Journal of Financial Econometrics, 2022, Vol. 20, No. 1, 187–218
doi: 10.1093/jjfinec/nbaa007
Advance Access Publication Date: 23 June 2020
Article

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

Olivier Ledoit[1,2] and Michael Wolf[1]

[1]University of Zurich and [2]AlphaCrest Capital Management

Address correspondence to Michael Wolf, Department of Economics, University of Zurich, Zürichbergstrasse 14, CH-8032 Zurich, Switzerland, or e-mail: michael.wolf@econ.uzh.ch.

Received May 10, 2019; revised February 6, 2020; editorial decision March 26, 2020; accepted April 6, 2020

Abstract

Many econometric and data-science applications require a reliable estimate of the covariance matrix, such as Markowitz's 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 article, 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's portfolio selection, rotation equivariance

JEL classification: C13, C58, G11

The covariance matrix is, arguably, the second most important object in all of statistics. In practical applications, such as in Markowitz's 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 15 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.[1] After three papers on linear shrinkage in the first decade of this century, and a creative break, there have been an additional ten papers (by now) 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. As 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 in a review paper. And to go one step beyond, it would also go too far to spell out the assumptions in detail for every method, respectively, approach. So we will take the liberty to be purposefully vague about 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).

The remainder of the article is organized as follows. Section 1 presents linear shrinkage to the identity matrix, which is the starting point of the journey. Section 2 adapts linear shrinkage to a variety of custom-tailored targets. Section 3 generalizes linear shrinkage to nonlinear shrinkage, which is more flexible and powerful. Section 4 presents an extension to dynamic models. Section 5 presents an extension to factor models. Section 6 discusses computational aspects. Finally, Section 7 concludes.

1 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 Σ_T and assumed to be positive definite. Even though the population covariance matrix is fixed in this section, we index it by T for notational consistency with some subsequent sections, where the dimension of the population covariance matrix will vary and go to infinity as a function of the sample size T. To simplify the notation, we assume that all variables have mean zero. In this way, the sample covariance matrix is given by

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

where the symbol $:=$ indicates that the left-hand side is defined to be equal to the right-hand side.

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

Remark 1.1 (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 6. ■

Remark 1.2 (Notational conventions). Population quantities are generally denoted by Greek letters, with sample counterparts denoted by the corresponding Latin letter: for example, Σ_T and S_T. Other estimators can be indicated by various superscripts: star (*) for linear shrinkage and for nonlinear shrinkage circle ($^\circ$) or bullet ($^\bullet$). They represent *bona fide* estimators if they have a hat ($\hat{}$) or tilde ($\tilde{}$) accent, but without that they are an *oracle*, and thus not feasible in practice. This system allows us to present progressively more sophisticated oracle estimators that are optimal in certain ways, and then obtain feasible counterparts for them that have the same asymptotic properties. ■

1.1 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 Charles 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 trade-off. 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 (i.e, 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, which included the suggestion of alternative shrinkage targets, and also empirical applications 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 moves all sample means toward 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 of the mean vector to the estimation of the covariance matrix, keeping the MSE

risk function. To this end, denote by $||\cdot||_{\mathrm{F}}$ the (scaled) Frobenius norm of a square matrix; more specifically, for a $N \times N$ matrix A, this norm is given by

$$||A||_{\mathrm{F}} := \sqrt{\langle A, A\rangle} := \sqrt{\mathrm{Tr}(A'A)/N} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}^2}, \tag{1.1}$$

where a_{ij} denotes the (i, j) entry of A and $\mathrm{Tr}(\cdot)$ 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 b

$$\mathcal{L}_{\mathrm{F}}\big(\hat{\Sigma}_T, \Sigma_T\big) := ||\hat{\Sigma}_T - \Sigma_T||_{\mathrm{F}}^2, \tag{1.2}$$

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

$$\mathbb{E}\big[\mathcal{L}_{\mathrm{F}}\big(\hat{\Sigma}_T, \Sigma_T\big)\big] = \mathbb{E}\Big[||\hat{\Sigma}_T - \Sigma_T||_{\mathrm{F}}^2\Big]. \tag{1.3}$$

The class of estimators we considered was that of linear combinations of the identity matrix, denoted by $\mathbb{I}_T$, and the sample covariance matrix, so that the optimization problem became:

$$\min_{\rho 1, \rho_2} \mathbb{E}\Big[||\hat{\Sigma}_T - \Sigma_T||_{\mathrm{F}}^2\Big] \tag{1.4},$$

$$\text{s.t. } \hat{\Sigma}_T = \rho_1 \mathbb{I}_T + \rho_2 S_T \tag{1.5}.$$

The solution to this problem turns out to be

$$\Sigma_T^* := \frac{\beta_T^2}{\delta_T^2}\mu_T \mathbb{I}_T + \frac{\alpha_T^2}{\delta_T^2} S_T, \tag{1.6}$$

with $\mu_T := \langle \Sigma_T, \mathbb{I}_T\rangle, \alpha_T^2 := ||\Sigma_T - \mu_T \mathbb{I}_T||_{\mathrm{F}}^2, \beta_T^2 := \mathbb{E}\Big[||S_T - \Sigma_T||_{\mathrm{F}}^2\Big]$, and $\delta_T^2 := ||S_T - \mu_T \mathbb{I}_T||_{\mathrm{F}}^2$; see Equation (1.1) for the definition of the operator $\langle\cdot,\cdot\rangle$. As it can be shown that $\alpha_T^2 + \beta_T^2 = \delta_T^2$, the solution (1.6) can also be written as a *convex* linear combination of a multiple of the identity matrix, $\mu_T \mathbb{I}_T$, and S_T, namely,

$$\Sigma_T^* = \gamma_T^* \mu_T \mathbb{I}_T + (1 - \gamma_T^*) S_T \text{ with } \gamma_T^* := \frac{\beta_T^2}{\delta_T^2}. \tag{1.7}$$

In this way, the solution Σ_T^* can be interpreted as shrinking the sample covariance matrix S_T toward the *shrinkage target* $\mu_T \mathbb{I}_T$ with *(shrinkage) intensity* $\gamma_T^* \in [0, 1]$. The multiplier μ_T in the shrinkage target makes intuitive sense: As μ_T is equal to the average diagonal entry of Σ_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 get multiplied by four, just like the sample covariance matrix and also the true covariance matrix.

An important feature of Σ_T^* is that it is positive definite, and thus invertible, even when $N > T$ (in contrast with the sample covariance matrix, which is rank-deficient and thus not

invertible, in this case). This is because Σ_T^* is a convex linear combination of a matrix that is positive definite, $\mu_T \mathbb{I}_T$, and another matrix that is positive semi-definite, S_T, where the weight given to the target $\mu_T \mathbb{I}_T$ is positive (in all cases of practical relevance).

1.2 Asymptotic Analysis

The optimal linear combination Σ_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. As we are only interested in situations where N is of the same magnitude as T, and not in situations where N is vastly larger than T, we may assume without loss of generality that $N/T \to c \in [0, \infty)$, where c is called the *limiting concentration (ratio)*.[2] 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 (1.7), we only need to estimate the three parameters μ_T, δ_T^2, and β_T^2. The respective solutions are as follows. First,

$$\hat{\mu}_T := \langle S_T, \mathbb{I}_T \rangle = \frac{1}{N} \sum_{i=1}^{N} s_{ii}^T, \tag{1.8}$$

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

$$\hat{\delta}_T^2 := ||S_T - \hat{\mu}_T \mathbb{I}_T||_{\mathrm{F}}^2. \tag{1.9}$$

Third, denote by $x_{t\cdot}^T$ the t^{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

$$S_T := \frac{1}{T} X_T' X_T = \frac{1}{T} \sum_{t=1}^{T} x_{t\cdot}^T \left(x_{t\cdot}^T \right)'.$$

The estimator of β_T^2 is then given by

$$\hat{\beta}_T^2 := \min \left\{ \tilde{\beta}_T^2, \hat{\delta}_T^2 \right\} \text{ with } \tilde{\beta}_T^2 := \frac{1}{T} \sum_{t=1}^{T} ||x_{t\cdot}^T \left(x_{t\cdot}^T \right)' - S_T||_{\mathrm{F}}^2, \tag{1.10}$$

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

$$\hat{\Sigma}_T^* := \hat{\gamma}_T^* \hat{\mu}_T \mathbb{I}_T + \left(1 - \hat{\gamma}_T^*\right) S_T \text{ with } \hat{\gamma}_T^* := \frac{\hat{\beta}_T^2}{\hat{\delta}_T^2}. \tag{1.11}$$

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

2 Ledoit and Wolf (2004b) assume the weaker condition that there exists a finite constant K_1 such that $N/T < K_1$ always, but there does not appear to be any practical benefit from having this weaker condition.

Under a set of regularity conditions, one can show that the three estimators (1.8)–(1.10) are consistent in quadratic mean, which implies that $\hat{\Sigma}_T^*$ is also a consistent estimator of Σ_T^* in quadratic mean, namely that

$$\mathbb{E}\left[||\hat{\Sigma}_T^* - \Sigma_T^*||_{\mathrm{F}}^2\right] \to 0. \tag{1.12}$$

(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 (1.12) is that the feasible estimator $\hat{\Sigma}_T^*$ has asymptotically the same risk as the infeasible optimal linear combination Σ_T^* for estimating Σ_T in the sense that

$$\mathbb{E}\left[||\hat{\Sigma}_T^* - \Sigma_T||_{\mathrm{F}}^2\right] - \mathbb{E}\left[||\Sigma_T^* - \Sigma_T||_{\mathrm{F}}^2\right] \to 0. \tag{1.13}$$

Remark 1.3 It is important not to mistake consistent estimation of Σ_T^*—that is, consistent estimation of the optimal linear combination (1.7)—for consistent estimation of Σ_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 $N/T \to 0$; 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, as one cannot expect to estimate $N(N+1)/2$ parameters (namely, the distinct entries of the symmetric matrix Σ_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 (2008a,b), which makes the strong (and unverifiable in practice) assumption that the true covariance matrix Σ_T is sparse. In such a restrictive setting, consistent estimation of Σ_T itself is possible even when $c > 0$. ■

The feasible estimator $\hat{\Sigma}_T^*$ shares with the optimal linear combination Σ_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 (1.11), $\hat{\Sigma}_T^*$ is also a convex linear combination of a positive definite (with probability 1) matrix, $\hat{\mu}_T \mathbb{I}_T$, and positive semi-definite matrix, S_T, where the weight given to the target $\hat{\mu}_T \mathbb{I}_T$ is positive (with probability 1).

1.3 Simulation Evidence

In Ledoit and Wolf (2004b), we compared the finite-sample performance of the feasible estimator $\hat{\Sigma}_T^*$ to that of three other estimators from the literature via Monte Carlo simulations:

- an empirical Bayes estimator proposed by Haff (1980);
- the better performing estimator, for any given simulated data set, of Stein (1975) and Haff (1982);
- and the minimax estimator derived independently by Stein (1982) and Dey and Srinivasan (1985).

(These three estimators were derived under loss functions different from the Frobenius loss; see Section 3.5. 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.)

Over a wide range of scenarios, all four estimators improved upon the sample covariance matrix in terms of empirical MSE, and the shrinkage estimator $\hat{\Sigma}_T^*$ was overall the best; in particular, our estimator improved upon 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.

1.4 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, Sun, and Zhang, 2009).

Cancer research: mapping out the influence of the human papillomavirus on gene expression (Pyeon et al., 2007).

Chemistry: estimating the temporal autocorrelation function for fluorescence correlation spectroscopy (Guo et al., 2012).

Civil engineering: detecting and identifying vibration-based bridge damage through random coefficient pooled models (Michaelides, Apostolellis, and Fassois, 2011).

Climatology: detecting trends in average global temperature through the optimal fingerprinting method (Ribes, Planton, and Terray, 2013).

Electromagnetics: studying correlation between reverberation chamber measurements collected at different stirrer positions (Pirkl, Remley, Lötbäck Patané, 2012)

Entertainment technology: designing a video game controlled by performing tricks on a skateboard (Anlauff et al., 2010).

Genetics: improving the accuracy of genomic-estimated-breeding-value predictions with low-density markers (Endelman and Jannink, 2012).

Geology: modeling multiphase flow in subsurface petroleum reservoirs with the iterative stochastic ensemble method on inverse problems (Elsheikh, Wheeler, and Hoteit, 2013).

Image recognition: detecting anomalous pixels in hyperspectral imagery (Bachega et al., 2011).

Macro-finance: improved generalized least squares AQregressions 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 electronic signals impinging upon an array of receptors (Abrahamsson, Selen, and Stoica, 2007).

Speech recognition: automatic transcription of phone conversation records (Bell and King, 2009).

2 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's portfolio selection. So why is finance largely missing in the previous list of fields? This is because we devised some alternative estimators, at the same time already, which 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 a 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 in form of the capital asset pricing model (CAPM) dating back to Sharpe (1964) and Lintner (1965). According to this model,

$$x_{ti} = \alpha_i + \beta_i x_{t0} + u_{ti} \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{2.1}$$

where x_{ti} is the return on stock i in period t, α_i and β_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 (2.1) implies that $\Sigma_T = \Phi_T$ with

$$\Phi_T := \sigma_{00}^T \beta_T \beta_T' + \Delta_T, \tag{2.2}$$

where $\sigma_{00}^T := \mathrm{Var}(x_{t0})$, $\beta_T := (\beta_1, \ldots, \beta_N)'$, and Δ_T is a diagonal matrix with typical entry $\delta_{ii}^T := \mathrm{Var}(u_{ti})$. The matrix Φ_T is unknown in practice but it can be estimated as

$$F_T := \hat{\sigma}_{00}^T \hat{\beta}_T \hat{\beta}_T' + \hat{\Delta}_T, \tag{2.3}$$

where $\hat{\sigma}_{00}^T$ is the sample variance of the $\{x_{t0}\}$, $\hat{\beta}_T$ is obtained through estimating model (2.1) by ordinary least squares (OLS) one stock at a time, and a typical diagonal entry of $\hat{\Delta}_T$ is given by $\hat{\delta}_{ii}^T$, 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 Σ_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 (i.e., variance). The problem is that no one believes in the CAPM anymore, at least not as a "perfect" model. Indeed, the CAPM serves as a 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 Σ_T. Intuitively, the closer F_T is to the true Σ_T, the larger should be the shrinkage intensity γ_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 Φ_T with typical entry $\phi_{ij}^T = \sqrt{\sigma_{ii}^T \sigma_{jj}^T}\rho_T$, where ρ_T is a common correlation, that is, $\mathrm{Cor}(x_{ti}, x_{tj}) \equiv \rho_T$; here, the symbol $\equiv$ indicates that the left-hand side is constantly equal to the right-hand side. The feasible shrinkage target is then given by

$$F_T \text{ with typical entry } f_{ij}^T := \sqrt{\hat{\sigma}_{ii}^T \hat{\sigma}_{jj}^T}\hat{\rho}_T, \tag{2.4}$$

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 Φ_T that has a common variance σ^2 on the diagonal and a common covariance η on the off-diagonal. The feasible shrinkage target is then given by

$$F_T \text{ with } f_{ii}^T := \hat{\sigma}_T^2 \text{ and } f_{ij}^T := \hat{\eta}_T, \tag{2.5}$$

where $\hat{\sigma}_T^2$ is the average of the N sample variances of the $\{x_{ti}\}$, for $i = 1, \ldots, N$, 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 (2.5) is a generalization of the generic target used in Equation (1.11), which has the same value $\hat{\sigma}_T^2$ on the diagonal but sets the off-diagonal entries 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 (2.5) was first suggested by (Ledoit, 1995, Appendix B.1), then was also proposed 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.

2.1 Finite-Sample Analysis

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

$$\min_{\gamma_T} \mathbb{E}\left[\|\hat{\Sigma}_T - \Sigma_T\|_{\mathrm{F}}^2\right] \tag{2.6}$$

$$\text{s.t. } \hat{\Sigma}_T = \gamma_T F_T + (1-\gamma_T) S_T \tag{2.7}$$

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

$$\gamma_T^* := \frac{\sum_{i=1}^N \sum_{j=1}^N \left[\mathrm{Var}\left(s_{ij}^T\right) - \mathrm{Cov}\left(f_{ij}^T, s_{ij}^T\right)\right]}{\sum_{i=1}^N \sum_{j=1}^N \left[\mathrm{Var}\left(f_{ij}^T - s_{ij}^T\right) + \left(\phi_{ij}^T - \sigma_{ij}^T\right)\right]}, \tag{2.8}$$

resulting in the optimal linear combination

$$\Sigma_T^* := \gamma_T^* F_T + \left(1 - \gamma_T^*\right) S_T. \tag{2.9}$$

(More generally, one needs to replace ϕ_{ij}^T with $\mathbb{E}\left(f_{ij}^T\right)$ in Equation (2.8) 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 Σ_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 γ_T^*; this is indeed the case as will be explained below.

2.2 Asymptotic Analysis

Again, the optimal linear combination Σ_T^* depends on unknown population quantities and must be estimated in practice to obtain a feasible estimator. Unfortunately, in this more general setting, we had 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

$$\gamma_T^* = \frac{1}{T}\frac{\pi_T - \rho_T}{\nu_T} + O\left(\frac{1}{T^2}\right) \tag{2.10}$$

$$\text{with } \pi_T := \sum_{i=1}^{N}\sum_{j=1}^{N} \text{AsyVar}\left(\sqrt{T}s_{ij}^T\right), \tag{2.11}$$

$$\rho_T := \sum_{i=1}^{N}\sum_{j=1}^{N} \text{AsyCov}\left(\sqrt{T}f_{ij}^T, \sqrt{T}s_{ij}^T\right), \tag{2.12}$$

$$\text{and} \quad \nu_T := \sum_{i=1}^{N}\sum_{j=1}^{N} \left(\phi_{ij}^T - \sigma_{ij}^T\right)^2, \tag{2.13}$$

where $\text{AsyVar}(\cdot)$ stands for asymptotic variance and $\text{AsyCov}(\cdot)$ stands for asymptotic covariance, tacitly assuming a set of regularity conditions to ensure that these quantities exist.[3]

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

First, the term π_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. Second, the term ρ_T measures the ("combined") covariance between the

3 Consider an estimator $\hat{\theta}_T$ of a parameter θ 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 analogously for the definition of the operator $\text{AsyCov}(\cdot,\cdot)$.

data-dependent shrinkage target F_T and S_T[4]; ceteris paribus, the larger the covariance between F_T and S_T, the smaller should be the shrinkage intensity, because a large covariance implies that F_T provides little independent information about Σ_T relative to S_T. Third, the term ν_T measures how close the population version of the shrinkage target, Φ_T, is to the population covariance matrix, Σ_T; ceteris paribus, the closer the two are to each other, the larger should be the shrinkage intensity.

Given formula (2.10), the estimation of the optimal shrinkage intensity γ_T^* is straightforward: estimate the three terms π_T, ρ_T, and ν_T separately and ignore higher-order terms.

First, a consistent estimator of π_T is standard and given by

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

(Note that this estimator is numerically equal to the estimator $\tilde{\beta}_T$ of Equation (1.10).)

Second, a consistent estimator of ρ_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 (2.3) and (2.4) in Ledoit and Wolf (2003, lemma 2) and Ledoit and Wolf (2004a, Appendix B), respectively. Some details for the choice (2.5) can be found in Schäfer and Strimmer (2005, Appendix A).

Third, a consistent estimator of ν_T is simply given by

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

and thus does not require any extra work.

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

$$\hat{\gamma}_T^* := \min\left\{ \max\left\{ \tilde{\gamma}_T^*, 0 \right\}, 1 \right\} \text{ with } \tilde{\gamma}_T^* := \frac{1}{T} \frac{\hat{\pi}_T - \hat{\rho}_T}{\hat{\nu}_T}, \tag{2.14}$$

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

$$\hat{\Sigma}_T^* := \hat{\gamma}_T^* F_T + \left(1 - \hat{\gamma}_T^*\right) S_T. \tag{2.15}$$

(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 ρ_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 Φ_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 γ_T can be derived using the methodology described above.

4 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.

2.3 Simulation Evidence

In Ledoit and Wolf (2004a), we carried out a "hybrid" empirical analysis that was part of real-life back-testing and part simulation study. The idea was to emulate a portfolio manager who has a certain amount of skill in predicting future stock returns. The setting was 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's 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 (i.e., 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 (2.3), linear shrinkage to the constant-correlation matrix (2.4), and an estimator 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 $N \geq 225$.

2.4 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's 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 (i.e., 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's 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's portfolio based on the estimator of the covariance matrix obtained by shrinkage to the single-factor matrix.[5]

An application outside of Markowitz's portfolio selection involves the famous distance test by Hansen and Jagannathan (1997) 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) showed that if

5 Backtest exercises involving (smaller) universes of portfolios as assets, such as ten industry portfolios or twenty-five Fama–French portfolios yielded inconclusive results, with neither methodology dominating the other.

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 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 Paolella (2009).

3 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 (1.7) is one of moving each entry of the sample covariance matrix S_T to the shrinkage target $\mu_T \mathbb{I}_T$ with *common* intensity γ_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 $\mu_T \mathbb{I}_T$. 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 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 turned out to be the *spectral decomposition* of the sample covariance matrix, which is given by $S_T = U_T \Lambda_T U_T'$. Here, $\Lambda_T := \mathrm{Diag}(\lambda_{T,1}, \ldots, \lambda_{T,N})$ is a diagonal matrix[6] whose diagonal entries are the sample eigenvalues $\lambda_{T,i}$, and $U_T := [u_{T,1}, \ldots, u_{T,N}]$ is an orthogonal matrix whose columns are the sample eigenvectors $u_{T,i}$. This starting point provides an alternative interpretation of the optimal linear combination (1.7) which we already pointed out in Ledoit and Wolf (2004b): one can also express this linear combination as

$$\Sigma_T^* := U_T \Delta_T^* U_T' \text{ with } \Delta_T^* := \mathrm{Diag}(\delta_{T,1}^*, \ldots, \delta_{T,N}^*) \text{ and } \delta_{T,i}^* := \gamma_T^* \mu_T + (1 - \gamma_T^*) \lambda_{T,i}. \quad (3.1)$$

Therefore, Σ_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 entries of S_T to the entries of the target $\mu_T \mathbb{I}_T$ with common intensity γ_T^* is equivalent to keeping the eigenvectors of S_T and shrinking its eigenvalues to the target μ_T (which is actually equal to the mean of the population eigenvalues) with *the same* common intensity γ_T^*. The generalization is now obvious: use different shrinkage intensities for different sample eigenvalues;

6 In a slight abuse of notation, we will use the operator $\mathrm{Diag}(\cdot)$ for two different purposes. On the one hand, for an $N \times 1$ vector $a := (a_1, \ldots, a_N)'$, $\mathrm{Diag}(a_1, \ldots, a_N)$ denotes an $N \times N$ diagonal matrix whose diagonal is the vector a. On the other hand, for an $N \times N$ matrix A with typical entry a_{ij}, $\mathrm{Diag}(A)$ denotes the $N \times 1$ vector $(a_{11}, \ldots, a_{NN})'$.

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 μ_T.) Clearly, this more general approach will be better than 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 fits into the following class of estimators already introduced by Stein (1975, 1986):

$$\hat{\Sigma}_T := U_T \Delta_T U_T' \text{ with } \Delta_T := \text{Diag}(\delta_{T,1}, \ldots, \delta_{T,N}) \ , \tag{3.2}$$

where, Δ_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. Note that the specific shrinkage formula injected by Stein into Equation (3.2) differs from our own choices reviewed in this article; see Ledoit and Wolf (2018, p. 3810) for a more detailed comparison between the shrinkage formulas.

One nice property of this class is that it only contains estimators that are *rotation-equivariant*. Let W_T be a rotation matrix, that is, an orthogonal matrix with determinant one. Also, let $\hat{\Sigma}_T(\cdot)$ denotes a generic estimator of the covariance matrix Σ_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 to rotation of the estimator. In the absence of any a priori knowledge on the structure of Σ_T, rotation-equivariance is a natural and desirable property of a covariance matrix estimator. Intuitively, we have to retain the sample eigenvectors because we have no idea in which direction to rotate them (with the goal of bringing them closer to their population counterparts), unless we make strong structural assumptions like sparsity.

3.1 Finite-Sample Analysis

It had been our original plan to do as before with linear shrinkage, that is, 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:

$$\min_{\Delta_T} ||U_T \Delta_T U_T' - \Sigma_T||_{\text{F}}^2, \tag{3.3}$$

$$\text{s.t. } \Delta_T = \text{Diag}(\delta_{T,1}, \ldots, \delta_{T,N}). \tag{3.4}$$

The optimal solution to this problem turns out to be

$$\mathring{\Delta}_T := \text{Diag}\left(\mathring{\delta}_{T,1}, \ldots, \mathring{\delta}_{T,N}\right) \text{ with } \mathring{\delta}_{T,i} := u_{T,i}' \Sigma_T u_{T,i}. \tag{3.5}$$

The optimal diagonal entries $\mathring{\delta}_{T,i}$, unsurprisingly, are not equal to the sample eigenvalues, as then we would recuperate the sample covariance matrix as the estimator. But,

surprisingly perhaps, the optimal diagonal entries 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 entries $\delta^{\circ}_{T,i}$ combine sample and population information, namely how the i^{th} sample eigenvector $u_{T,i}$ relates to the population covariance matrix Σ_T.

Remark 3.1 (Relation to PCA). Another way to look at the optimal entry $\delta^{\circ}_{T,i}$ is that it is equal to the variance of the i^{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 $\delta^{\circ}_{T,i}$, 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

$$\Sigma^{\circ}_T := U_T \Delta^{\circ}_T U_T \quad \text{with } \Delta^{\circ}_T \text{ as defined in Equation (3.5).} \tag{3.6}$$

An important feature of Σ°_T is that it is positive definite, and thus invertible, even in the case when $N > T$. This is because $\delta^{\circ}_{T,i} > 0$, for $i = 1, \ldots, N$.

We actually already mentioned the entries $\delta^{\circ}_{T,i}$ 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.

3.2 Asymptotic Analysis

The "estimator" (3.6) is not feasible in practice (hence the use of the single quotation marks), as the optimal diagonal entries $\delta^{\circ}_{T,i}$ depend on the very object that we want to estimate: the true covariance matrix Σ_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 entries of the matrix Δ°_T; clearly, such a function does not depend on the number of diagonal entries, 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 (nonstochastic) limit, as T tends to infinity. The goal is then to find the optimal limiting function, and estimate it consistently from the observed data. As any meaningful 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

$$\hat{\Sigma}_T := U_T \hat{\Delta}_T U_T \text{ with } \hat{\Delta}_T := \text{Diag}\left(\hat{\delta}_{T,1}, \ldots, \hat{\delta}_{T,N}\right), \tag{3.7}$$

where $\hat{\delta}_{T,i} := \hat{\phi}_T(\lambda_{N,i})$ and $\hat{\phi}_T : \mathbb{R} \to \mathbb{R}_+$ 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 ϕ. The goal is to find the "optimal" ϕ.

In order to accomplish this goal, we had to invoke some 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 $c \in (0,1) \cup (1,\infty)$,[7] the empirical distribution of the sample eigenvalues converges almost surely to a non-stochastic limit distribution F. If $c < 1$, this limit distribution is continuous with positive support; if $c > 1$, it is a "mixture" distribution with a discrete part at $\{0\}$ and a continuous part with positive support. The limit distribution F 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 existences (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 later restated in an alternative expression by Silverstein (1995).

A key step in finding the optimal limiting shrinkage function ϕ was the realization that the loss function

$$\mathcal{L}_{\text{F}}\left(\hat{\Sigma}_T, \Sigma_T\right) := ||\hat{\Sigma}_T - \Sigma_T||_{\text{F}}^2$$

is non-stochastic in the limit for estimators $\hat{\Sigma}_T$ in the class (3.7). That is, under the assumed set of regularity conditions, $\mathcal{L}_{\text{F}}\left(\hat{\Sigma}_T, \Sigma_T\right)$ converges almost surely to a non-stochastic limiting expression, as T goes to infinity.[8] This limiting expression, as to be expected, depends on the limiting shrinkage function ϕ and can be minimized with respect to it. The corresponding minimizer is then the optimal ϕ, which we shall henceforth denote as $\phi^\bullet$. Over the years, we have established different formulas for $\phi^\bullet$; they are all equivalent and just look different, as they are based on different mathematical tools, respectively,

7 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 $N/T = 1$.

8 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, as the loss is already non-stochastic in the limit; in other words, in the limit, the loss is equal to the risk.

ingredients. For the sake of space, we shall limit ourselves here to the formula given in Ledoit and Wolf (2020):

$$\phi^{\bullet}(x) := \begin{cases} \dfrac{1}{\pi(c-1)\mathcal{H}_{\underline{f}}(0)} & \text{if } x = 0 \text{ and } c > 1 \\ \dfrac{x}{\left[\pi c x f(x)\right]^2 + \left[1 - c - \pi c x \mathcal{H}_f(x)\right]^2} & \text{otherwise} \end{cases} \tag{3.8}$$

There are three ingredients in this formula that need to be explained: f, $\underline{f}$, and the operator $\mathcal{H}$. First, $f(x)$ for $x > 0$ denotes the density of F on $(0, +\infty)$; second, $\underline{f}$ is the density of $\underline{F} := (1 - c)\ \mathbb{1}_{[0,+\infty)} + cF$, which is continuous everywhere when $c > 1$; third, for a real function g, $\mathcal{H}_g$ denotes its *Hilbert transform*, defined as

$$\forall x \in \mathbb{R} \qquad \mathcal{H}_g(x) := \frac{1}{\pi} PV \int_{-\infty}^{+\infty} g(t) \frac{dt}{t - x}. \tag{3.9}$$

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

$$PV \int_{-\infty}^{+\infty} g(t) \frac{dt}{t - x} := \lim_{\varepsilon \to 0^+} \left[\int_{-\infty}^{x-\varepsilon} g(t) \frac{dt}{t - x} + \int_{x+\varepsilon}^{+\infty} g(t) \frac{dt}{t - x} \right]. \tag{3.10}$$

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 (3.8), in a different mathematical expression, was first discovered by Ledoit and Péché (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 Σ_T is then given by

$$\Sigma_T^{\bullet} := U_T \Delta_T^{\bullet} U_T' \text{ with } \Delta_T^{\bullet} := \text{Diag}\left(\phi^{\bullet}(\lambda_{T,1}), \ldots, \phi^{\bullet}(\lambda_{T,N})\right), \tag{3.11}$$

where the quantities $\left(\phi^{\bullet}(\lambda_{T,1}), \ldots, \phi^{\bullet}(\lambda_{T,N})\right)$ represent large-dimensional asymptotic counterparts to the finite-sample optimal quantities $\left(\delta_{T,1}^{\circ}, \ldots, \delta_{T,1}^{\circ}\right)$ of Equation (3.5).

What have we gained by moving from the finite-sample optimal estimator Σ_T° to the oracle estimator $\Sigma_T^{\bullet}$? Nothing in the sense of feasibility, as the oracle estimator is also infeasible in practice, as it also depends on unknown population quantities, such as the density f. But unlike Σ_T°, the oracle estimator serves as a useful starting point for deriving a feasible estimator; this is because it is possible to consistently estimate the oracle shrinkage function $\phi^{\bullet}$ of Equation (3.8).

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 estimators of f, $\underline{f}$, and their two respective Hilbert transforms. (As a technical detail, these estimators need to be *uniformly* consistent.)

In Ledoit and Wolf (2020), we detail how this task can be accomplished by means of kernel estimation.[9] The feasible nonlinear shrinkage estimator of Σ_T is then given by

$$\hat{\Sigma}_T^{\bullet} := U_T \hat{\Delta}_T^{\bullet} U_T' \text{ with } \hat{\Delta}_T^{\bullet} := \text{Diag}\Big(\hat{\phi}_T^{\bullet}(\lambda_{T,1}), \ldots, \hat{\phi}_T^{\bullet}(\lambda_{T,N})\Big), \tag{3.12}$$

where

$$\hat{\phi}_T^{\bullet}(x) := \begin{cases} \dfrac{1}{\pi(\hat{c}_T - 1)\hat{\mathcal{H}}_{T,\underline{f}}(0)} & \text{if } x = 0 \text{ and } \hat{c}_T > 1 \\ \dfrac{x}{\Big[\pi \hat{c}_T x \hat{f}_T(x)\Big]^2 + \Big[1 - \hat{c}_T - \pi \hat{c}_T x \hat{\mathcal{H}}_{T,f}(x)\Big]^2} & \text{otherwise} \end{cases}, \tag{3.13}$$

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

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

$$\mathcal{L}_{\text{F}}\Big(\hat{\Sigma}_T^{\bullet}, \Sigma_T\Big) - \mathcal{L}_{\text{F}}\big(\Sigma_T^{\bullet}, \Sigma_T\big) \xrightarrow{p} 0.$$

Remark 3.2 (Other estimation strategies). The title of our paper (Ledoit and Wolf, 2020) stems from the fact that the outlined method is the first to "directly" estimate the oracle shrinkage function $\phi^{\bullet}$ with the analytical formula (3.13).

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 $\phi^{\bullet}$ used in those previous papers. The most demanding step in practice is the first one, the estimation of H, as it involves a large-dimensional optimization problem that has no analytical solution and must be solved by large-scale optimization software instead. Therefore, these alternative strategies can also be characterized as numerical strategies.

Such numerical strategies work well in practice but they are cumbersome to implement[10] 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 $N = 1000$.

A completely different strategy to consistently estimate the oracle, based on repeated sample splitting, has been suggested by Abadir, Distaso, and Žikeš (2014) and Lam (2016); this strategy is of completely numerical nature and bypasses estimation of the oracle

9 This idea goes back to Jing et al. (2010), although they only considered estimation of f.

10 In fact, we had to write a separate paper to detail the implementation; see Ledoit and Wolf (2017b).

shrinkage function $\phi^{\bullet}$ 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 $N = 1000$. ■

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 *toward* 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, toward local "centers of attraction." Nevertheless, nonlinear shrinkage still reduces the overall spread compared to the sample eigenvalues.

Figure 1 provides a graphical illustration of 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. In this example, the average eigenvalue is equal to 1.25. Sample eigenvalues below the average but above 1 are "shrunk" downward because they are attracted by the cluster to their immediate left; this is because these sample eigenvalues generally correspond to a population eigenvalue of 0.8. Similarly, sample eigenvalues above the average but below 1.75 are "shrunk" upward 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 toward the grand mean of 1.25.

Generally speaking, when the population eigenvalues are tightly clustered in a bulk, the optimal nonlinear shrinkage is nearly linear; however, when the population eigenvalues are dispersed, lumpy, or otherwise unruly, the optimal nonlinear shrinkage can be highly nonlinear.

3.3 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 (2020). 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 *Quantized Eigenvalues Sampling Transform* (QuEST).

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 $\phi^{\bullet}$ is (nearly) a linear function; for example, this is the case when the population eigenvalues are (nearly) identical. 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, as the performance difference also depends on the distribution of the population eigenvalues; but, as a rule of thumb, one

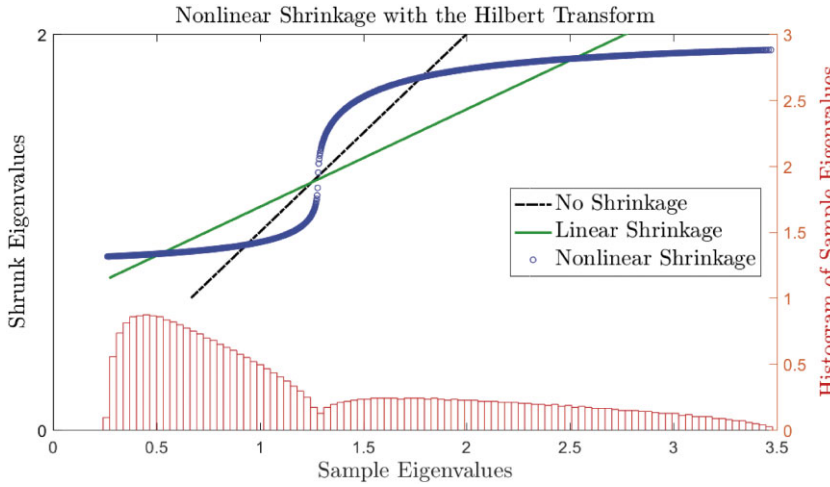

Figure 1 Local attraction effect. A total of 2500 population eigenvalues are equal to 0.8, and 1500 are equal to 2, so that $N = 4000$. The sample size is $T = 18{,}000$. At the bottom of the figure is a histogram displaying the location of the sample eigenvalues.

generally needs both N and T to be above 50 for nonlinear shrinkage to outperform linear shrinkage when $\phi^\bullet$ is "markedly" nonlinear, on the one hand, and for nonlinear shrinkage to be, basically, as good as linear shrinkage when $\phi^\bullet$ is (nearly) linear, on the other hand. As a consequence, when both N and T are above 50, 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 $N = 1000$. As discussed in Remark 3.2, the analytical strategy of Ledoit and Wolf (2020) fixes both problems. In terms of accuracy, extensive Monte Carlo studies in Ledoit and Wolf (2020) showed 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!

3.4 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:

Chemometrics: multivariate analysis of variance and PCA (Engel, Buydens, and Blanchet, 2017).

Finance: Markowitz's portfolio selection (Agrawal, Roy, and Uhler, 2019; Choi Lim, and Choi, 2019; Moura, Santos, and Ruiz, 2019; Trucios et al., 2019).
Neuroscience: likelihood estimation of drug occupancy for brain positron emission tomography studies (Schain, Zanderigo, and Ogden, 2018).
Signal processing: developing detection mechanisms for high-dimensional signals (Robinson, 2019).

A major application of interest to us was, again, Markowitz's portfolio selection. Needless to say, one can argue that the generic (scaled) Frobenius loss function $\mathcal{L}_{\mathrm{F}}$ of Equation (1.2) 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 $\phi^{\bullet}$ of Equation (3.8) and, therefore, also *the same* feasible nonlinear shrinkage estimator $S_T^{\bullet}$ of Equation (3.12).

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

$$\mathcal{L}_{\mathrm{MV}}\left(\hat{\Sigma}_T, \Sigma_T\right) := \frac{\mathrm{Tr}\left(\hat{\Sigma}_T^{-1}\Sigma_T\hat{\Sigma}_T^{-1}\right)/N}{\left[\mathrm{Tr}\left(\hat{\Sigma}_T^{-1}\right)/N\right]^2} - \frac{1}{\mathrm{Tr}\left(\Sigma_T^{-1}\right)/N}. \tag{3.14}$$

Roughly speaking, $\mathcal{L}_{\mathrm{MV}}$ 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 MV loss function is given in Engle and Colacito (2006) and Ledoit and Wolf (2017a).

Remark 3.3 (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 $\phi^{\bullet}$ of Equation (3.8) and, therefore, also *the same* feasible nonlinear shrinkage estimator $S_T^{\bullet}$ of Equation (3.12). ■

In Ledoit and Wolf (2017a), we compared the nonlinear shrinkage estimator $S_T^{\bullet}$ to a variety of other covariance matrix estimators from the literature for the purpose of Markowitz's 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.

3.5 Different Oracle Shrinkage Formulas

We have seen that the oracle shrinkage formula $\phi^{\odot}$ of Equation (3.8) holds for two different loss functions: the Frobenius loss $\mathcal{L}_{\mathrm{F}}$ of Equation (1.2) and the MV loss $\mathcal{L}_{\mathrm{MV}}$ of Equation (3.14). 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

$$\mathcal{L}_{\mathrm{St}}(\hat{\Sigma}_T, \Sigma_T) := \frac{1}{N}\mathrm{Tr}\left(\Sigma_T^{-1}\hat{\Sigma}_T\right) - \frac{1}{N}\log\left[\det\left(\Sigma_T^{-1}\hat{\Sigma}_T\right)\right] - 1. \tag{3.15}$$

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

$$\mathcal{L}_{\mathrm{S-St}}(\hat{\Sigma}_T, \Sigma_T) := \frac{1}{2N}\mathrm{Tr}\left(\Sigma_T^{-1}\hat{\Sigma}_T + \Sigma_T\hat{\Sigma}_T^{-1}\right) - 1. \tag{3.16}$$

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

4 Extension to Dynamic Models

So far, we have used the assumption that the T observations (i.e., 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-varying 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 is multivariate Generalized Autoregressive Condtional Heteroskedasticity (GARCH) models; for example, see Bauwens, Laurent, and Rombouts (2006) for a review.

Unfortunately, these models suffer from the curse of dimensionality: As they can be quite complex and contain a large number of parameters, they do not work well, or cannot be even estimated to begin with, when the number of assets is large, such as $N = 1000$; indeed, in basically all prior applications in the literature, the number of assets had been at most $N = 100$ and often even been 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, $N = 1000$ 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_{ti}: observed return for asset i at date t, stacked into $r_t := (r_{t1}, \ldots, r_{tN})'$;
- $d_{ti}^2 := \mathrm{Var}(r_{ti}|\mathcal{F}_{t-1})$: conditional variance of the i^{th} return at date t;
- D_t: N-dimensional diagonal matrix whose i^{th} diagonal entry is d_{ti};
- $H_t := \mathrm{Cov}(r_t|\mathcal{F}_{t-1})$: conditional covariance matrix at date t; thus, $\mathrm{Diag}(H_t) = D_t^2$.
- $s_{ti} := r_{ti}/d_{ti}$: devolatized return, stacked into $s_t := (s_{t1}, \ldots, s_{tN})'$;
- $R_t := \mathrm{Cor}(r_t|\mathcal{F}_{t-1}) = \mathrm{Cov}(s_t|\mathcal{F}_{t-1})$: conditional correlation matrix at date t; and
- $C := \mathbb{E}(R_t) = \mathrm{Cor}(r_t) = \mathrm{Cov}(s_t)$: 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:

$$Q_t = (1-\alpha-\beta)C + \alpha\, s_{t-1}s_{t-1}' + \beta\, Q_{t-1}. \tag{4.1}$$

Here, (α, β) 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 (4.1) is that the implied unconditional covariance matrix of Q_t is guaranteed to be equal to C, no matter what the values of (α, β) 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 devolatized residuals. It cannot be used directly because its diagonal entries, although generally close to 1, are not exactly equal to 1. From this representation, one obtains the conditional correlation matrix and the conditional covariance matrix as

$$R_t := \mathrm{Diag}(Q_t)^{-1/2} Q_t\, \mathrm{Diag}(Q_t)^{-1/2}, \tag{4.2}$$

$$H_t := D_t R_t D_t, \tag{4.3}$$

and the data-generating process is driven by the multivariate normal law[11]

$$r_t | \mathcal{F}_{t-1} \sim \mathcal{N}(0, H_t). \tag{4.4}$$

How can model (4.1) 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,[12] 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 the DCC parameters (α, β) based on the feasible relation

$$Q_t = (1-\alpha-\beta)\hat{C} + \alpha\, \hat{s}_{t-1}\hat{s}_{t-1}' + \beta\, Q_{t-1}, \tag{4.5}$$

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

Shrinkage estimation helps with step 2. The problem here is that C is an $N \times N$ matrix, which can be large-dimensional. The original proposal of Engle (2002) was to use the sample correlation matrix of the "feasible" devolatized return series $\{\hat{s}_t\}$; this approach works well for dimensions $N \leq 100$, but not for dimensions $N = 1000$ and above. A superior approach in large dimensions is to use a shrinkage estimator on the return series $\{\hat{s}_t\}$; linear shrinkage to (a multiple of) the identity already works well but nonlinear shrinkage works even better. Note that the resulting estimator needs to be post-processed along the lines of Equation (4.2) to produce a proper correlation matrix $\hat{C}$.

11 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 (α, β).

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

Having estimators $\hat{s}_{t-1}$ and $\hat{C}$, one can now estimate the DCC parameters (α, β) based on the feasible relation (4.5). The natural approach is to use maximum likelihood, based on assumption (4.4). This becomes a computational problem in large dimensions, as 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. (2020) 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 (α, β) in a manner that is computationally feasible.

The resulting DCC-NL estimator (where NL stands for Nonlinear Shrinkage) of Engle, Ledoit, and Wolf (2019) unfolds in a three-stage process:

1. For each asset, fit a univariate GARCH(1,1) model and use the fitted models to devolatize 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 two DCC parameters (α, β).

In Engle, Ledoit, and Wolf (2019), we studied the properties of the DCC-NL estimator when used for the purpose of Markowitz's 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, as 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.[13]

As a further application, in Ledoit, Wolf, and Zhao (2019), we showed how to use the DCC-NL estimator to construct more powerful tests for cross-sectional anomalies, that is, more 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 portfolios can be constructed by incorporating an accurate estimator of this covariance matrix in the spirit of Markowitz's portfolio selection; in practice, we proposed the use of the DCC-NL estimator

13 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 $N = 1000$; see Zumbach (2007). Unfortunately, this model performed poorly and cannot be recommended for the purpose of Markowitz's portfolio selection.

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.

5 Extension to Factor Models

Factor models have a long history in finance, with a wide range of applications in both theory and practice. Examples of theoretical applications are asset pricing models, such as the CAPM and the *Arbitrage Pricing Theory* 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 2. 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, Ledoit, and Wolf (2021), 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, Ledoit, and Wolf (2021), 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 $i = 1, \ldots, N$,

$$r_{ti} = \alpha_i + \beta_i' f_t + u_{ti}, \tag{5.1}$$

where $f_t := (f_{t1}, \ldots, f_{tK})'$ is a vector of returns on K observed (or explicit) factors, $\beta_i := (\beta_{i1}, \ldots, \beta_{iK})'$ is a vector of time-invariant factor loadings, and u_{ti} is an error term that satisfies $\mathbb{E}(u_{ti}|f_t) = 0$.

Let $\Sigma_{f,t} := \mathrm{Cov}(f_t|\mathcal{F}_{t-1})$ and $\Sigma_{u,t} := \mathrm{Cov}(u_t|\mathcal{F}_{t-1})$, with $u_t := (u_{t1}, \ldots, u_{tN})'$. 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.[14]

Using again the notation $H_t := \mathrm{Cov}(r_t|\mathcal{F}_{t-1})$, and denoting by B, the $K \times N$ matrix whose i^{th} column is the vector β_i, model (5.1) in conjunction with our stated assumption implies that

$$H_t = B'\Sigma_f B + \Sigma_{u,t}, \tag{5.2}$$

which specializes to

$$H_t \equiv H := B'\Sigma_f B + \Sigma_u \tag{5.3}$$

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

14 We also tried models with $\Sigma_{f,t}$ time-varying but the performance was not better.

In practice, we need to estimate three ingredients: B, Σ_f, and $\Sigma_{u,t}$, respectively, Σ_u. For every asset $i = 1, \ldots, N$, we estimate model (5.1) by a time-series regression using OLS, resulting in estimators $\hat{\alpha}_i$ and $\hat{\beta}_i$ and in residuals $\{\hat{u}_{ti}\}$. The estimator of B, denoted by $\hat{B}$, is then the $K \times N$ matrix whose i^{th} column is the vector $\hat{\beta}_i$.

The estimator of Σ_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 $\hat{u}_t := (\hat{u}_{t1}, \ldots, \hat{u}_{tN})'$. In general, we estimate $\Sigma_{u,t}$ by applying DCC-NL to the residuals $\{\hat{u}_t\}$ and denote the resulting estimator by $\hat{\Sigma}_{u,t}$; in the specialized case $\Sigma_{u,t} \equiv \Sigma_u$, we estimate Σ_u by applying nonlinear shrinkage to the residuals $\{\hat{u}_t\}$ and denote the resulting estimator by $\hat{\Sigma}_u$. Doing so results in the estimator

$$\hat{H}_t := \hat{B}' \hat{\Sigma}_f \hat{B} + \hat{\Sigma}_{u,t}, \tag{5.4}$$

which specializes to

$$\hat{H}_t \equiv \hat{H} := \hat{B}' \hat{\Sigma}_f \hat{B} + \hat{\Sigma}_u \tag{5.5}$$

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

Remark 5.1 (Comparison with more traditional estimators). For the special case (5.3), we can compare our estimator (5.5) with more traditional approaches. The earliest approach in the literature was to assume an *exact* factor model (EFM), which corresponds to assuming that Σ_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 entry to zero. The problem with this approach is that the assumption of an EFM is often violated in practice and thus the resulting estimator $\hat{\Sigma}_f$ can suffer from severe biases, leading to unsatisfactory performance.

In this day and age, the assumption of an *approximate* factor model (AFM) is more common, which corresponds to assuming that Σ_u is a sparse matrix. In this setting, one obtains an estimator of Σ_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 Principal Orthogonal Complement Thresholding (POET) covariance matrix estimator of Fan, Liao, and Mincheva (2013).[15] 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, Ledoit, and Wolf (2021), we studied the properties of a variety of covariance matrix estimators when used for the purpose of Markowitz's 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[16] 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 article is the first one that includes (sophisticated) estimators from all four

15 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.

16 A static estimator assumes that the conditional covariance matrix is time-invariant, whereas a dynamic estimator assumes that it is time-varying.

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, AFMs, unsurprisingly, performed better than EFMs. The overall best estimator, which we call AFM–DCC-NL, was estimator (5.4).

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 EFM worked better when using all five factors as opposed to using the first factor (i.e., the market factor) only. An important advantage of the AFM–DCC-NL estimator is that it worked just as well when using the first factor only as when using all five factors. Apparently, 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. This is good news, in particular for managers who want to invest outside of the United States, 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 5.1, both POET and estimator (5.5), 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 was somewhat better. The interesting find was that, similar to AFM–DCC-NL, the performance of AFM-NL was 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. (We only tried using $K = 5$ principals components for POET in our empirical analysis; following the recommendations of Fan, Liao, and Mincheva (2013), we did not try using (the first) $K = 1$ principal component only.)

Therefore, one can conclude 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 a desirable property not commonly shared by other factor models.

6 Computational Aspects and 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. To this end, denote by $x_{\cdot i}^T$ the i^{th} column of the matrix X_T and by $\bar{x}_{\cdot i}^T := T^{-1} \sum_t x_{ti}^T$ its mean; denote further by Y_T the $T \times N$ matrix with typical entry $y_{ti}^T := x_{ti}^T - \bar{x}_{\cdot i}^T$. With this notation, 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 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, at least for nonlinear shrinkage.

As a more general situation, consider the setting where the data constitute OLS residuals based on a linear model with K regressors (including a possible constant). In such a setting, T has to be replaced with $T-K$ as the "effective" sample size. (Note that simple demeaning corresponds to using OLS residuals based on a linear model with the constant as the only regressor, in which case $K=1$.)

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 article can be downloaded at www.econ.uzh.ch/en/people/faculty/wolf/publications.html under the header "Programming Code." There also exists an R package nlshrink, available at cran.r-project.org/web/packages/, which implements the QuEST version of nonlinear shrinkage, though note that this package was not written by us.

7 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 article, 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; here, the 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 trade-off: 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 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, at the time, he lacked

the mathematical machinery to solve the problem in a satisfactory fashion. This machinery is called random matrix theory (RMT) 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 a 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 researchers in many fields, particularly in finance, to solve real-life problems. Let they, and time, be our judge.

References

Abadir, K., W. Distaso, and F. Žikešš. 2014. Design-Free Estimation of Variance Matrices. *Journal of Econometrics* 181: 165–180.

Abrahamsson, R., Y. Selen, and P. Stoica. 2007. "Enhanced Covariance Matrix Estimators in Adaptive Beamforming." *2007 IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP*, Honolulu, HI, USA, *volume* II, pp. 969–972, IEEE.

Agrawal, R., U. Roy, and C. Uhler. 2019. Covariance Matrix Estimation under Total Positivity for Portfolio Selection. Preprint arXiv:1909.04222.

Anlauff, J., E. Weitnauer, A. Lehnhardt, S. Schirmer, S. Zehe, and K. Tonekaboni. 2010. "A Method for Outdoor Skateboarding Video Games." *Proceedings of the 7th International Conference on Advances in Computer Entertainment Technology*, Taipei, Taiwan, pp. 40–44, ACM.

Bachega, L. R., J. Theiler, and , C. Bouman et al2011. "Evaluating and Improving Local Hyperspectral Anomaly Detectors." *Applied Imagery Pattern Recognition Workshop (AIPR), Washington DC, USA*, pp. 1–8, IEEE.

Bauwens, L., S. Laurent, and J. V. Rombouts. 2006. Multivariate GARCH Models: A Survey. *Journal of Applied Econometrics* 21: 79–109.

Bell, P., and S. King. 2009. "Diagonal Priors for Full Covariance Speech Recognition." *IEEE Workshop on Automatic Speech Recognition & Understanding, 2009. ASRU 2009*, pp. 113–117. IEEE.

Bickel, P. J., and E. Levina. 2008a. Covariance Regularization by Thresholding. *The Annals of Statistics* 36: 2577–2604.

Bickel, P. J., and E. Levina. 2008b. Regularized Estimation of Large Covariance Matrices. *The Annals of Statistics* 36: 199–227.

Broda, S. A., and M. S. Paolella. 2009. CHICAGO: A Fast and Accurate Method for Portfolio Risk Calculation. *Journal of Financial Econometrics* 7: 412–436.

Chincarini, L. B., and D. Kim. 2006. *Quantitative Equity Portfolio Management: An Active Approach to Portfolio Construction and Management*. New York, NY: McGraw-Hill.

Choi, Y.-G., J. Lim, and S. Choi. 2019. High-Dimensional Markowitz Portfolio Optimization Problem: Empirical Comparison of Covariance Matrix Estimators. *Journal of Statistical Computation and Simulation* 89: 1278–1300.

Connor, G., and R. A. Korajczyk. 1988. Risk and Return in an Equilibrium APT: Application of a New Test Methodology. *Journal of Financial Economics* 21: 255–289.

Connor, G., and R. A. Korajczyk. 1993. A Test for the Number of Factors in an Approximate Factor Model. *The Journal of Finance* 48: 1263–1291.

De Nard, G., O. Ledoit, and M. Wolf. 2021. Factor Models for Portfolio Selection in Large Dimensions: The Good, the Better and the Ugly. *Journal of Financial Econometrics* 19: 236–257.

DeMiguel, V., L. Garlappi, F. J. Nogales, and R. Uppal. 2009. A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms. *Management Science* 55: 798–812.

Dey, D. K., and C. Srinivasan. 1985. Estimation of a Covariance Matrix under Stein's Loss. *The Annals of Statistics* 13: 1581–1591.

Efron, B., and C. Morris. 1973. Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach. *Journal of the American Statistical Association* 68: 117–130.

Efron, B., and C. Morris. 1975. Using Stein's Estimator and Its Generalizations. *Journal of the American Statistical Association* 70: 311–319.

Efron, B., and C. Morris. 1977. Stein's Paradox in Statistics. *Scientific American* 236: 119–127.

Elsheikh, A. H., M. F. Wheeler, and I. Hoteit. 2013. An Iterative Stochastic Ensemble Method for Parameter Estimation of Subsurface Flow Models. *Journal of Computational Physics* 242: 696–714.

Endelman, J. B., and J.-L. Jannink. 2012. Shrinkage Estimation of the Realized Relationship Matrix. *G3: Genes, Genomes, Genetics* 2: 1405–1413.

Engel, J., L. Buydens, and L. Blanchet. 2017. An Overview of Large-Dimensional Covariance and Precision Matrix Estimators with Applications in Chemometrics. *Journal of Chemometrics* 31: e2880.

Engle, R. F. 2002. Dynamic Conditional Correlation—A Simple Class of Multivariate GARCH Models. *Journal of Business & Economic Statistics* 20: 339–350.

Engle, R. F., and R. Colacito. 2006. Testing and Valuing Dynamic Correlations for Asset Allocation. *Journal of Business & Economic Statistics* 24: 238–253.

Engle, R. F., O. Ledoit, and M. Wolf. 2019. Large Dynamic Covariance Matrices. *Journal of Business & Economic Statistics* 37: 363–375.

Fama, E. F., and K. R. French. 2015. A Five Factor Asset Pricing Model. *Journal of Financial Economics* 116: 1–22.

Fan, J., Y. Liao, and M. Mincheva. 2013. Large Covariance Estimation by Thresholding Principal Orthogonal Complements (with Discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75: 603–680.

Frost, P. A., and J. E. Savarino. 1986. An Empirical Bayes Approach to Portfolio Selection. *The Journal of Financial and Quantitative Analysis* 21: 293–305.

Guo, S.-M., J. He, N. Monnier, G. Sun, T. Wohland, and M. Bathe. 2012. Bayesian Approach to the Analysis of Fluorescence Correlation Spectroscopy Data II: Application to Simulated and in Vitro Data. *Analytical Chemistry* 84: 3880–3888.

Haff, L. R. 1980. Empirical Bayes Estimation of the Multivariate Normal Covariance Matrix. *The 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 R. Jagannathan. 1997. Assessing Specification Errors in Stochastic Discount Factor Models. *The Journal of Finance* 52: 557–590.

Haufe, S., M. S. Treder, M. F. Gugler, M. Sagebaum, G. Curio, and B. Blankertz. 2011. EEG Potentials Predict Upcoming Emergency Brakings During Simulated Driving. *Journal of Neural Engineering* 8: 056001.

James, W., and C. Stein. 1961. "Estimation with Quadratic Loss." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Oakland, CA, USA: University of California Press, pp. 361–380.

Jing, B.-Y., G. Pan, Q.-M. Shao, and W. Zhou. 2010. Nonparametric Estimate of Spectral Density Functions of Sample Covariance Matrices: A First Step. *The Annals of Statistics* 38: 3724–3750.

Korniotis, G. M. 2008. Habit Formation, Incomplete Markets, and the Significance of Regional Risk for Expected Returns. *Review of Financial Studies* 21: 2139–2172.

Lam, C. 2016. Nonparametric Eigenvalue-Regularized Precision or Covariance Matrix Estimator. *The Annals of Statistics* 44: 928–953.

Ledoit, O. 1995. Essays on Risk and Return in the Stock Market. PhD thesis, Massachusetts Institute of Technology, Sloan School of Management. Available at http://dspace.mit.edu/handle/1721.1/11875 (accessed April 2020).

Ledoit, O., and S. Péché. 2011. Eigenvectors of Some Large Sample Covariance Matrix Ensembles. *Probability Theory and Related Fields* 150: 233–264.

Ledoit, O., and M. Wolf. 2003. Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection. *Journal of Empirical Finance* 10: 603–621.

Ledoit, O., and M. Wolf. 2004a. Honey, I Shrunk the Sample Covariance Matrix. *The Journal of Portfolio Management* 30: 110–119.

Ledoit, O., and M. Wolf. 2004b. A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *Journal of Multivariate Analysis* 88: 365–411.

Ledoit, O., and M. Wolf. 2012. Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices. *The Annals of Statistics* 40: 1024–1060.

Ledoit, O., and M. Wolf. 2015. Spectrum Estimation: A Unified Framework for Covariance Matrix Estimation and PCA in Large Dimensions. *Journal of Multivariate Analysis* 139: 360–384.

Ledoit, O., and M. Wolf. 2017a. Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks. *The Review of Financial Studies* 30: 4349–4388.

Ledoit, O., and M. Wolf. 2017b. Numerical Implementation of the QuEST Function. *Computational Statistics & Data Analysis* 115: 199–223.

Ledoit, O., and M. Wolf. 2018. Optimal Estimation of a Large-Dimensional Covariance Matrix under Stein's Loss. *Bernoulli* 24: 3791–3832.

Ledoit, O., and M. Wolf. 2020. Analytical Nonlinear Shrinkage of Large-Dimensional Covariance Matrices. *Annals of Statistics* (in press).

Ledoit, O., M. Wolf, and Z. Zhao. 2019. Efficient Sorting: A More Powerful Test for Cross-Sectional Anomalies. *Journal of Financial Econometrics* 17: 645–686.

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: 13–37.

Lotte, F., and C. Guan. 2009. "An Efficient p300-Based Brain-Computer Interface with Minimal Calibration Time." *Assistive Machine Learning for People with Disabilities Symposium (NIPS'09 Symposium), Vancouver, Canada, December 2009*.

Marčenko, V. A., and L. A. Pastur. 1967. Distribution of Eigenvalues for Some Sets of Random Matrices. *Mathematics of the Ussr-Sbornik* 1: 457–483.

Markon, K. 2010. Modeling Psychopathology Structure: A Symptom-Level Analysis of Axis I and II Disorders. *Psychological Medicine* 40: 273–288.

Meucci, A. 2005. *Risk and Asset Allocation*. Berlin Heidelberg and New York, NY: Springer Finance.

Michaelides, P., P. Apostolellis, and S. Fassois. 2011. Vibration-Based Damage Diagnosis in a Laboratory Cable–Stayed Bridge Model via an RCP–ARX Model Based Method. *Journal of Physics: Conference Series* 305: 012104.

Moura, G. V., A. P. Santos, and E. Ruiz. 2019. "Comparing Forecasts of Extremely Large Conditional Covariance Matrices." Statistics and Econometrics Working Paper 2019, Universidad Carlos III de Madrid. Available at https://e-archivo.uc3m.es/handle/10016/14 (accessed April 2020).

Pakel, C., N. Shephard, K. Sheppard, and R. F. Engle. 2020. Fitting Vast Dimensional Time-Varying Covariance Models. *Journal of Business & Economic Statistics*; doi: 10.1080/07350015.2020.1713795.

Pirkl, R. J., K. A., Remley, and C. S., Lötbäck Patané, 2012. Reverberation Chamber Measurement Correlation. *IEEE Transactions on Electromagnetic Compatibility* 54: 533–545.

Pourahmadi, M. 2013. *High-Dimensional Covariance Estimation*. Hoboken: John Wiley & Sons.

Pyeon, D., M. A. Newton, P. F. Lambert, J. A. Den Boon, S. Sengupta, C. J. Marsit, C. D. Woodworth 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: 4605–4619.

Ren, Y., and K. Shimotsu. 2009. Improvement in Finite-Sample Properties of the Hansen-Jagannathan Distance Test. *Journal of Empirical Finance* 16: 483–506.

Ribes, A., S. Planton, and L. Terray. 2013. Application of Regularised Optimal Fingerprinting to Attribution. Part I: Method, Properties and Idealised Analysis. *Climate Dynamics* 41: 2817–2836.

Robinson, B. D. 2019. "Optimal Rotation-Equivariant Covariance Estimation for Detection of High-Dimensional Signals." *2019 IEEE Radar Conference (RadarConf)*, Boston, MA, USA, pp. 1–6, IEEE.

Ross, S. A. 1976. The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory* 13: 341–360.

Schäfer, J., and K. Strimmer. 2005. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology* 4.

Schain, M., F. Zanderigo, and R. T. Ogden. 2018. Likelihood Estimation of Drug Occupancy for Brain PET Studies. *Neuroimage* 178: 255–265.

Sharpe, W. F. 1964. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance* 19: 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.

Stein, C. 1956. "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution." *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA, pp. 197–206, University of California Press.

Stein, C. 1975. "Estimation of a Covariance Matrix." *Rietz Lecture, 39th Annual Meeting IMS*, Atlanta, Georgia, 28 August 1975.

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 Soviet Mathematics* 34: 1373–1403.

Trucios, C., M. Zevallos, L. K. Hotta, and A. P. Santos. 2019. Covariance Prediction in Large Portfolio Allocation. *Econometrics* 7: 19.

Wigner, E. P. 1955. Characteristic Vectors of Bordered Matrices with Infinite Dimensions. *The Annals of Mathematics* 62: 548–564.

Wolf, M., and D. Wunderli. 2012. "Fund-of-Funds Construction by Statistical Multiple Testing Methods." In B. Scherer and K. Winston (eds.), *The Oxford Handbook of Quantitative Asset Management*, pp. 116–135. Oxford: Oxford University Press.

Zhang, Y., D. Sun, and D. Zhang. 2009. Robust Adaptive Acoustic Vector Sensor Beamforming Using Automated Diagonal Loading. *Applied Acoustics* 70: 1029–1033.

Zumbach, G. O. 2007. The RiskMetrics 2006 Methodology. Technical report, RiskMetrics Group. Available at https://ssrn.com/abstract=1420185 (accessed April 2020).