University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 231

Large Dynamic Covariance Matrices

Robert F. Engle, Olivier Ledoit and Michael Wolf

July 2016

Large Dynamic Covariance Matrices*

Robert F. Engle
Department of Finance
New York University
New York, NY 10012, USA
rengle@stern.nyu.edu

Olivier Ledoit
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
olivier.ledoit@econ.uzh.ch

Michael Wolf
Department of Economics
University of Zurich
CH-8032 Zurich, Switzerland
michael.wolf@econ.uzh.ch

July 2016

Abstract

Second moments of asset returns are important for risk management and portfolio selection. The problem of estimating second moments can be approached from two angles: time series and the cross-section. In time series, the key is to account for conditional heteroskedasticity; a favored model is Dynamic Conditional Correlation (DCC), derived from the ARCH/GARCH family started by Engle (1982). In the cross-section, the key is to correct in-sample biases of sample covariance matrix eigenvalues; a favored model is nonlinear shrinkage, derived from Random Matrix Theory (RMT). The present paper aims to marry these two strands of literature in order to deliver improved estimation of large dynamic covariance matrices.

KEY WORDS: Composite likelihood, dynamic conditional correlations, GARCH, Markowitz portfolio selection, nonlinear shrinkage.

JEL CLASSIFICATION NOS: C13, C58, G11.

*The authors would like to thank Zhao Zhao (Department of Economics, Huazhong University of Science and Technology, China) for providing research assistance. The authors would also like to thank Kevin Sheppard for having made publicly available the UCSD GARCH Toolbox as well as its successor, the Oxford MFE Toolbox. Any errors are ours.

1 Introduction

Multivariate GARCH models derived from the ARCH/GARCH family started by Engle (1982) are popular tools for risk management and portfolio selection. However, the number of assets in the investment universe generally poses a challenge to such models. When this number is large, say on the order of a thousand, many multivariate GARCH models exhibit unsatisfactory performance or cannot even be estimated in the first place due to computational problems. In other words, many multivariate GARCH models suffer from the curse of dimensionality.

The aim of this paper is to robustify the Dynamic Conditional Correlation (DCC) model originally proposed by Engle (2002) against large dimensions. To this end we combine two tools. The first tool is the composite likelihood method of Pakel et al. (2014) which makes the estimation of a DCC model in large dimensions computationally feasible: Composite likelihood ensures that DCC can be used in the first place when the number of assets is large. The second tool is the nonlinear shrinkage method of Ledoit and Wolf (2012) which results in improved estimation of the correlation targeting matrix of a DCC model: Nonlinear shrinkage ensures that DCC performs well when the number of assets is large.

The aim of this paper, on the other hand, is not to carry out an extensive comparison of our proposal to all other multivariate GARCH models in the literature. Such a comparison is beyond the scope of the paper and left to future research.

The remainder of the paper is organized as follows. Section 2 gives a brief description of the DCC model including the composite likelihood method. Section 3 gives a description of the nonlinear shrinkage method. Section 4 details our loss function which is custom-tailored to the problem of portfolio selection. Section 5 contains Monte Carlo simulations while Section 6 provides backtesting results based on real-life stock return data. Section 7 concludes. Appendix A provides additional details and motivation regarding the nonlinear shrinkage method. Appendix B gives an extension of our approach to the BEKK model presented in Engle and Kroner (1995).

2 The DCC Model

Our exposition of the DCC model is primarily based on the work of Engle (2002) and Engle (2009, Section 11.2).

2.1 Notation

In what follows, the subscript i indexes the variables and covers the range of integers from 1 to N, where N denotes the dimension of the covariance matrix. The subscript t indexes the dates and covers the range of integers from 1 to T, where T denotes the sample size. The notation $\mathsf{Diag}(\cdot)$ represents the function that sets to zero all the off-diagonal elements of

a matrix.

- $r_{i,t}$: observed data series for variable i at date t, stacked into $\mathbf{r}_t := (r_{1,t}, \ldots, r_{N,t})'$
- $d_{i,t}^2 := \mathsf{Var}(r_{i,t}|\mathcal{F}_{t-1})$: conditional variance of the i^{th} variable at date t
- D_t is the N-dimensional diagonal matrix whose i^{th} diagonal element is $d_{i,t}$
- $H_t := \mathsf{Cov}(\mathbf{r}_t|\mathcal{F}_{t-1})$: conditional covariance matrix at date t; thus $\mathsf{Diag}(H_t) = D_t^2$
- $s_{i,t} := r_{i,t}/d_{i,t}$: devolatilized series, stacked into $\mathbf{s}_t := (s_{1,t}, \ldots, s_{N,t})'$
- $R_t := \mathsf{Corr}(\mathbf{r}_t|\mathcal{F}_{t-1}) = \mathsf{Cov}(\mathbf{s}_t|\mathcal{F}_{t-1})$: conditional correlation matrix at date t
- $\sigma_i^2 := \mathsf{E}(d_{i,t}^2) = \mathsf{Var}(r_{i,t})$: unconditional variance of the i^{th} series
- $C := \mathsf{E}(R_t) = \mathsf{Corr}(\mathbf{r}_t) = \mathsf{Cov}(\mathbf{s}_t)$: unconditional correlation matrix

2.2 Model Definition

This exposition is not meant to review the full generality of the DCC family, but to describe a representative member. Certain specific choices have been made for simplicity, tractability, and scalability. For the dynamics of the univariate volatilities, we use a GARCH(1,1) process:

$$d_{i,t}^2 = \omega_i + a_i r_{i,t-1}^2 + b_i d_{i,t-1}^2 \ , \tag{2.1}$$

where (ω_i, a_i, b_i) are the variable-specific GARCH(1,1) parameters.

Remark 2.1. We use the standard GARCH(1,1) specification here for simplicity. However, it is possible to upgrade to more sophisticated GARCH-type models, for example models incorporating asymmetry effects. ■

The square roots $d_{i,t}$ of the conditional variances go into the diagonal of the matrix D_t. These N separate univariate models also yield the vector of devolatilized residuals $\mathbf{s}_t = (r_{1,t}/d_{1,t}, \ldots, r_{N,t}/d_{N,t})'$.

We assume that the evolution of the correlation matrix over time is governed by the DCC model with correlation targeting. This is an adaptation of the variance targeting idea of Engle and Mezrich (1996); see equation (11.7) of Engle (2009). In our notation, it is expressed as

$$Q_t = (1 - \alpha - \beta)C + \alpha\, \mathbf{s}_{t-1}\mathbf{s}_{t-1}' + \beta\, Q_{t-1} \ , \tag{2.2}$$

where (α, β) are the DCC parameters analogous to (a_i, b_i), but in correlation space instead of variance space. 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 elements, although close to one, are not exactly equal to one. From this representation we obtain the conditional correlation matrix and the conditional covariance matrix as

$$R_t = \mathsf{Diag}(Q_t)^{-1/2}\, Q_t\, \mathsf{Diag}(Q_t)^{-1/2} \tag{2.3}$$

$$H_t = D_t R_t D_t \ , \tag{2.4}$$

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

$$\mathbf{r}_t | \mathcal{F}_{t-1} \sim \mathcal{N}(0, H_t) \ . \tag{2.5}$$

2.3 Estimation in Large Dimensions

It is well known that estimating the DCC model with a large number of assets is challenging. One of the difficulties is to invert the conditional correlation matrix. Fortunately, a way to overcome this hurdle has been found by Pakel et al. (2014). The basic idea is that, instead of inverting the full conditional correlation matrix, it is easier to look at a selection of two-by-two blocks. They call it the *composite likelihood* method. The composite likelihood is computed by summing up the quasi-likelihoods of pairs of assets.

Of the different variants proposed by the authors, the most scalable one is the 2MSCLE method, which is the composite likelihood estimator based on all contiguous pairs. Thus, there are only $N-1$ bivariate quasi-likelihoods to compute. Pakel et al. (2014) show that maximizing the composite likelihood thus constructed yields consistent estimators of the two correlation dynamics parameters α and β.

To summarize, the estimation unfolds as a three-stage process:

1. For each asset, fit a univariate GARCH(1,1) model and use the fitted model to devolatilize the return series.
2. Estimate the unconditional correlation matrix and use it for correlation targeting.
3. Maximize the composite likelihood to estimate correlation dynamics.

The focus of the present paper, starting with the next section, is to improve the second step.

3 Large Unconditional Correlation Matrices

The critical juncture where the time-series approach of the DCC model meets the cross-sectional results of Random Matrix Theory (RMT) is in the estimation of the unconditional correlation matrix C, which serves as the target or intercept of the ARCH/GARCH dynamics in equation (2.2).

3.1 Sample Correlation Matrix

It is widely acknowledged that the sample correlation matrix works poorly in large dimensions. This observation goes back at least to Michaud (1989), who famously described portfolio

optimization as "error maximization". The sample correlation (and covariance) matrix is a good fit in-sample but it suffers from overfitting, so it underperforms out-of-sample. Portfolio managers invest out-of-sample. Therefore, the sample correlation matrix should be shunned for decision making.

The reason is that the sample correlation matrix has $N(N-1)/2$ parameters, and the data set has $N \times T$ noisy observations. When N is of the same order of magnitude as T, these two quantities are similar-sized. It is not possible to estimate accurately $O(N^2)$ parameters from $O(N^2)$ noisy data points. This is the curse of dimensionality in action.

RMT teaches us that the ratio N/T, generally called the *concentration (ratio)*, is the determinant factor. If the concentration is small, say less than 1/100, then standard (fixed-dimensional) asymptotics represent a good approximation of the true behavior, and the sample correlation matrix can be trusted out of sample. As the concentration gets higher, special methods need to be employed to address the issue of in-sample overfitting due to the excessive number of free parameters.

The most egregious and easily understood example is when the concentration ratio $c := N/T$ exceeds one, which means that there are more assets than time-series observations, so the sample correlation matrix is singular. In this case, inverting it blows up optimal portfolio weights to plus or minus infinity, which obviously is an economically incorrect solution. But by continuity, as N gets closer to T, problems start to creep in. Indeed, the main lesson of RMT is that, unless N/T is negligible, the sample correlation (and covariance) matrix will systematically misbehave out of sample.

3.2 Shrinkage Estimator

Fortunately, recent developments have provided effective solutions to this problem. It is now possible to rectify the in-sample bias of the sample correlation (or covariance) matrix due to overfitting. This rectification takes place at the level of the eigenvalues. The small sample eigenvalues are too small and the large ones too large. So it is just a matter of pushing up the small ones and pulling down the large ones. Since this transformation reduces the spread of the cross-sectional distribution of eigenvalues, it is generally called *shrinkage*.

In this paper, we will focus on the nonlinear shrinkage formula of Ledoit and Wolf (2012). The intuition is as follows. Let Σ denote the population covariance matrix, S the sample covariance matrix, and u an eigenvector of S. Then by basic linear algebra, the corresponding sample eigenvalue is equal to $u'Su$. It is the in-sample variance of a portfolio with weights given by the vector u. This is the quantity that needs to be rectified due to overfitting. Nonlinear shrinkage replaces it with (a consistent estimator of) $u'\Sigma u$, the out-of-sample variance of the same portfolio. Clearly, we want to decide our portfolio allocation in the direction of the vector u based on its true out-of-sample risk $u'\Sigma u$, rather than its in-sample counterpart $u'Su$, which is heavily biased due to the curse of dimensionality.

This may seem like magic: given that we do not know Σ, how could we know $u'\Sigma u$? However, Ledoit and Péché (2011) show that we do not need to know *all* of the true covariance matrix Σ — which would be a hopeless task — but only its eigenvalues. There are just N of them. Given a data set of dimension $N \times T$, it is clearly impossible to estimate accurately a whole matrix of dimension $N \times N$, but it is theoretically possible to estimate its N eigenvalues. The ratio of parameters to data entries is $1/T$, which is a small number, regardless of how big the matrix dimension N is.

Recovering the population eigenvalues from the sample eigenvalues requires inverting the Marčenko and Pastur (1967) equation, which governs their asymptotic relationship when the dimension is large. El Karoui (2008) and Mestre (2008) were the first to make an attempt in this direction. The solutions they proposed suffered from some limitations that made them unsuitable for general use. Subsequently Ledoit and Wolf (2015) introduced an effective method based on numerical inversion of what they call the QuEST function. This acronym stands for Quantized Eigenvalues Sampling Transform. It is a deterministic N-dimensional function that discretizes the Marčenko-Pastur equation and lends itself to numerical inversion. This is the technology that we will use here.

The basic idea of the paper is to use the nonlinear shrinkage estimator of Ledoit and Wolf (2012) to estimate the target (or intercept) correlation matrix C in equation (2.2) instead of the sample correlation matrix. When the dimension N is large and of the same order of magnitude as the sample size T, this approach is expected to generate an estimator of the conditional covariance matrix H_t that has better out-of-sample properties.

3.3 Mathematical Formulation

Let $S \coloneqq [s_{i,t}]$ denote the $N \times T$ matrix of devolatilized returns. Assuming mean zero, its sample covariance matrix is

$$\widehat{C} \coloneqq \frac{1}{T} SS'. \tag{3.1}$$

Decompose $\widehat{C}$ it into a set of eigenvalues $(\lambda_1, \lambda_2 \ldots, \lambda_N)$, sorted in descending order without loss of generality, and corresponding eigenvectors $(u_1, u_2, \ldots, u_N)$; consequently

$$\widehat{C} = \sum_{i=1}^{N} \lambda_i u_i u_i'. \tag{3.2}$$

Let $Q_{N,T}$ denote the QuEST function defined in Section 2.2 of Ledoit and Wolf (2015). It is a multivariate deterministic function that maps $[0, \infty)^N$ onto itself. Given a set of population eigenvalues $\mathbf{t} \coloneqq (t_1, \ldots, t_N)$ as input, it returns as output a deterministic equivalent of the sample eigenvalues $Q_{N,T}(\mathbf{t}) = \left(q^1_{N,T}(\mathbf{t}), q^2_{N,T}(\mathbf{t}), \ldots, q^N_{N,T}(\mathbf{t})\right)$. Thus, population eigenvalues can be estimated by numerically inverting the QuEST function:

$$\widetilde{\boldsymbol{\tau}} \coloneqq \underset{\mathbf{t} \in [0,\infty)^N}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left[q^i_{N,T}(\mathbf{t}) - \lambda_i \right]^2. \tag{3.3}$$

Given estimated population eigenvalues $\widetilde{\boldsymbol{\tau}} = (\widetilde{\tau}_1, \widetilde{\tau}_2, \ldots, \widetilde{\tau}_N)$, we can then use Theorem 4 of Ledoit and Péché (2011) to compute a nonlinear shrinkage formula that is *asymptotically* optimal under large-dimensional asymptotics. Denote the shrunk eigenvalues by $\widetilde{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\tau}}) := \big(\widetilde{\lambda}_1(\widetilde{\boldsymbol{\tau}}), \widetilde{\lambda}_2(\widetilde{\boldsymbol{\tau}}), \ldots, \widetilde{\lambda}_N(\widetilde{\boldsymbol{\tau}})\big)$.

The mathematical expressions for the QuEST function and the nonlinear shrinkage formula are quite involved, as they depend on the Stieltjes transform, an integral transform defined on the upper half of the complex plane, so they will not be repeated here. The way to intuitively understand $\widetilde{\lambda}_i(\widetilde{\boldsymbol{\tau}})$ is that it is a consistent estimator for the out-of-sample variance $u_i'\Sigma u_i$ under large-dimensional asymptotics. The shrinkage estimator of the covariance matrix can then be reconstructed as

$$\widetilde{C} := \sum_{i=1}^{N} \widetilde{\lambda}_i(\widetilde{\boldsymbol{\tau}}) u_i u_i'. \tag{3.4}$$

Two important advantages of this approach are (i) that it does not require the assumption of normality and (ii) that it can handle the challenging case where the number of assets exceeds the sample size.

Appendix A contains a primer on this technology imported from Probability Theory and from Multivariate Statistics; see Ledoit and Wolf (2014) for a more rigorous exposition.

3.4 Linear Shrinkage

A simpler alternative is to use the linear shrinkage formula of Ledoit and Wolf (2004). Define the cross-sectional average of sample eigenvalues as

$$\overline{\lambda} := \frac{1}{N} \sum_{i=1}^{N} \lambda_i \ . \tag{3.5}$$

This method provides a consistent estimator ρ for the optimal *shrinkage intensity*, which is a number between zero and one controlling the amount by which the sample eigenvalues are dragged towards their cross-sectional average $\overline{\lambda}$. The linear shrinkage estimator is expressed as

$$\overline{C} := \sum_{i=1}^{N} \left[\rho\overline{\lambda} + (1-\rho)\lambda_i\right] u_i u_i' \ . \tag{3.6}$$

This estimator has proven quite popular with applied researchers but it is optimal only in a class of dimension two, which is nested inside the class of dimension N over which nonlinear shrinkage is optimal. Therefore, asymptotically, when N and T grow large together, nonlinear shrinkage should perform better in the generic case. The main difference is that the shrinkage intensity ρ is constrained to be the same for all eigenvalues under linear shrinkage, whereas it is *individually fitted* to each eigenvalue under nonlinear shrinkage. This approach is obviously more complicated but due to the highly nonlinear nature of the problem, it is also more powerful. Haffner and Reznkikova (2012) propose the application of linear shrinkage to the estimation

of DCC models. However, they restrict themselves to smaller dimensions and throw shrinkage into a horse race against composite likelihood, instead of harnessing the combined powers of the two methods. In addition, there is theoretical justification for upgrading from linear to nonlinear shrinkage (Ledoit and Wolf, 2012).

3.5 Renormalization

In practice, the diagonal elements of $\widehat{C}$, $\widetilde{C}$, and $\overline{C}$ tend to deviate from one slightly, in spite of the fact that devolatilized returns are used as inputs. Therefore every column and every row has to be divided by the square root of the corresponding diagonal entry, so as to produce a proper correlation matrix.

4 Loss Function

This section builds upon Engle and Colacito (2006), hereafter abbreviated by EC. It is couched in terms of the unconditional covariance matrix initially, and the adaptation to conditional covariance matrices is described in Section 4.4. To make the exposition more fluid, the word "return" stands for the raw return on a risky asset minus the risk-free interest rate.

4.1 Out-of-Sample Portfolio Variance

Let Σ denote the N-dimensional covariance matrix of returns, $\widehat{\Sigma}$ a generic estimator of Σ, and m an assumed vector of expected returns. EC's equation (3) gives the out-of-sample variance of the optimal portfolio based on the estimator $\widehat{\Sigma}$ as

$$\mathcal{L}_V(\widehat{\Sigma}, \Sigma, m) := \frac{m'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}m}{\left(m'\widehat{\Sigma}^{-1}m\right)^2} \ . \tag{4.1}$$

This loss function corresponds to the quintessential risk or error minimization objective. It was also adopted by Ledoit and Wolf (2014, Definition 2.1), up to rescaling by the squared Euclidian norm of m. In addition, it captures the performance of a covariance matrix estimator for mathematically equivalent problems where variance minimization decisions must be taken, such as Capon (1969) beamforming in signal processing and optimal fingerprinting in climatology (IPCC, 2007).

EC's Theorem 1 demonstrates that $\mathcal{L}_V$ is minimized when $\widehat{\Sigma} = \Sigma$, so if we want a loss function that is always nonnegative, and is equal to zero when the covariance matrix estimator happens to be equal to the truth — as is customary —, then we can compute the excess portfolio variance caused by covariance matrix estimation error:

$$\mathcal{L}_E(\widehat{\Sigma}, \Sigma, m) := \mathcal{L}_V(\widehat{\Sigma}, \Sigma, m) - \mathcal{L}_V(\Sigma, \Sigma, m) = \frac{m'\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}m}{\left(m'\widehat{\Sigma}^{-1}m\right)^2} - \frac{1}{m'\Sigma^{-1}m} \geq 0 \ . \tag{4.2}$$

4.2 Expected Returns

The vector of expected returns that we have assumed is not required to be equal to the true one. Different investors will have different models of expected returns. This suggests investigating over a wide range of alternatives for m. The idea is to integrate the excess portfolio variance $\mathcal{L}_E$ over a relevant manifold of expected return vectors.

The relative efficiency of two covariance matrix estimators is unaffected by the norm of the vector m, so we can take $\|m\| = 1$ without loss of generality, where $\|\cdot\|$ denotes the Euclidian norm. For this reason, in dimension 2, EC's Section 2.1 considers expected returns of the form $m = [\cos(\theta), \sin(\theta)]'$ where θ is some angle. The generalization in dimension N is to consider vectors m that belong to the N-dimensional unit sphere $\mathcal{U}_N := \{x \in \mathbb{R}^N : \|x\| = 1\}$. Averaging out the excess portfolio variance across all possible m's in this manifold leads to a loss function that depends solely on covariance matrices:

$$\mathcal{L}_I(\widehat{\Sigma}, \Sigma) := \int_{\mathcal{U}_N} \mathcal{L}_E(\widehat{\Sigma}, \Sigma, \mu)\, d\mu = \int_{\mathcal{U}_N} \left[\frac{\mu' \widehat{\Sigma}^{-1} \Sigma \widehat{\Sigma}^{-1} \mu}{\left(\mu' \widehat{\Sigma}^{-1} \mu\right)^2} - \frac{1}{\mu' \Sigma^{-1} \mu} \right] d\mu \ , \tag{4.3}$$

where $\int_{\mathcal{U}_N}$ denotes the integral over the N-dimensional unit sphere. The intuition is that we want a covariance matrix estimator that is the best "all-rounder" and thus performs well across the board. Given that this paper focuses explicitly on the second moments, we are not in a position to take a stance on the orientation of the linear constraint vector.

4.3 Equivalent Formulation Under Large-Dimensional Asymptotics

The integral in $\mathcal{L}_I$ is not easy to evaluate analytically. At this juncture, we can get help from a foundational lemma in Random Matrix Theory (RMT):

$$x' A x \approx \frac{\mathsf{Tr}(A)\, \|x\|^2}{N} \ , \tag{4.4}$$

where $\mathsf{Tr}(\cdot)$ denotes the trace of a matrix, A is some symmetric random matrix of large dimension N, and x is an N-dimensional vector not statistically related to A.[1] Rigorous versions of equation (4.4) appear decisively as Lemma 1 of Marčenko and Pastur (1967), Lemma 3.1 of Silverstein and Bai (1995), and Lemma 1 of Ledoit and Péché (2011). This is basically a cross-sectional law of large numbers. Indeed if, conditional on A, x follows the standard multivariate normal distribution, then even in finite samples the relation

$$\mathsf{E}\left[x' A x \middle| A \right] = \mathsf{E}\left[\frac{\mathsf{Tr}(A)\, \|x\|^2}{N} \middle| A \right] \tag{4.5}$$

holds exactly. Injecting (4.4) into the loss function $\mathcal{L}_I$ naturally suggests the alternative loss function $\mathcal{L}$, which is significantly easier to evaluate.

[1] x is distributed independently of A and its distribution is rotation-invariant.

Definition 4.1. *The loss function is defined as*

$$\mathcal{L}\big(\widehat{\Sigma},\Sigma\big) := \frac{\mathsf{Tr}\big(\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}\big)/N}{\Big[\mathsf{Tr}\big(\widehat{\Sigma}^{-1}\big)/N\Big]^2} - \frac{1}{\mathsf{Tr}\big(\Sigma^{-1}\big)/N} \ . \tag{4.6}$$

This is the loss function that will be used throughout the rest of the paper. The formulations $\mathcal{L}_I$ and $\mathcal{L}$ are interchangeable under large-dimensional asymptotics, as the following proposition attests.

Proposition 4.1. *Under the assumptions of Theorem 3.1 of Ledoit and Wolf (2014), both loss functions $\mathcal{L}_I\big(\widehat{\Sigma},\Sigma\big)$ and $\mathcal{L}\big(\widehat{\Sigma},\Sigma\big)$ converge a.s. to the same non-stochastic limit.*

Proof. Let m be a random vector distributed according to the N-dimensional standard multivariate normal distribution, independently from $\widehat{\Sigma}$. Then $m/\|m\|$ is uniformly distributed on $\mathcal{U}_N$. This implies that

$$\mathcal{L}_I(\widehat{\Sigma},\Sigma) = \mathsf{E}\left[\mathcal{L}_E\left(\widehat{\Sigma},\Sigma,m/\|m\|\right)\middle|\,\widehat{\Sigma}\right] \ . \tag{4.7}$$

Proposition 4.1 then follows directly by injecting Lemma 1 of Ledoit and Péché (2011) into the proof of Theorem 3.1 of Ledoit and Wolf (2014). ■

A traditional property of a loss function is that if one plugs the population parameters into the loss function, the value of the loss is zero. By convention, there is no loss from estimation if somehow you happen to know the truth; the loss should only be strictly positive if the estimator has error in it. The following theorem, which is equivalent in spirit to EC's Theorem 1, shows that the loss function in Definition 4.1 possesses the desired property.

Theorem 4.1. $\mathcal{L}\big(\widehat{\Sigma},\Sigma\big) \geq 0$*; and* $\mathcal{L}\big(\widehat{\Sigma},\Sigma\big) = 0 \iff \exists\,\gamma > 0$ *s.t.* $\widehat{\Sigma} = \gamma\Sigma$.

Proof. It is a standard result in linear algebra that, for any two symmetric positive-definite matrices $\widehat{\Sigma}$ and Σ, there exists an invertible matrix A such that

$$A'\Sigma A = \mathbb{I}_N \tag{4.8}$$

$$A'\widehat{\Sigma}A = \Delta \ , \tag{4.9}$$

where the superscript $'$ denotes the transpose of a matrix, $\mathbb{I}_N$ denotes the identity matrix of dimension N, and Δ denotes some diagonal matrix with strictly positive diagonal elements. Substituting equations (4.8)–(4.9) into the definition (4.6) yields

$$\mathcal{L}\big(\widehat{\Sigma},\Sigma\big) \geq 0 \iff \mathsf{Tr}\big(\Delta^{-2}A'A\big) \times \mathsf{Tr}\big(A'A\big) \geq \big[\mathsf{Tr}\big(\Delta^{-1}A'A\big)\big]^2 \ . \tag{4.10}$$

Denote the i^{th} diagonal element of Δ by δ_i and the i^{th} diagonal element of $A'A$ by α_i, for $i = 1,\ldots,N$. Note that $\forall i = 1,\ldots,N$, $\alpha_i > 0$ and $\delta_i > 0$. Then,

$$\mathcal{L}\Big(\widehat{\Sigma},\Sigma\Big) \geq 0 \iff \left(\sum_{i=1}^{N}\frac{\alpha_i}{\delta_i^2}\right)\left(\sum_{i=1}^{N}\alpha_i\right) \geq \left(\sum_{i=1}^{N}\frac{\alpha_i}{\delta_i}\right)^2 . \tag{4.11}$$

Define $x_i := \sqrt{\alpha_i}/\delta_i$ and $y_i := \sqrt{\alpha_i}$, for $i = 1, \ldots, N$. Then

$$\mathcal{L}(\widehat{\Sigma}, \Sigma) \geq 0 \iff \left(\sum_{i=1}^{N} x_i^2\right)\left(\sum_{i=1}^{N} y_i^2\right) \geq \left(\sum_{i=1}^{N} x_i y_i\right)^2. \tag{4.12}$$

On the right-hand side of the equivalency we recognize the Cauchy-Schwarz inequality; therefore, the loss function is always non-negative. For equality to hold, we need the vectors $(x_1, \ldots, x_N)'$ and $(y_1, \ldots, y_N)'$ to be collinear, which implies that the diagonal matrix Δ is a scalar multiple of the identity. Inspecting equations (4.8)–(4.9) reveals that this can only happen if the estimator $\widehat{\Sigma}$ is a scalar multiple of the true covariance matrix Σ. ■

4.4 Adaptation for Conditional Covariance Matrices

Given that the loss function from Definition 4.1 is suitable for unconditional covariance matrices, it must be adapted for conditional ones. Let $\widehat{H}_t$ denote a generic estimator of the true conditional covariance matrix H_t (for $t = 1, \ldots, T$). Then the average loss is given by

$$\widehat{L} := \frac{1}{T}\sum_{i=1}^{T} \mathcal{L}(\widehat{H}_t, H_t). \tag{4.13}$$

This is the quantity that will be reported in all Monte Carlo simulations below.

5 Monte Carlo Simulations

5.1 Overall Setup

In order to generate realistic Monte Carlo simulations, we estimate the unconditional population covariance matrix from the $N \in \{100, 500, 1000\}$ most liquid stocks in the CRSP database using ten years of daily data from 2005 through 2014. This matrix will be taken as the 'true' unconditional covariance matrix in all the simulations.

We then simulate a DCC model with parameters $\alpha = 0.05$ and $\beta = 0.93$ as in Table 4 of Pakel et al. (2014). The variates are drawn either from a multivariate standard normal distribution, or from a multivariate "Student" t-distribution with 5 degrees of freedom rescaled so that all individual variances are equal to one. The univariate volatility dynamics are GARCH(1,1) models with identical parameters $a_i = 0.05$ and $b_i = 0.90$ across all stocks $i = 1, \ldots, N$.

For each simulation, we generate an $N \times T$ matrix of simulated returns, where the sample size is $T = 1250$. This sample size corresponds to approximately five years of daily data, an estimation window which is commonly used in practice. Thus, the concentration ratio for the broadest universe is $c := N/T = 0.8$ which is not negligible, and we can expect substantial shrinkage effects. Even though we could potentially accomodate the case $c > 1$, we chose a value of c less than one so that the sample correlation matrix can be used as a benchmark.

In any case, having a universe of 1000 assets in which to invest is sufficiently broad for many if not most portfolio managers.

5.2 Numerical Details

The nonlinear shrinkage estimator of the unconditional correlation matrix was computed in Matlab using version 025 of the QuEST package. It contains the QuEST function itself, a routine to invert it numerically by using a nonlinear optimizer, and another routine to compute the optimal nonlinear shrinkage formula. The end user does not have to do anything except input the raw data and collect the shrinkage estimator. The linear shrinkage estimator was also computed in Matlab using code that implements the formula of Ledoit and Wolf (2004). All codes are available from the university faculty webpage of one of the authors.[2]

The program to simulate the DCC model was adapted from Kevin Sheppard's legacy UCSD GARCH Toolbox.[3] The program to estimate the DCC model was based on the successor to the UCSD GARCH Toolbox, which is the Oxford MFE Toolbox, also by Kevin Sheppard.[4]

We made two important modifications to the latter. The first one was to add the flexibility to shrink the eigenvalues of the sample correlation matrix used for correlation targeting as delineated in Section 3. The second was to rewrite the function that computes the composite likelihood in a way that was more optimized for speed and memory management. This required, among other things, translating the Matlab code into the C language, which in this case can generate substantial speed advantage, if used judiciously. The end result is that we can go to $N = 1000$ assets without running out of memory, and it takes less than three minutes to estimate the DCC model with nonlinear shrinkage, a speed gain by a factor of at least 20.[5]

5.3 Results

Let $\widehat{H}_t$, $\widetilde{H}_t$, and $\overline{H}_t$ denote the estimators of the true conditional covariance matrix H_t obtained by using as target respectively the sample correlation matrix, the nonlinear shrinkage estimator, and the linear shrinkage estimator, for $t = 1, \ldots, T$. Following Section 4.4, we will denote the average loss corresponding to these three estimators by $\widehat{L}$, $\widetilde{L}$, and $\overline{L}$, respectively. The results averaged across $10^5/N$ Monte Carlo simulations are presented in Table 1.[6]

[2]The Matlab codes can be downloaded at http://www.econ.uzh.ch/en/people/faculty/wolf/publications.html under the link "Programming Code". Pratik Ramprasad from the Master's in Financial Statistics and Risk Management (FSRM) at Rutgers University implemented nonlinear shrinkage in the (free) open-source language R, and posted the code on the Comprehensive R Archive Network (CRAN) repository at http://cran.rstudio.com/web/packages/nlshrink/.

[3]This toolbox is no longer supported, but it can still be downloaded from its author's personal website at http://www.kevinsheppard.com/UCSD_GARCH.

[4]It can be downloaded from the following code repository: http://github.com/bashtage/mfe-toolbox.

[5]Using Matlab R2014b on an Apple Mac Pro with a 3.5 GHz Intel Xeon E5 processor and 60GB of memory.

[6]In large dimensions, such as $N = 1000$, the results are extremely consistent from one simulation to the next, so there is no need to go to thousands of simulations.

N	Sample	Linear	Nonlinear
100	0.612	0.596	0.547
500	0.120	0.098	0.066
1000	0.421	0.147	0.079

N	Sample	Linear	Nonlinear
100	0.637	0.621	0.572
500	0.123	0.101	0.068
1000	0.427	0.152	0.083

Table 1: Average loss for DCC estimators using three different methods for correlation targeting in dimensions $N \in \{100, 500, 1000\}$ with sample size $T = 1250$. The unit is 10^{-3}. The left panel is for normally distributed variates and the right for the "Student" t-distribution with 5 degrees of freedom rescaled to have unit variance.

An intuitive way to quantify the improvement is to compute the Percentage Relative Improvement in Average Loss (PRIAL). For example, the PRIAL of nonlinear shrinkage relative to the sample correlation matrix is defined as:

$$100 \times \left\{ 1 - \frac{\mathsf{E}\left[\widetilde{L}\right]}{\mathsf{E}\left[\widehat{L}\right]} \right\} \% \ , \tag{5.1}$$

where the expectation is taken across Monte Carlo simulations. By construction, the PRIAL of the true covariance matrix is 100%, which is the maximum attainable; and 0% means no improvement at all.

The economic interpretation is that, as long as we do not know the true conditional correlation matrix, estimation error will cause excess out-of-sample portfolio variance, and we want to eliminate as much of it as possible: the PRIAL says what percentage of it we managed to eliminate by nonlinear shrinkage. Table 2 gives the PRIAL of nonlinear shrinkage with respect to the sample covariance matrix and linear shrinkage respectively.

N	Sample	Linear
100	10.6%	8.3%
500	45.4%	33.2%
1000	81.2%	46.0%

N	Sample	Linear
100	10.1%	7.9%
500	44.3%	32.4%
1000	80.6%	45.5%

Table 2: PRIAL of nonlinear shrinkage with respect to the sample covariance matrix and linear shrinkage. The setup is the same as in Table 1. The left panel is for normally distributed variates and the right for the "Student" t distribution with 5 degrees of freedom rescaled to have unit variance.

One can see that shrinking the eigenvalues of the correlation target matrix nonlinearly generates substantial improvements in the estimation of the DCC model. The improvement is stronger the higher the number of assets. This effect is robust against fat-tailedness.

5.4 Dynamic Correlation Parameters

There are two possible channels through which the improvement could be happening: either directly because nonlinear shrinkage just gives a better estimate of the correlation targeting matrix or indirectly because it yields better estimates of the dynamic correlation parameters α and β in equation (2.2). In order to distinguish between these two hypotheses, we report the mean and standard deviations (across Monte Carlo simulations) of the estimates of the parameters α and β produced by the various methods. Remember that the true parameters are $\alpha = 0.05$ and $\beta = 0.93$ as per Section 5.1. Table 3 has statistics for the α estimates and Table 4 for the β estimates.

N	Sample	Linear	Nonlinear
100	0.0485 (0.0026)	0.0485 (0.0026)	0.0485 (0.0026)
500	0.0489 (0.0019)	0.0490 (0.0019)	0.0490 (0.0019)
1000	0.0490 (0.0023)	0.0490 (0.0023)	0.0490 (0.0023)

N	Sample	Linear	Nonlinear
100	0.0483 (0.0026)	0.0484 (0.0026)	0.0484 (0.0026)
500	0.0486 (0.0020)	0.0487 (0.0020)	0.0487 (0.0020)
1000	0.0488 (0.0021)	0.0489 (0.0021)	0.0489 (0.0021)

Table 3: Estimated parameter α from correlation matrix dynamics. The panels show the mean and (in parentheses) the standard deviation across Monte Carlo simulations. The setup is the same as in Table 1. The left panel is for normally distributed variates and right for the "Student" t-distribution with 5 degrees of freedom rescaled to have unit variance.

N	Sample	Linear	Nonlinear
100	0.9301 (0.0035)	0.9301 (0.0035)	0.9301 (0.0035)
500	0.9296 (0.0026)	0.9296 (0.0026)	0.9297 (0.0026)
1000	0.9292 (0.0030)	0.9293 (0.0030)	0.9293 (0.0030)

N	Sample	Linear	Nonlinear
100	0.9299 (0.0036)	0.9299 (0.0036)	0.9300 (0.0036)
500	0.9297 (0.0026)	0.9297 (0.0026)	0.9297 (0.0026)
1000	0.9293 (0.0029)	0.9294 (0.0029)	0.9294 (0.0029)

Table 4: Estimated parameter β from correlation matrix dynamics. The panels show the mean and (in parentheses) the standard deviation across Monte Carlo simulations. The setup is the same as in Table 1. The left panel is for normally distributed variates and right for the "Student" t-distribution with 5 degrees of freedom rescaled to unit variance.

One can see that there is no pattern. Therefore the explanation for the outperformance of linear shrinkage lies not in the better estimation of the dynamic correlation parameters but in the better estimation of the correlation targeting matrix (that is, the DCC model intercept).

Remark 5.1. An alternative way of summarizing the same set of results would be to say that the composite likelihood method manages to obtain accurate estimates of the α and β parameters even when the targeting correlation matrix is the sample correlation matrix, with all the associated difficulties in large dimensions. This is an advantage of composite likelihood, as we would expect the maximum likelihood estimators of α and β to be hampered in the presence of a rank-deficient or at least ill-conditioned targeting correlation matrix. ■

6 Empirical Results

The goal of this section is to examine the out-of-sample properties of Markowitz portfolios based on our newly suggested covariance matrix estimator. There are a myriad of popular investment strategies by now and it is not our goal to compare to an extensive list of them. The only focus of this section is compare nonlinear DCC to linear DCC and sample DCC.

For compactness of notation, we do not use the subscript T in denoting the covariance matrix itself, an estimator of the covariance matrix, or a return-predictive signal that proxies for the vector of expected returns.

6.1 Data and General Portfolio-Formation Rules

We download daily data from the Center for Research in Security Prices (CRSP) starting in 01/08/1980 and ending in 12/31/2015. We restrict attention to stocks from the NYSE and NASDAQ stock exchanges. For simplicity, we adopt the common convention that 21 consecutive trading days constitute one 'month'. The out-of-sample period ranges from 01/01/1986 through 12/31/2015, resulting in a total of 360 'months' (or 7560 days). All portfolios are updated 'monthly'.[7] We denote the investment dates by $h = 1, \ldots, 360$. At any investment date h, a covariance matrix is estimated using the most recent $T = 1250$ daily returns, which roughly corresponds to using to five years of past data.

We consider the following portfolio sizes: $N \in \{100, 500, 1000\}$. For a given combination (h, n), the investment universe is obtained as follows. We find the set of stocks that have have a complete return history over the most recent $T = 1250$ days as well as a complete return 'future' over the next 21 days.[8] We then look for possible pairs of highly correlated stocks, that is, pairs of stocks that returns with a sample correlation exceeding 0.95 over the past 1250 days. With such pairs, if they should exist, we remove the stock with the lower volume of the two on investment date h.[9] Of the remaining set of stocks, we then pick the largest N stocks

[7] 'Monthly' updating is common practice to avoid an unreasonable amount of turnover and thus transaction costs. During a 'month', from one day to the next, we hold number of shares fixed rather than portfolio weights; in this way, there are no transactions at all during a 'month'.

[8] The latter, 'forward-looking' restriction is not a feasible one in real life but is commonly applied in the related finance literature on the out-of-sample evaluation of portfolios.

[9] The reason is that we do not include highly similar stocks, or even the same stock, listed under two different

(as measured by their market volume on investment date h) as our investment universe. In this way, the investment universe changes slowly from one investment date to the next.

6.2 Global Minimum Variance Portfolio

We consider the problem of estimating the global minimum variance (GMV) portfolio, in the absence of short-sales constraints. The problem is formulated as

$$\min_{w} w'\Sigma w \tag{6.1}$$

$$\text{subject to} \quad w'\mathbf{1} = 1 \ , \tag{6.2}$$

where $\mathbf{1}$ denotes a vector of ones of dimension $N \times 1$. It has the analytical solution

$$w = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}} \ . \tag{6.3}$$

The natural strategy in practice is to replace the unknown Σ by an estimator $\widetilde{\Sigma}$ in formula (6.3), yielding a feasible portfolio

$$\widetilde{w} := \frac{\widetilde{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\widetilde{\Sigma}^{-1}\mathbf{1}} \ . \tag{6.4}$$

Estimating the GMV portfolio is a 'clean' problem in terms of evaluating the quality of a covariance matrix estimator, since it abstracts from having to estimate the vector of expected returns at the same time. In addition, researchers have established that estimated GMV portfolios have desirable out-of-sample properties not only in terms of risk but also in terms of reward-to-risk (that is, in terms of the Sharpe ratio); for example, see Haugen and Baker (1991), Jagannathan and Ma (2003), and Nielsen and Aylursubramanian (2008). As a result, such portfolios have become an addition to the large array of products sold by the mutual-fund industry.

The following four portfolios are included in the study.

- **$1/N$**: the equal-weighted portfolio. This portfolio is a standard benchmark and has been promoted by DeMiguel et al. (2009), among others.
- **DCC-S:** the portfolio (6.4), where the estimator $\widetilde{\Sigma}$ is obtained from DCC based on the sample correlation matrix.
- **DCC-L:** the portfolio (6.4), where the estimator $\widetilde{\Sigma}$ is obtained from DCC based on linear shrinkage.
- **DCC-NL:** the portfolio (6.4), where the estimator $\widetilde{\Sigma}$ is obtained from DCC based on nonlinear shrinkage.

We report the following three out-of-sample performance measures for each scenario. (All of them are annualized and in percent for ease of interpretation.)

permanent issue identification numbers (PERMNOs) in the CRSP data base. In the early years, there are no such pairs; in the most recent years, there are never more than three such pairs.

- **AV:** We compute the average of the 7560 out-of-sample log returns and then multiply by 252 to annualize.
- **SD:** We compute the standard deviation of the 7560 out-of-sample log returns and then multiply by $\sqrt{252}$ to annualize.
- **IR:** We compute the (annualized) information ratio as the ratio AV/SD.[10]

Our stance is that in the context of the GMV portfolio, the most important performance measure is the out-of-sample standard deviation, SD. The true (but unfeasible) GMV portfolio is given by (6.3). It is designed to minimize the variance (and thus the standard deviation) rather than to maximize the expected return or the information ratio. Therefore, any portfolio that implements the GMV portfolio should be primarily evaluated by how successfully it achieves this goal. A high out-of-sample average return, AV, and a high out-of-sample information ratio, IR, are naturally also desirable, but should be considered of secondary importance from the point of view of evaluating the quality of a covariance matrix estimator.

We also consider the question of whether DCC-NL delivers a lower out-of-sample standard deviation than DCC-S at a level that is statistically significant. For a given universe size N, a two-sided p-value for the null hypothesis of equal standard deviations is obtained by the prewhitened HAC_{PW} method described in Ledoit and Wolf (2011, Section 3.1).[11]

The results are presented in Table 5 and can be summarized as follows; unless stated otherwise, the findings are with respect to the out-of-sample standard deviation as performance measure.

- All DCC variants consistently outperform $1/N$ by a wide margin.
- DCC-L consistently outperforms DCC-S but, in turn, is consistently outperformed by DCC-NL.
- The outperformance of DCC-NL over DCC-S is always statistically significant and it is also economically meaningful for $N = 500, 1000$.
- DeMiguel et al. (2009) claim that it is difficult to outperform $1/N$ in terms of the out-of-sample Sharpe ratio with 'sophisticated' portfolios (that is, with Markowitz portfolios that estimate input parameters). It can be seen that all DCC variants consistently outperform $1/N$ in terms of the out-of-sample Information ratio, which translates into outperformance in terms of the out-of-sample Sharpe ratio. Again, DCC-NL is consistently best among the DCC variants.

[10] This version of the information ratio, which simply uses zero as the benchmark, is widely used in the mutual fund industry.

[11] Since the out-of-sample size is very large at 7560, there is no need to use the computationally more involved bootstrap method described in Ledoit and Wolf (2011, Section 3.2), which is preferred for small sample sizes.

Period: 01/08/1986–12/31/2015				
	$1/N$	DCC-S	DCC-L	DCC-NL
$N = 100$				
AV	12.10	9.92	9.91	9.95
SD	21.56	13.36	13.33	**13.17*****
IR	0.56	0.74	0.74	0.76
$N = 500$				
AV	13.46	13.94	13.88	13.38
SD	19.53	10.57	10.40	**9.64*****
IR	0.69	1.32	1.33	1.39
$N = 1000$				
AV	14.21	11.77	12.15	12.17
SD	19.04	10.59	9.14	**8.02*****
IR	0.75	1.11	1.33	1.52

Table 5: Annualized performance measures (in percent) for various estimators of the GMV portfolio. AV stands for average; SD stands for standard deviation; and IR stands for information ratio. All measures are based on 7560 daily out-of-sample returns from 01/08/1986 through 12/31/2015. In the rows labeled SD, the lowest number appears in **bold face**. In the columns labeled DCC-S and DCC-NL, significant outperformance of one of the two portfolios over the other in terms of SD is denoted by asterisks: *** denotes significance at the 0.01 level; ** denotes significance at the 0.05 level; and * denotes significance at the 0.1 level.

6.3 Markowitz Portfolio with Momentum Signal

We now turn attention to a 'full' Markowitz portfolio with a signal.

By now a large number of variables have been documented that can be used to construct a signal in practice. For simplicity and reproducibility, we use the well-known momentum factor (or simply momentum for short) of Jegadeesh and Titman (1993). For a given investment period h and a given stock, the momentum is the the geometric average of the previous 252 returns on the stock but excluding the most recent 21 returns; in other words, one uses the geometric average over the previous 'year' but excluding the previous 'month'. Collecting the individual momentums of all the N stocks contained in the portfolio universe yields the return-predictive signal m.

In the absence of short-sales constraints, the investment problem is formulated as

$$\min_{w} w'\Sigma w \tag{6.5}$$

$$\text{subject to} \quad w'm = b \ , \text{ and} \tag{6.6}$$

$$w'\mathbf{1} = 1 \ , \tag{6.7}$$

where b is a selected target expected return. The problem has the analytical solution

$$w = c_1\Sigma^{-1}\mathbf{1} + c_2\Sigma^{-1}m \ , \tag{6.8}$$

$$\text{where } c_1 := \frac{C - bB}{AC - B^2} \quad \text{and} \quad c_2 := \frac{bA - B}{AC - B^2} \ , \tag{6.9}$$

$$\text{with } A := \mathbf{1}'\Sigma^{-1}\mathbf{1} \ , \quad B := \mathbf{1}'\Sigma^{-1}b \ , \quad \text{and} \quad C := m'\Sigma^{-1}m \ . \tag{6.10}$$

The natural strategy in practice is to replace the unknown Σ by an estimator $\widetilde{\Sigma}$ in formulas (6.8)–(6.10), yielding a feasible portfolio

$$\widetilde{w} := c_1\widetilde{\Sigma}^{-1}\mathbf{1} + c_2\widetilde{\Sigma}^{-1}m \ , \tag{6.11}$$

$$\text{where } c_1 := \frac{C - bB}{AC - B^2} \quad \text{and} \quad c_2 := \frac{bA - B}{AC - B^2} \ , \tag{6.12}$$

$$\text{with } A := \mathbf{1}'\widetilde{\Sigma}^{-1}\mathbf{1} \ , \quad B := \mathbf{1}'\widetilde{\Sigma}^{-1}b \ , \quad \text{and} \quad C := m'\widetilde{\Sigma}^{-1}m \ . \tag{6.13}$$

The following four portfolios are included in the study.

- **EW-TQ** The equal-weighted portfolio of the top-quintile stocks according to momentum m. This strategy does not make use of the momentum signal beyond sorting of the stocks in quintiles.

  The value of the target expected return b for the remaining four portfolios below is then given by the arithmetic average of the momentums of the stocks included in this portfolio (that is, the expected return of EW-TQ according to the signal m).
- **DCC-S:** the portfolio (6.8)–(6.10), where the estimator $\widetilde{\Sigma}$ is obtained from DCC based on the sample correlation matrix.
- **DCC-L:** the portfolio (6.8)–(6.10), where the estimator $\widetilde{\Sigma}$ is obtained from DCC based on linear shrinkage.
- **DCC-NL:** the portfolio (6.8)–(6.10), where the estimator $\widetilde{\Sigma}$ is obtained from DCC based on nonlinear shrinkage.

Our stance is that in the context of a 'full' Markowitz portfolio, the most important performance measure is the out-of-sample information ratio, IR. In the 'ideal' investment problem (6.8)–(6.10), minimizing the variance (for a fixed target expected return b) is equivalent to maximizing the information ratio (for a fixed target expected return b). In practice, because of estimation error in the signal, the various strategies do not have the same expected return and, thus, focusing on the out-of-sample standard deviation is inappropriate.

We also consider the question whether DCC-NL delivers a higher out-of-sample information ratio than DCC-S at a level that is statistically significant. For a given universe size N, a two-sided p-value for the null hypothesis of equal information ratios is obtained by the prewhitened HAC_{PW} method described in Ledoit and Wolf (2008, Section 3.1).[12]

The results are presented in Table 6 and can be summarized as follows; unless stated otherwise, the findings are with respect to the out-of-sample Information ratio as performance measure.

- All DCC variants consistently outperform EW-TQ by a wide margin.
- DCC-L consistently outperforms DCC-S but, in turn, is consistently outperformed by DCC-NL.
- The outperformance of DCC-NL over DCC-S is statistically significant and also economically meaningful for $N = 500, 1000$.
  DeMiguel et al. (2009) claim that it is difficult to outperform $1/N$ in terms of the out-of-sample Sharpe ratio with 'sophisticated' portfolios (that is, with Markowitz portfolios that estimate input parameters). Comparing table 6 with Table 5, it can be seen that all DCC variants consistently outperform $1/N$ in terms of the out-of-sample Information ratio, which translates into outperformance in terms of the out-of-sample Sharpe ratio. Even though momentum is not a very powerful return-predictive signal, the differences can be enormous. For example, for $N = 1000$, the information ratio of $1/N$ is only 0.75 while the information ratio of DCC-NL is 1.62 and thus more than twice as large.
- Engle and Colacito (2006) argue for the use of the out-of-sample standard deviation, SD, as a performance measure also in the context of a 'full' Markowitz portfolio. Also for this alternative performance measure, all DCC variants consistently outperform EW-TQ by a wide margin. Furthermore, DCC-NL always has the smallest out-of-sample standard deviation and its outperformance over DCC-S is always statistically significant.[13]

[12] Since the out-of-sample size is very large at 7560, there is no need to use the computationally more expensive bootstrap method described in Ledoit and Wolf (2008, Section 3.2), which is preferred for small sample sizes.

[13] Statistical significance is again judged by the prewhitened HAC_{PW} method of Ledoit and Wolf.

Period: 01/08/1986–12/31/2015				
	EW-TQ	DCC-S	DCC-L	DCC-NL
$N = 100$				
AV	17.13	15.79	15.79	15.77
SD	28.43	17.05	17.03	16.90***
IR	0.60	0.93	0.93	**0.93**
$N = 500$				
AV	17.15	16.60	16.66	16.78***
SD	24.42	12.36	12.16	11.31
IR	0.70	1.34	1.37	**1.48****
$N = 1000$				
AV	17.35	12.78	13.96	14.92
SD	22.89	13.07	10.76	9.20***
IR	0.76	0.98	1.30	**1.62*****

Table 6: Annualized performance measures (in percent) for various estimators of the Markowitz portfolio with momentum signal. AV stands for average; SD stands for standard deviation; and IR stands for information ratio. All measures are based on 7560 daily out-of-sample returns from 01/08/1986 until 12/31/2015. In the rows labeled IR, the largest number appears in **bold face**. In the columns labeled DCC-S and DCC-NL, significant outperformance of one of the two portfolios over the other in terms of IR is denoted by asterisks: *** denotes significance at the 0.01 level; ** denotes significance at the 0.05 level; and * denotes significance at the 0.1 level.

7 Conclusion

This paper demonstrates that there is a 'division of labor' between composite likelihood and nonlinear shrinkage in the estimation of a Dynamic Conditional Correlation (DCC) model: The former takes care of the dynamic correlation parameters (time series) while the latter takes care of the correlation targeting matrix (cross-section). Their actions complement each other. Together they enable DCC to conquer high dimensions on the order of a thousand, such as are frequently encountered in modern portfolio theory and risk management. The attractive performance of our new proposal has been established both by simulations studies and by backtesting on real-life stock return data.

References

Bai, Z. D. and Silverstein, J. W. (2010). *Spectral Analysis of Large-Dimensional Random Matrices*. Springer, New York, second edition.

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy? *Review of Financial Studies*, 22:1915–1953.

El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Annals of Statistics*, 36(6):2757–2790.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50:987–1007.

Engle, R. F. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350.

Engle, R. F. (2009). *Anticipating Correlations: A New Paradigm for Risk Management*. Princeton University Press.

Engle, R. F. and Colacito, R. (2006). Testing and valuing dynamic correlations for asset allocation. *Journal of Business and Economic Statistics*, 24(2):238–253.

Engle, R. F. and Kroner, K. (1995). Multivariate simultaneous GARCH. *Econometric Theory*, 11:122–150.

Engle, R. F. and Mezrich, J. (1996). GARCH for groups. *Risk*, 9:36–40.

Haffner, C. M. and Reznkikova, O. (2012). On the estimation of dynamic conditional correlation models. *Computational Statistics & Data Analysis*, 56(11):3533–3545.

Haugen, R. A. and Baker, N. L. (1991). The efficient market inefficiency of capitalization-weighted stock portfolios. *The Journal of Portfolio Management*, 17(3):35–40.

IPCC (2007). Climate change 2007: the scientific basis. In Solomon, S., Qin, D., Manning, M., Marquis, M., Averyt, K., Tignor, M. M., Miller, H. L., and Chen, Z., editors, *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. 996pp.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 54(4):1651–1684.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91.

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 150(1–2):233–264.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, 15:850–859.

Ledoit, O. and Wolf, M. (2011). Robust performance hypothesis testing with the variance. *Wilmott Magazine*, September:86–89.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*, 40(2):1024–1060.

Ledoit, O. and Wolf, M. (2014). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. Working Paper ECON 137, Department of Economics, University of Zurich.

Ledoit, O. and Wolf, M. (2015). Spectrum estimation: a unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis*, 139(2):360–384.

Ledoit, O. and Wolf, M. (2016). Numerical implementation of the QuEST function. Working Paper ECON 215, Department of Economics, University of Zurich.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4):457–483.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7:77–91.

Mestre, X. (2008). On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices. *IEEE Transactions on Signal Processing*, 56(11):5353–5368.

Michaud, R. (1989). The Markowitz optimization enigma: Is optimized optimal? *Financial Analysts Journal*, 45:31–42.

Nielsen, F. and Aylursubramanian, R. (2008). Far from the madding crowd — Volatility efficient indices. Research insights, MSCI Barra.

Pakel, C., Shephard, N., Sheppard, K., and Engle, R. F. (2014). Fitting vast dimensional time-varying covariance models. Technical report.

Perlman, M. D. (2007). *STAT 542: Multivariate Statistical Analysis*. University of Washington (On-Line Class Notes), Seattle, Washington.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *Journal of Multivariate Analysis*, 55:331–339.

Silverstein, J. W. and Bai, Z. D. (1995). On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *Journal of Multivariate Analysis*, 54:175–192.

Silverstein, J. W. and Choi, S. I. (1995). Analysis of the limiting spectral distribution of large-dimensional random matrices. *Journal of Multivariate Analysis*, 54:295–309.

Stein, C. (1975). Estimation of a covariance matrix. Rietz lecture, 39th Annual Meeting IMS. Atlanta, Georgia.

Stieltjes, T. J. (1894). Recherches sur les fractions continues. *Annales de la Faculté des Sciences de Toulouse 1re Série*, 8(4):J1–J122.

Wigner, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62(3):548–564.

Yin, Y. (1986). Limiting spectral distribution for a class of random matrices. *Journal of Multivariate Analysis*, 20(1):50–68.

A Primer on Nonlinear Shrinkage

Nonlinear shrinkage estimation of the unconditional covariance matrix is a burgeoning field of probability and statistics which may not be very accessible to applied researchers in economics and finance. This appendix provides a self-contained introduction. It is intended to be descriptive, qualitative, and as non-technical as the nature of the subject matter will allow. It is not intended as a substitute for the rigorous treatment provided in Ledoit and Wolf (2014).

The exposition here is couched in terms of the covariance matrix, but in the DCC context the described estimator should be applied to the *devolatilized* residuals. The resulting estimate should then be renormalized as per Section 3.5 in order to generate a proper correlation matrix.

Σ denotes the population covariance matrix. Y_T denotes a stationary data set of dimension $T \times N$ with covariance matrix Σ. We assume mean zero for simplicity.

A.1 Importance of the Eigenvalues for Portfolio Selection

Following Markowitz (1952), if μ denotes a vector of expected returns, then the weights of the tangency portfolio are

$$w^{\text{TANGENCY}} = \text{scalar} \times \Sigma^{-1}\mu \ . \tag{A.1}$$

Inverting a matrix is not a particularly intuitive operation, and when an experienced practitioner like Michaud (1989) warns that it leads to "error maximization", it is hard to see what is going wrong or how to fix it.

Fortunately, the covariance matrix is not just *any* matrix, it is a *symmetric* matrix. The covariance of the return on Intel shares with Nike shares is the same as Nike with Intel by definition. Symmetric matrices enjoy a very special property: they always admit a *spectral decomposition*. It is given by

$$\Sigma =: V \begin{bmatrix} \tau_1 & & & \\ & \tau_2 & & 0 \\ & & \ddots & \\ & 0 & & \ddots \\ & & & & \tau_N \end{bmatrix} V' \ , \tag{A.2}$$

where $\boldsymbol{\tau} := (\tau_1, \ldots, \tau_N)$ are the population eigenvalues and V is a rotation matrix, meaning that $V' = V^{-1}$. The ith column of V is the population eigenvector v_i.

The best way to interpret this decomposition is to look at the dimension $N = 2$. Figure 1 gives a graphical illustration.

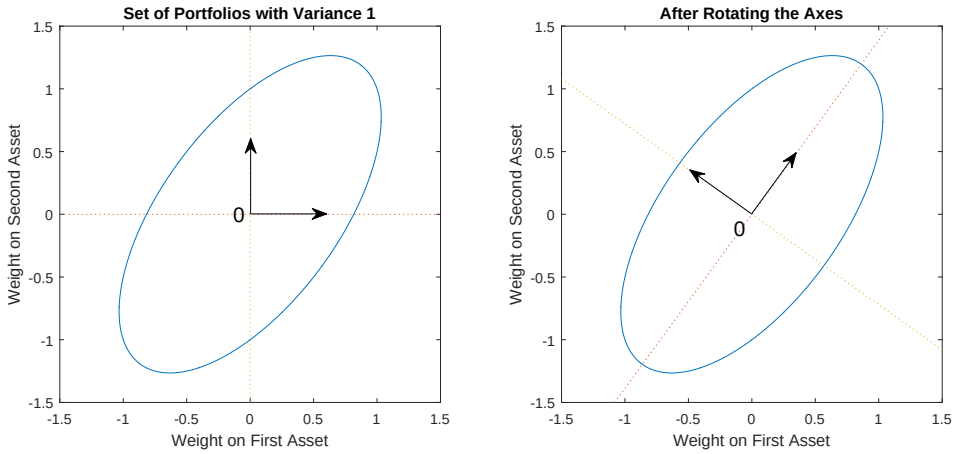

Figure 1: Decomposition into eigenvalues and eigenvectors in dimension $N = 2$.

On the left panel, the ellipsis represents all the portfolios that have the same variance (which we can take as normalized to one). The fact that it bulges out into the northeast and southwest quadrants is due to nonzero covariance between the two asset returns. What the eigenvectors do is define a change of basis, a rotation of the axes, so that (when viewed from this new angle) all quadrants look the same. This is shown on the right panel. One eigenvector corresponds to the axis across which the ellipsis is narrowest, and another corresponds to the axis across which the ellipsis is widest. The ellipsis is not a perfect circle because the two eigenvalues are not equal to each other.

Economically, what this rotation does is to repackage the original menu of assets into N funds whose returns are all mutually uncorrelated. The weights of each fund are given by the corresponding eigenvector; and the corresponding eigenvalue is the variance of the return on the fund. Since they represent fund return variances, all eigenvalues must be non-negative, meaning that the covariance matrix is *positive semi-definite*.

The N funds span the same space of investment opportunities as the original assets, therefore we can rewrite equation (A.1) as

$$w^{\text{TANGENCY}} = \text{scalar} \times \sum_{i=1}^{N} \frac{v_i' \mu}{\tau_i} v_i \ . \tag{A.3}$$

Equation (A.3) demonstrates that the tangency portfolio is best viewed not as a combination of the N original assets, but as a combination of the N uncorrelated eigenvector funds. The capital assigned to each fund is proportional to its expected return and inversely proportional to its variance, which makes economic sense.

It is easy to justify that the spectral decomposition is important for portfolio selection. Consider the hypothetical covariance matrix of monthly stock returns in Table 7.

	Apple	Boeing	Disney	IBM
Apple	0.2694	0.5714	0.2900	0.3080
Boeing	0.5714	1.3910	0.6674	0.6964
Disney	0.2900	0.6674	0.3275	0.3433
IBM	0.3080	0.6964	0.3433	0.4822

Table 7: Hypothetical covariance matrix between four US stocks.

To the naked eye, it looks fine. However, its eigenvalues are $(0, 0.0299, 0.1072, 2.3329)$. Even to the naked eye, the first eigenvalue looks wrong. The tangency portfolio does not exist when an eigenvalue is equal to zero. This is why extracting eigenvalues and eigenvectors is called the *spectral* decomposition: It enables us to penetrate right through the outer appearance of the matrix into its inner structure.

In practice, we do not know the true covariance matrix Σ, therefore we must use some estimator of it. It is known that the sample covariance matrix $S_T := Y_T' Y_T / T$ is a consistent estimator of Σ when the sample size T goes to infinity *while the dimension* N *remains fixed* (an often overlooked yet crucial assumption, to which we will return later). Mirroring equation (A.2), define the spectral composition of S_T as

$$S_T =: U_T \begin{bmatrix} \lambda_{1,T} & & & \\ & \lambda_{2,T} & 0 & \\ & & \ddots & \\ & 0 & & \ddots \\ & & & \lambda_{N,T} \end{bmatrix} U_T' \ , \tag{A.4}$$

where $\boldsymbol{\lambda}_T := (\lambda_{1,T}, \ldots, \lambda_{N,T})$ are the sample eigenvalues and U_T is a rotation matrix ($U_T' = U_T^{-1}$) whose ith column is the sample eigenvector $u_{i,T}$. An implementable version of equation (A.3) is

$$w^{\text{TANGENCY}} = \text{scalar} \times \sum_{i=1}^{N} \frac{u_{i,T}' \mu}{\lambda_{i,T}} u_{i,T} \ . \tag{A.5}$$

This formulation leads to a fundamental insight: In the denominator, we have $\lambda_{i,T}$, which is the *in-sample* variance of the ith eigenvector fund $u_{i,T}$, whereas for investment purposes we need its *out-of-sample* variance $u_{i,T}' \Sigma u_{i,T}$ instead. The whole point of the procedure advocated in this paper is to replace the former with (a consistent estimate of) the latter. Investment decisions are always evaluated out of sample.

One little-known mathematical fact about the eigenvalues is that they are the *most dispersed* diagonal elements that can be obtained through rotation.[14] Given that the group

[14] Ledoit and Wolf (2004, Section 2.3)

of rotations has dimensionality of order N^2, the potential for overfitting is tremendous when N is large. Overfitting causes excess dispersion: The smallest sample eigenvalues are too small, leading to over-investment, and the largest sample eigenvalues too large, leading to under-investment. The overall result is mal-investment. This insight goes a long way towards explaining the observation by Michaud (1989) about "error maximization". However, to fix it requires a detour through multivariate statistics.

A.2 Importance of the Eigenvalues for Covariance Matrix Estimation

If we are going to look for estimators that improve upon the sample covariance matrix, the first task is to decide *where* to look. We need to specify a class of eligible estimators, and search within this class for one that beats the sample covariance matrix. In mathematics, a standard way to approach this kind of problem is to say that we want estimators that have certain appealing properties. One such property initially championed by Stein (1975) and subsequently adopted by many other authors is called *rotation equivariance*. A covariance matrix estimator $\widehat{\Sigma}_T(Y_T)$ is said to be rotation-equivariant if and only if for any N-dimensional rotation matrix W,

$$\widehat{\Sigma}_T(Y_T W) = W' \, \widehat{\Sigma}_T(Y_T) \, W \ . \tag{A.6}$$

That is, the estimate based on the rotated data equals the rotation of the estimate based on the original data. Absent any *a priori* knowledge about the orientation of the true eigenvectors, it is natural to consider only covariance matrix estimators that are rotation-equivariant.

It can be proven that the class of rotation-equivariant estimators that are a function of the sample covariance matrix is the class of estimators of the form

$$\widehat{\Sigma}_{\Psi_T} \coloneqq U_T \begin{bmatrix} \psi_{1,T} & & & \\ & \psi_{2,T} & & 0 \\ & & \ddots & \\ & 0 & & \ddots \\ & & & & \psi_{N,T} \end{bmatrix} U_T' \ , \tag{A.7}$$

where $\Psi_T \coloneqq (\psi_{1,T}, \ldots, \psi_{N,T})$ can be any vector in $[0, +\infty)^N$.[15] Thus, we preserve the sample eigenvectors, but are free to modify the sample eigenvalues in any way needed to improve upon the sample covariance matrix. Given Section A.1, the basic idea will be to set $\psi_{i,T}$ equal to $u_{i,T}' \Sigma u_{i,T}$, or if it is unavailable (the more likely scenario, given that it depends on the population covariance matrix, which is unobservable), a consistent estimator thereof.

In summary, the key intuition is that we have to preserve the sample eigenvectors because we lack *a priori* information about the orientation of the true eigenvectors, and the goal is to modify the sample eigenvalues so we can beat the sample covariance matrix.

[15] See for example Perlman (2007, Section 5.4).

A.3 General Asymptotics

As mentioned before, the sample covariance matrix S_T is a consistent estimator of the population covariance matrix Σ when the sample size T goes to infinity while the dimension N remains fixed. This is strange: why is T allowed to move but not N? When we have five years of daily data ($T = 1250$) on the components of the Russell 1000 stock index ($N = 1000$), it is easy to believe that T goes to infinity, as 1250 is a large number by any measure in statistics, but who is to say that N is finite? Shouldn't numbers that go to infinity be much bigger than those that are assumed to remain finite?

The answer is to simply relax the constraining assumption that N is fixed and instead allow the dimension to move along with the sample size: $N := N(T)$. This is called *general asymptotics*, large-dimensional asymptotics, or Kolmogorov asymptotics. Notation-wise, this kind of asymptotics requires appending the subscript T to the population covariance matrix, and also its eigenvalues and eigenvectors, a convention that we will uphold from here onwards.

Given that the number of eigenvalues N goes to infinity, it is no longer possible to make statements about individual eigenvalues. This is why it is necessary to introduce what is known as *spectral distributions*. The population and sample spectral distributions are defined respectively as

$$\forall x \in \mathbb{R} \qquad H_T(x) := \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{x \leq \tau_{i,T}\}} \tag{A.8}$$

$$\forall x \in \mathbb{R} \qquad F_T(x) := \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{x \leq \lambda_{i,T}\}} \ , \tag{A.9}$$

where $\mathbf{1}$ denotes the indicator function. The spectral distribution can be interpreted as a cross-sectional cumulative distribution function (c.d.f.): it is a nondecreasing function having values between zero and one that returns the proportion of eigenvalues lower than its argument.

Two standard assumptions under general asymptotics are (i) that the population spectral distribution converges to a well-defined limit H called the *limiting spectral distribution* and (ii) that the ratio N/T converges to a finite limit c called the *concentration*:

$$H_T(x) \longrightarrow H(x) \qquad \text{at all points of continuity of } H \tag{A.10}$$

$$\frac{N}{T} \longrightarrow c < +\infty \ . \tag{A.11}$$

Along with other technical assumptions that can vary from author to author, these two assumptions imply the fundamental result of general asymptotics, which is that the sample spectral distribution converges to a *nonrandom* limit F called the limiting spectral distribution:

$$F_T(x) \xrightarrow{\text{a.s.}} F(x) \qquad \text{at all points of continuity of } F. \tag{A.12}$$

Remark A.1. The fact that the matrix is random but its eigenvalues are not is a remarkable mathematical phenomenon first discovered by Wigner (1955) while investigating

the properties of the wave functions of complicated quantum mechanical systems; see Figure 2 for an illustration of this influential result. ■

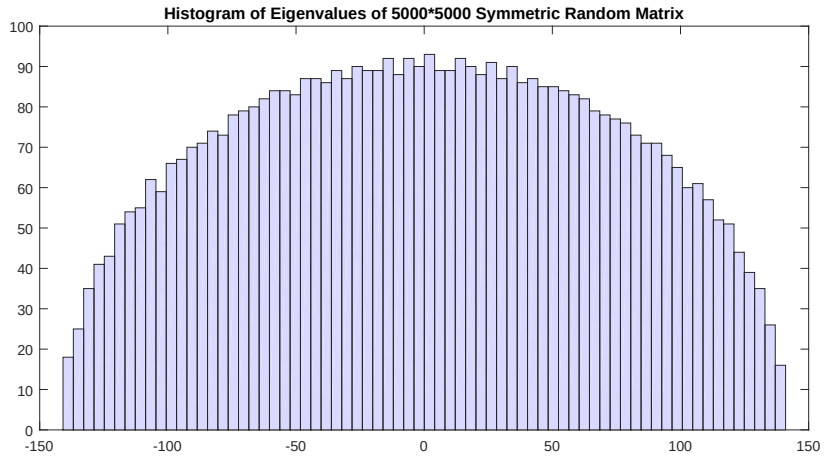

Figure 2: The eigenvalues of a large Wigner matrix follow Wigner's semi-circular law. Wigner matrices are random symmetric matrix with i.i.d. standard normal entries. (They are different from covariance matrices.) This picture does not represent an average across Monte Carlo simulations: it is just the result of one single draw.

The limiting sample spectral distribution F is the key to knowing where the sample eigenvalues lie. There are a few things we can immediately say about this important object:

(a) F is uniquely determined by H and c.[16]

(b) $F = H \iff c = 0$

(c) $\int_{-\infty}^{+\infty} x\, dF(x) = \int_{-\infty}^{+\infty} x\, dH(x)$

(d) $\int_{-\infty}^{+\infty} x^2\, dF(x) = \int_{-\infty}^{+\infty} x^2\, dH(x) + c\left[\int_{-\infty}^{+\infty} x\, dH(x)\right]^2$

Statement (b) confirms that finite-dimensional asymptotics are included as a special case of general asymptotics. When N remains fixed and finite, N/T converges to zero as T goes to infinity. In this case, the eigenvalues of the sample covariance matrix are consistent estimators of their sample counterparts. This remains true even if N goes to infinity along with T, as long as it grows sufficiently slowly (say in $\log(T)$ or $\sqrt{T}$). When $c = 0$ or, practically speaking, when N/T is minuscule, the sample covariance matrix works fine.

For five years of history on the Russell 1000, the ratio N/T is equal to 0.8, so it is definitely not minuscule. $c > 0$ is the relevant case for all large covariance matrices, because when N is large it is very difficult to have a sample size such that the ratio is N/T minuscule. In this case, the sample eigenvalues never get close to their population counterparts, so we

[16] Silverstein and Choi (1995)

enter a qualitatively different regime where improvement over the sample covariance matrix is possible.

Statement (c) means that the cross-sectional average of the sample eigenvalues is in the right place even when $c > 0$: it never needs fixing. However, Statement (d) shows their cross-sectional dispersion around the average is systematically inflated, and excess dispersion increases in c.[17] This confirms formally the intuition developed at the end of Section A.1. To fix this problem, the filtering applied to sample eigenvalues will have to 'shrink' their cross-sectional dispersion towards the center.[18]

A.4 Single Mass Point

One way to get information about the limiting sample spectral distribution is to study what happens in the simplest case, when the population covariance matrix Σ_T is equal to the identity matrix. In this case, a closed-form solution exists: F is differentiable, and its derivative f, called the limiting spectral density, follows the so-called Marčenko-Pastur Law:

$$\forall x \in [a, b] \qquad f(x) := \frac{\sqrt{(b-x)(x-a)}}{2\pi c x} \ ,$$

where the bounds of the support of f are $a := (1 - \sqrt{c})^2$ and $b := (1 + \sqrt{c})^2$ respectively.[19]

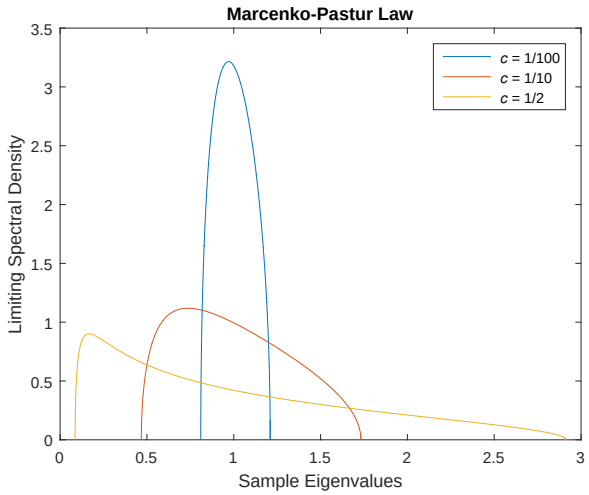

Figure 3: Limiting cross-sectional density of sample eigenvalues for various concentration levels.

Some lessons can be drawn, that reinforce and complement what we have learned from the first two moments:

1. Sample eigenvalues are smudged to the left and the right of the population eigenvalues.
2. The amount of excess spread increases in the concentration ratio c.

[17] Yin (1986, Equation (4.14))
[18] Ledoit and Wolf (2004)
[19] Marčenko and Pastur (1967)

3. The smallest sample eigenvalues are too small, the largest ones too large.
4. This systematic bias must be filtered out by *shrinking* the distribution of sample eigenvalues towards the center.
5. The center (cross-sectional average) of the sample eigenvalues distribution is accurate: it matches its population counterpart.
6. The density is right-skewed: there are many small sample eigenvalues and few large ones.
7. It is only in the limit $c \to 0$ that sample eigenvalues start to conform with standard (fixed-dimension) asymptotics.

The concentration c needs to be *very* close to 0 before concerns about excess dispersion can be safely dismissed. For $c = 1/10$, meaning that we have 10 times more observations than variables, which many people would deem sufficient, some sample eigenvalues are still less than half their population counterparts. Even when $c = 1/100$, out-of-sample portfolio variances can be under/overestimated by 20%, inducing over/under-allocation of risk capital by the same percentage.

A.5 Two Mass Points

These seven observations carry over to the case where a fraction α of the population eigenvalues are equal to one, while the rest are equal to some $\tau > 1$. The behavior of the limiting spectral density is pretty much as one would expect intuitively: There are two clusters formed around each population eigenvalue. The clusters are either close to each other or distant, depending on how far τ is from 1. This finding adds one more qualitative observation to the list:

8. The locations of clusters of sample eigenvalues match the locations of the underlying population eigenvalues.

When one mass point in the distribution of population eigenvalues is heavier than the other, the weights of the clusters of population eigenvalues adjust in proportion, as one would intuitively expect. This finding yields yet one more common-sense observation:

9. The weights of clusters of sample eigenvalues match the multiplicities of the underlying population eigenvalues.

The only non-obvious mathematical phenomenon is a so-called "phase transition" that takes place when two clusters merge with each other to form a single one. It happens either because the underlying population eigenvalues are too close to each other, or the concentration c is too high. This phenomenon is illustrated in Figure 4 for the case $\alpha = 1/2$ and $\tau = 2$.

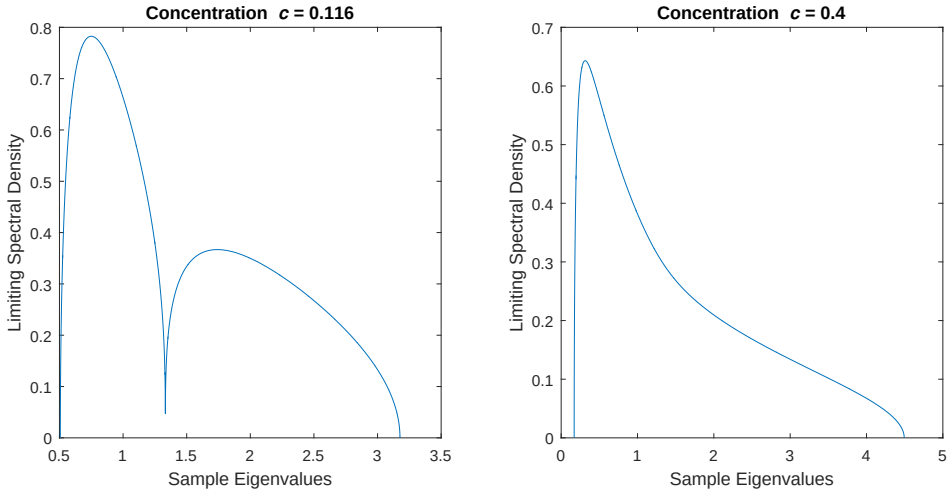

Figure 4: Phase transition: two clusters merge into one.

The tenth and final in our series of qualitative observations, which will all carry over to thegeneral case, is:

10. Clusters that are too close to one another merge for sufficiently high concentration ratios.

On the right panel, the merger of the two clusters is so complete that we no longer even have a bimodal distribution. The naked eye cannot discern that half of the population eigenvalues are equal to one, and the other half to two. Only a purpose-built estimation process resting on advances in probability theory can. This estimation process is what Sections A.6–A.8 detail.

A.6 Limiting Sample Spectral Distribution

In what follows, we describe only the case $0 < c < 1$. The case $c = 0$ is excluded because it is trivial: The sample eigenvalues converge to their population counterparts, and the sample covariance matrix is optimal. The case $c \geq 1$ is excluded because it would render the exposition less fluid, but it poses no great difficulty and can be dealt with just as effectively.[20] One convenience that the assumption $0 < c < 1$ buys us is that the limiting sample spectral distribution admits a countinuous derivative f.

In order to relate F to H quantitatively, a new mathematical object must be introduced: the Stieltjes (1894) transform. The Stieltjes transform of the distribution function F is defined on the half-plane of complex numbers with strictly positive imaginary part $\mathbb{C}^+$ by

$$\forall z \in \mathbb{C}^+ \qquad m_F(z) := \int_{-\infty}^{+\infty} \frac{1}{\lambda - z} dF(\lambda) \ . \tag{A.13}$$

Although it is difficult to visualize a complex function, there is an important result valid specifically for F that will help us gain an intuitive understanding of its Stieltjes transform.

[20] Ledoit and Wolf (2014).

Indeed, F is smooth enough for the limit

$$\lim_{z \in \mathbb{C}^+ \to x} m_F(z) =: \breve{m}_F(x) \tag{A.14}$$

to exist for all $x \in \mathbb{R}$. $\breve{m}_F$ extends the Stieltjes transform from the upper half of the complex plane onto the real line. Now $\breve{m}_F$, being a complex-valued function of real argument, is much easier to comprehend. In particular, its imaginary part is simply the limiting spectral density f divided by π. Thus, if we have $\breve{m}_F$, we get the density f and then also the distribution F (by integration of f). The quantiles of F tell us where the sample eigenvalues are located in the limit.

$\breve{m}_F$ can be deduced from H in only one known way, which is the following. For all $x \in \mathbb{R}$, $m := \breve{m}_F(x)$ is the unique solution in $\mathbb{C}^+$ to the equation

$$\boxed{m = \int_{-\infty}^{+\infty} \frac{1}{\tau\big[1 - c - c\,x\,m\big] - x}\, dH(\tau)} \tag{A.15}$$

Equation (A.15) is the fundamental building block of all research in large-dimensional covariance matrix estimation. Although it may look daunting, and indeed there is generally no closed-form solution, it can be solved numerically in a matter of seconds for $N = 1000$ eigenvalues.[21] This equation has been around for half a century in some form or other, and all researchers in the field are fully confident that "it does exactly what it says on the tin", even though it is difficult to provide intuitive insight.[22]

F is smooth, it is more spread out than H, and the excess spread increases in c. In the limit, as $c \to 0$, we can recognize on the right-hand side the Stieltjes transform of H, so F becomes identical to H, which was to be expected from finite-dimensional asymptotics. All ten of the qualitative observations gathered from the study of simple cases in Sections A.4–A.5 carry over to the solution of equation (A.15). Indeed they pretty much encompass all the intuition that can be extracted from this equation. Figure 5 provides an illustration of the difference between F and H.

[21] Ledoit and Wolf (2016)

[22] See Marčenko and Pastur (1967), Silverstein and Bai (1995), Silverstein (1995), as well as the authoritative monograph by Bai and Silverstein (2010).

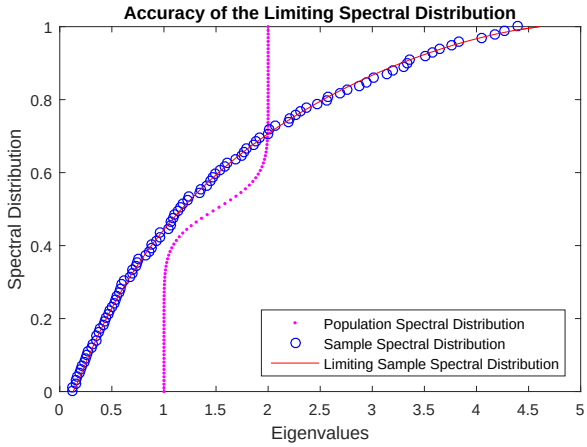

Figure 5: Three spectral distributions for $N = 100$ and $T = 200$. The population spectral distribution is the Beta$(0.1, 0.1)$ distribution shifted so the support is $[1, 2]$. Circles show the sample spectral distribution from one Monte Carlo simulation. One can observe that the sample eigenvalues are nowhere near their population counterparts, but their location is well predicted by the limiting spectral distribution F that comes out of equation (A.15).

A.7 Discretization

Although equation (A.15) theoretically solves the problem, it is not formulated in a directly usable way, as it relates the limiting spectral distributions F and H, whereas ideally we would want to relate the sample eigenvalues $(\lambda_{1,T}, \ldots, \lambda_{N,T})$ to their population counterparts $(\tau_{1,T}, \ldots, \tau_{N,T})$. A practical implementation is achieved through discretization.

Start from a family of N population eigenvalues $\boldsymbol{\tau} := (\tau_{1,T}, \ldots, \tau_{N,T})$, which can be any vector in $(0, +\infty)^N$. Construct the population spectral distribution H_T as per equation (A.8) and inject it into equation (A.15), replacing c with the ratio N/T. A trivial simplification shows that, for all $x \in \mathbb{R}$, $m := \breve{m}_F(x)$ is the unique solution in $\mathbb{C}^+$ to the equation

$$m = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\tau_{i,T} \left[1 - \frac{N}{T} - \frac{N}{T} x\, m \right] - x} \,. \tag{A.16}$$

Having solved for the Stieltjes transform $\breve{m}_F$ numerically, we multiply its imaginary part by π to obtain the limiting sample spectral density f. Integration yields the limiting sample spectral distribution F. Finally, we can invert the function F to compute the N distribution quantiles.

This is exactly how the QuEST function is constructed. It maps an N-dimensional vector $\boldsymbol{\tau}$ of population eigenvalues into another N-dimensional vector $\left(q^1_{N,T}(\boldsymbol{\tau}), q^2_{N,T}(\boldsymbol{\tau}), \ldots, q^N_{N,T}(\boldsymbol{\tau})\right)$ that represents a deterministic equivalent of the sample eigenvalues, by discretizing equation (A.15). The output (quantiles of F) can be interpreted as the expectation of the sample eigenvalues, although this correspondence is rigorous only in the large-dimensional asymptotic

limit.

A.8 Recovering Population Eigenvalues

As can be gathered from the above exposition, it is more straightforward to go from population to sample eigenvalues than the other way around. This is an intrinsic feature of equation (A.15), which is our only tractable hook into the underlying mathematical truth.

However, given that equation (A.15) has been made more practical through discretization inside the QuEST function as seen in Section A.7, inverting it becomes a simple numerical problem. Find the vector $\boldsymbol{\tau} \in (0, +\infty)^N$ such that the function's output $\left(q^1_{N,T}(\boldsymbol{\tau}), q^2_{N,T}(\boldsymbol{\tau}), \ldots, q^N_{N,T}(\boldsymbol{\tau})\right)$ matches most closely the observed sample eigenvalues $(\lambda_{1,T}, \ldots, \lambda_{N,T})$. Any capable off-the-shelf nonlinear optimizer can solve this problem. Two recommended optimizers are Stanford Business Software's `SNOPT` and Matlab's `fmincon`; they both can easily handle dimensions up to $N = 1000$ within a reasonable amount of time, a few minutes at most. In a nutshell, the problem of recovering the population eigenvalues from the sample eigenvalues has been resolved by numerically inverting the multivariate function that discretizes equation (A.15). Figure 6 illustrates the accuracy of this procedure.

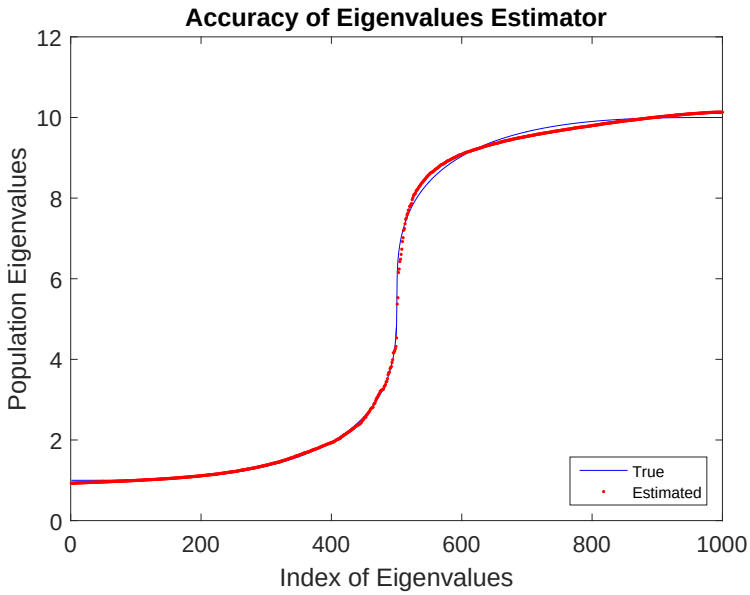

Figure 6: The matrix dimension is $N = 1000$ and the sample size is $T = 3000$. This graph is based on a single Monte Carlo simulation. Numerically inverting the discretized version of equation (A.15) asymptotically recovers the population eigenvalues.

A.9 Nonlinear Shrinkage

At this juncture, it might be tempting to conclude that replacing the observed sample eigenvalues with the estimated population eigenvalues that come from inverting the QuEST

function yields the optimal estimator of the covariance matrix. However this is not the case. The population eigenvalues (or consistent estimators thereof) are only optimal when recombined with the population eigenvectors. The latter are unobservable and, unlike the eigenvalues, there is no hope of recovering them through some advanced mathematics. The reason is that they live in a space of dimension $N(N-1)/2$, which is infinitely too large given that we only collect $N \times T$ noisy data points, T being of the same order of magnitude as N.

In the terminology of Section A.1, we do not want $\lambda_{i,T}$, the in-sample variance of the sample eigenvector $u_{i,T}$; but we also do not want $\tau_{i,T}$, the out-sample variance of the population eigenvector $v_{i,T}$. This is because we do not have $v_{i,T}$. What we want is a hybrid: $u'_{i,T}\Sigma u_{i,T}$, the out-sample variance of the sample eigenvector $u_{i,T}$. This quantity is estimated consistently under general asymptotics by the following "nonlinear shrinkage" formula:

$$u'_{i,T}\Sigma u_{i,T} \approx \frac{\lambda_{i,T}}{\left|1 - \frac{N}{T} - \frac{N}{T}\,\lambda_{i,T}\,\breve{m}_F(\lambda_{i,T})\right|^2}\ , \tag{A.17}$$

where $|\cdot|$ denotes the modulus of a complex number, and $\breve{m}_F$ is the function defined in (A.16).[23] Similar to equation (A.15), it is hard to give intuition: this is just what comes out of the underlying mathematics. We are fortunate to have any explicit equation at all; it basically comes from a generalization of equation (A.15). Extensive Monte Carlo simulations confirm the accuracy of this formula.[24] One can see that when N/T is negligible, there is a negligible amount of shrinkage, as expected from finite-dimensional asymptotics.

Calling the right-hand side $\psi_{i,T}$ and injecting it into equation (A.7) yields an estimator of the covariance matrix that improves upon the sample covariance matrix when the dimension N is not negligible with respect to the sample size T. A graphical illustration is given in Figure 7.

[23] Ledoit and Péché (2011, Theorem 3)
[24] Ledoit and Wolf (2012, Section 6)

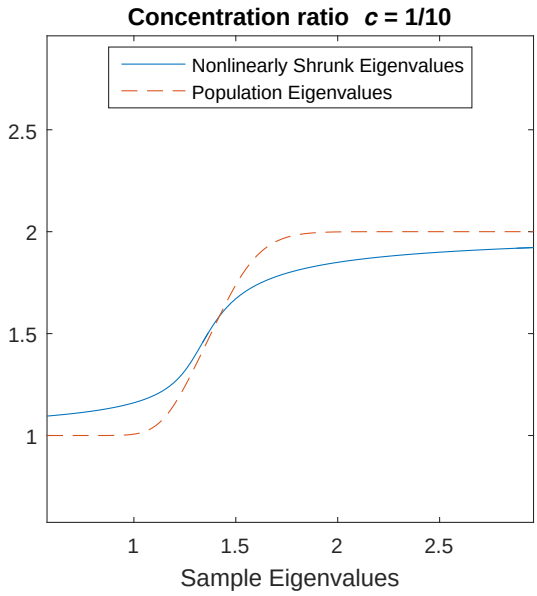

Figure 7: Nonlinearly shrunk eigenvalues as a function of population eigenvalues. The population eigenvalues come from the Beta$(0.1, 0.1)$ distribution, shifted so the support is $[1, 2]$. The optimal shrinkage transformation is highly nonlinear. The population eigenvalues are also plotted for reference as the dashed line. They are not the same as the shrunk eigenvalues, and are more spread out.

To summarize, the overall procedure consists of three consecutive steps:

Step 1 Given the sample eigenvalues $(\lambda_{1,T}, \ldots, \lambda_{N,T})$, invert the QuEST function defined in Section A.7 to obtain consistent estimates of the population eigenvalues;

Step 2 Plug the (estimated) population eigenvalues into equation (A.16) to compute the complex-valued function $\breve{m}_F$;

Step 3 Replace the sample eigenvalues with the nonlinear shrinkage formula on the right-hand side of equation (A.17), while preserving the sample eigenvectors.

All of this is handled automatically by the QuEST software, available from one of the authors' web page.[25] Ledoit and Wolf (2014) prove that the resulting covariance matrix estimator is optimal for portfolio selection under general asymptotics within the rotation-equivariant class.

[25] http://www.econ.uzh.ch/en/people/faculty/wolf/publications.html.

B The BEKK-NL Model

While the present paper focuses on the DCC model, which works at the level of correlations and devolatilized returns, an alternative approach involves the BEKK model presented in Engle and Kroner (1995), which works in an analogous way at the level of covariances and straight returns. The most scalable version of the BEKK model, and the one most similar to the particular version of DCC presented in Section 2, is the one with scalar dynamics and covariance targeting. Using the notation of Section 2.1, equations (2.2)–(2.4) are replaced with

$$H_t = (1 - \ddot{\alpha} - \ddot{\beta})\Sigma + \ddot{\alpha}\,\mathbf{r}_{t-1}\mathbf{r}_{t-1}' + \ddot{\beta}\,H_{t-1}\ , \tag{B.1}$$

where Σ is the unconditional covariance matrix and $(\ddot{\alpha}, \ddot{\beta})$ are BEKK dynamic parameters analogous to (α, β), but in covariance space instead of correlation space.

BEKK is simpler compared to DCC, but it does not handle well investment universes that include correlated assets with wildly different variances. For example, if we have gold and short-term government bonds, both of which can be considered 'safe havens' in times of financial crises, volatilities vary by one or two orders of magnitudes, so putting them on the same footing (as BEKK does) may not be the best modeling strategy. To put it another way, if we replace one asset, say the S&P 500 index, with a 2-to-1 leveraged version of itself (and such ETFs do exist), then the set of investment opportunities remains the same and DCC adapts automatically; whereas any portfolio allocation based on BEKK will be impacted. In other words, BEKK will be favored by a homogenous, unlevered investment universe.

Estimation of the BEKK model in large dimensions using the composite likelihood method is described by Pakel et al. (2014, Example 2.1 and Section 3). The BEKK-NL model is obtained by inserting the nonlinear shrinkage estimator of the covariance matrix developed by Ledoit and Wolf (2012, 2014) in place of the covariance targeting matrix Σ.