Lancewicki, T., Goodrich, B., and Arel, I. (2015). Sequential covariance-matrix estimation with application to mitigating catastrophic forgetting. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 628–633.
- Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 150(1–2):233–264.
- 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(5):603–621.
- Ledoit, O. and Wolf, M. (2004a). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*, 30(4):110–119.
- Ledoit, O. and Wolf, M. (2004b). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.
- Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*, 40(2):1024–1060.
- Ledoit, O. and Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis*, 139(2):360–384.



- Ledoit, O. and Wolf, M. (2017a). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *Review of Financial Studies*, 30(12):4349–4388.
- Ledoit, O. and Wolf, M. (2017b). Numerical implementation of the QuEST function. *Computational Statistics & Data Analysis*, 115:199–223.
- Ledoit, O. and Wolf, M. (2018a). Analytical nonlinear shrinkage of large-dimensional covariance matrices. Working Paper ECON 264, Department of Economics, University of Zurich.
- Ledoit, O. and Wolf, M. (2018b). Optimal estimation of a large-dimensional covariance matrix under Stein’s loss. *Bernoulli*, 24(4B). 3791–3832.
- Ledoit, O., Wolf, M., and Zhao, Z. (2019). Efficient sorting: A more powerful test for cross-sectional anomalies. *Journal of Financial Econometrics*. Forthcoming.
- Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, 47(1):13–37.
- Lotte, F. and Guan, C. (2009). An efficient p300-based brain-computer interface with minimal calibration time. In *Assistive Machine Learning for People with Disabilities Symposium (NIPS’09 Symposium)*.
- Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4):457–483.
- Markon, K. (2010). Modeling psychopathology structure: a symptom-level analysis of axis I and II disorders. *Psychological Medicine*, 40(02):273–288.
- Meucci, A. (2005). *Risk and Asset Allocation*. Springer Finance, Berlin Heidelberg New York.
- Meyer-Nieberg, S. and Kropat, E. (2015). A new look at the covariance matrix estimation in evolution strategies. In Pinson, E., Valente, F., and Vitoriano, B., editors, *Operations Research and Enterprise Systems*, pages 157–172, Cham. Springer International Publishing.
- Michaelides, P., Apostolellis, P., and Fassois, S. (2011). Vibration-based damage diagnosis in a laboratory cable-stayed bridge model via an rcp-arx model based method. In *Journal of Physics: Conference Series*, volume 305, page 012104. IOP Publishing.

- Pakel, C., Shephard, N., Sheppard, K., and Engle, R. F. (2017). Fitting vast dimensional time-varying covariance models. Working Paper FIN-08-009, NYU. Available at <https://ssrn.com/abstract=1354497>.
- Pirkl, R. J., Remley, K. A., and Lötbäck Patané, C. S. (2012). Reverberation chamber measurement correlation. *IEEE Transactions on Electromagnetic Compatibility*, 54(3):533–545.
- Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation*. John Wiley & Sons, Hoboken.
- Pyeon, D., Newton, M. A., Lambert, P. F., Den Boon, J. A., Sengupta, S., Marsit, C. J., Woodworth, C. D., Connor, J. P., Haugen, T. H., et al. (2007). Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Research*, 67(10):4605–4619.
- Ren, Y. and Shimotsu, K. (2009). Improvement in finite-sample properties of the Hansen-Jagannathan distance test. *Journal of Empirical Finance*, 16:483–506.
- Ribes, A., Planton, S., and Terray, L. (2013). Application of regularised optimal fingerprinting to attribution. part I: method, properties and idealised analysis. *Climate Dynamics*, 41(11-12):2817–2836.
- Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360.
- 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(1). Article 32.
- Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442.
- Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *Journal of Multivariate Analysis*, 55:331–339.
- Silverstein, J. W. and Bai, Z. D. (1995). On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *Journal of Multivariate Analysis*, 54:175–192.
- Simo-Serra, E., Torras, C., and Moreno-Noguer, F. (2017). 3d human pose tracking priors using geodesic mixture models. *International Journal of Computer Vision*, 122(2):388–408.

- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206. University of California Press.
- Stein, C. (1975). Estimation of a covariance matrix. Rietz lecture, 39th Annual Meeting IMS. Atlanta, Georgia.
- Stein, C. (1982). Series of lectures given at the University of Washington, Seattle. Technical Report.
- Stein, C. (1986). Lectures on the theory of estimation of many parameters. *Journal of Mathematical Sciences*, 34(1):1373–1403.
- Vakili, M., Kitaura, F.-S., Feng, Y., Yepes, G., Zhao, C., Chuang, C.-H., and Hahn, C. (2017). Accurate halogalaxy mocks from automatic bias estimation and particle mesh gravity solvers. *Monthly Notices of the Royal Astronomical Society*, 472(4):4144–4154.
- Wigner, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564.
- Wolf, M. and Wunderli, D. (2012). Fund-of-funds construction by statistical multiple testing methods. In Scherer, B. and Winston, K., editors, *The Oxford Handbook of Quantitative Asset Management*, pages 116–135. Oxford University Press, Oxford.
- Zhang, Y., Sun, D., and Zhang, D. (2009). Robust adaptive acoustic vector sensor beamforming using automated diagonal loading. *Applied Acoustics*, 70(8):1029–1033.
- Zumbach, G. O. (2007). The RiskMetrics 2006 methodology. Technical report, RiskMetrics Group. Available at <https://ssrn.com/abstract=1420185>.