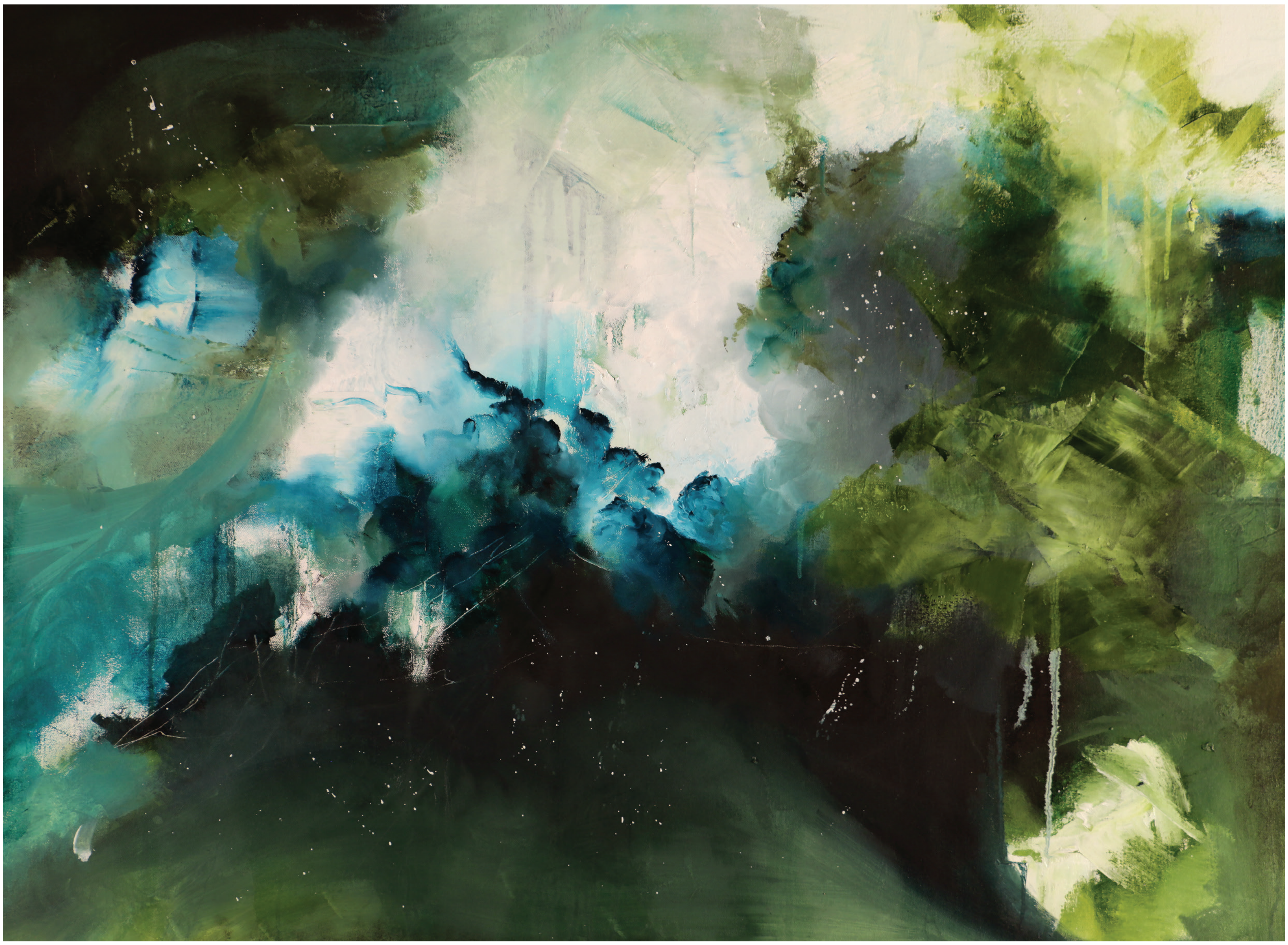
Risk publications

REPRINTED FROM
Risk
RISK MANAGEMENT • DERIVATIVES • REGULATION

Risk.net April 2016

Cutting edge
Refining Markowitz



Cleaning correlation matrices

A new cleaning recipe that outperforms all existing estimators in terms of the out-of-sample risk of synthetic portfolios

Cleaning correlation matrices

The determination of correlation matrices is typically affected by in-sample noise. Joël Bun, Jean-Philippe Bouchaud and Marc Potters propose a simple, yet optimal, estimator of the true underlying correlation matrix and show that this new cleaning recipe outperforms all existing estimators in terms of the out-of-sample risk of synthetic portfolios

The concept of correlations between different assets is a cornerstone of Markowitz's optimal portfolio theory, especially for risk management purposes (Markowitz 1968). In a nutshell, correlations measure the tendency of different assets to vary together, and it is well known that large losses at a portfolio level are indeed mostly due to correlated moves of its constituents (see, for example, Bouchaud & Potters 2003). Efficient and robust diversification is needed to alleviate such events. Markowitz's theory is the simplest and best known method for constructing a portfolio with a given level of risk, using the correlations between assets as inputs. However, the Markowitz solution results in over-allocation on low variance modes (ie, eigenvectors) of the correlation matrix. It is therefore of crucial importance to use a correlation matrix that faithfully represents future, and not past, risks – otherwise the over-allocation on spurious low risk combinations of assets might prove disastrous (see, for example, Bouchaud & Potters (2011) for a detailed discussion of this point).

The question of building reliable estimators of covariance or of correlation matrices has a long history in finance, and more generally in multivariate statistical analysis. The problem comes from the high-dimensionality of these matrices, as in many applications. When the size of the time series T is very large, each of the coefficients of the covariance/correlation matrix can be estimated with negligible error (if is assumed not to vary with time). But if N is also large and of the order of T, as is often the case in finance, the large number of noisy variables creates important systematic errors in the computation of the inverse of the matrix, which is a direct input of Markowitz's optimisation formula. These systematic errors lead to sub-optimal portfolios with grossly underestimated out-of-sample risk (Bouchaud & Potters 2011).

The aim of this note is to provide the reader with a short review of the different cleaning 'recipes' that have been proposed in the literature to cope with the problem of in-sample noise in the determination of correlation matrices. While some of these recipes are simple and well known, recent progress has been made that allows one to derive an optimal and fully observable estimator of the 'true' underlying correlation matrix, valid in the limit of large matrices. We want to popularise these new results and test the corresponding estimators on real financial data, with very satisfactory benefits: both the realised risk and risk-adjusted returns of synthetic portfolios are improved for a wide class of investment strategies.

Setting the stage

We consider a universe made of N different financial assets that we observe at (say) the daily frequency, defining a vector of returns $\boldsymbol{r}_t = (r_{1t}, r_{2t}, \dots, r_{Nt})$ for each day $t = 1, \dots, T$. Because the aim of the present study is to focus specifically on correlations and not on volatilities, we standardise these returns as follows: (i) we remove the sample mean of each asset; (ii) we normalise each return by an estimate $\hat{\sigma}_{it}$ of its daily volatility, $\tilde{r}_{it} := r_{it}/\hat{\sigma}_{it}$. There are many possible choices for $\hat{\sigma}_{it}$, based on, for example, Garch or Figarch models historical returns, or simply implied volatilities from option markets, and the reader can choose his/her favourite estimator which can easily be combined with the correlation matrix cleaning schemes discussed below. For simplicity, we have chosen here the cross-sectional daily volatility, that is:

$$\hat{\sigma}_{it} = \sqrt{\sum_j r_{jt}^2}$$

to remove a substantial amount of non-stationarity in the volatilities. The final standardised return matrix $\boldsymbol{X} = (X_{it}) \in \mathbb{R}^{N \times T}$ is then given by $X_{it} := \tilde{r}_{it}/\sigma_i$, where σ_i is the sample estimator of the volatility of $\tilde{r}_i$, which is now, to a first approximation, stationary.

The most common (and simplest) estimator of the 'true' underlying correlation matrix (that we henceforth denote by $\boldsymbol{C}$) is to use the sample estimator:

$$\boldsymbol{E} := \frac{1}{T}\boldsymbol{X}\boldsymbol{X}^* \tag{1}$$

We will use the following notation throughout:

$$\boldsymbol{E} = \sum_{k=1}^{N} \lambda_k \boldsymbol{u}_k \boldsymbol{u}_k^* \tag{2}$$

for the eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_N \geqslant 0$ and the associated eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_N$ of $\boldsymbol{E}$. When $q = N/T \to 0$, ie, when the data set is very long, one expects that $\boldsymbol{E} \to \boldsymbol{C}$, whereas important distortions survive when $q = O(1)$, even if $T \to \infty$. In fact, since the work of Marčenko & Pastur (1967) decades ago, one can show that the spectrum of $\boldsymbol{E}$ is a broadened version of that of $\boldsymbol{C}$, with an explicit q-dependent formula relating the two. In particular, small λ_is are too small and large λ_is are too large compared to the true eigenvalues $\boldsymbol{C}$. Thus, recalling that Markowitz's optimal strategy overweight low variance modes (see, for example, Bouchaud & Potters 2011), we understand why using the sample estimator $\boldsymbol{E}$ can lead to disastrous results, and why some cleaning procedure should be applied to $\boldsymbol{E}$ before any application to portfolio construction. Note that the Marčenko and Pastur results do not require multivariate normality of the returns, which can have fat-tailed distributions. In fact, the above normalisation by the cross-sectional volatility can be seen as a proxy for a robust estimator of the covariance matrix (see, for example, Couillet, Kammoun & Pascal 2016).

Five cleaning recipes

We list here five estimators that have been proposed in the literature, the last one being the most recent and, as we shall see in the next section, also the most efficient.[1] In the following, $\boldsymbol{\Xi}$ denotes a cleaned estimator of the true correlation matrix $\boldsymbol{C}$ and $\alpha \in [0, 1]$ is a parameter that must be determined through optimisation or analytical arguments.

■ 1. *Basic linear shrinkage*: a linear combination of the sample estimate and the identity matrix:

$$\Xi^{\text{bas.}} := \alpha \boldsymbol{E} + (1-\alpha) I_N \tag{3}$$

This is probably the oldest method proposed in the literature (Haff 1980). In a financial context, this method can be seen as a heuristic way to control the diversification of the optimal Markowitz portfolio (Bouchaud & Potters 2003).

■ 2. *Advanced linear shrinkage*: a linear combination of the sample estimate and a correlation matrix containing the market mode:

$$\Xi^{\text{adv.}} := \alpha \boldsymbol{E} + (1-\alpha)[(1-\rho) I_N + \rho \boldsymbol{e}\boldsymbol{e}^*] \tag{4}$$

where $\boldsymbol{e} = (1, 1, \dots, 1)$ is the unit vector and ρ an average correlation that can be optimally and self-consistently estimated (see Ledoit & Wolf 2003). It turns out, however, that (4) provides similar performance to (3) (Ledoit & Wolf 2014).

■ 3. *Eigenvalues clipping* (Bouchaud & Potters 2011): keep the $\lceil N\alpha \rceil$ top eigenvalues and shrink the others to a constant γ that preserves the trace, $\text{Tr}(\Xi^{\text{clip.}}) = \text{Tr}(\boldsymbol{E}) = N$:

$$\Xi^{\text{clip.}} := \sum_{k=1}^{N} \xi_k^{\text{clip.}} \boldsymbol{u}_k \boldsymbol{u}_k^*, \quad \xi_k^{\text{clip.}} := \begin{cases} \lambda_k & \text{if } k \leqslant \lceil N\alpha \rceil \\ \gamma & \text{otherwise} \end{cases} \tag{5}$$

A simple (but ad hoc) procedure to choose α is to assume that all empirical eigenvalues beyond the Marčenko and Pastur upper edge (see Box 1) can be deemed to contain some signal and are therefore kept without change. However, this cleaning overlooks the fact that the large empirical eigenvalues are overestimated (see below).

■ 4. *Eigenvalues substitution*: sample eigenvalues λ_k are replaced by an estimation μ_k of the true ones, obtained by inverting the general Marčenko and Pastur equation relating the spectrum of $\boldsymbol{C}$ to that of $\boldsymbol{E}$:

$$\Xi^{\text{sub.}} := \sum_{k=1}^{N} \mu_k^{\text{sub.}} \boldsymbol{u}_k \boldsymbol{u}_k^* \tag{6}$$

The inversion of the Marčenko-Pastur equation is numerically unstable and requires either a parametric description of the spectrum of $\boldsymbol{C}$ (eg, a power law, as proposed in Bouchaud & Potters (2011)) or some prior knowledge of the location of the 'true' eigenvalues as in El Karoui (2008). This method is also sub-optimal on theoretical grounds as it fails to take into account the noise in the eigenvector determination.

■ 5. *Rotationally invariant, optimal shrinkage*: this method was first proposed by Ledoit & Péché (2011), who realised that one can actually compute the 'overlap' between true and sample eigenvectors, precisely the ingredient missing from method 4. The method was extended to discrete eigenvalues (including outliers) in Bun & Knowles (2016) which also provided a practical implementation. The method was further studied in Bun, Bouchaud & Potters (2016), where, in particular, an ad hoc method was proposed to correct for a systematic bias for small eigenvalues. For large N, the optimal rotational invariant estimator (RIE) of $\boldsymbol{C}$ reads:

$$\Xi^{\text{RIE}} := \sum_{k=1}^{N} \xi_k^{\text{RIE}} \boldsymbol{u}_k \boldsymbol{u}_k^*$$

with (Bun & Knowles 2016):

$$\xi_k^{\text{RIE}} := \frac{\lambda_k}{|1 - q + q z_k s(z_k)|^2} \tag{7}$$

where $s(z) := N^{-1} \text{tr}(z I_N - \boldsymbol{E})^{-1}$ and $z_k = \lambda_k - \text{i}\eta$ (see below for the chosen value of η). The rotational invariant hypothesis on an estimator Ξ assumes that one does not have any prior knowledge of the structure of the true eigenvectors, and therefore that the best one can do is to keep the eigenvectors $\boldsymbol{u}_i$ of $\boldsymbol{E}$ untouched.[2]

We will not discuss methods 2 and 4 any further, because they turn out to be less efficient than the simpler schemes 1 and 3 on real data. Also, we do not consider the algorithm of Ledoit & Wolf (2014) for method 5 because it requires heavy numerical operations with $O(N)$ parameters to fit. Moreover, the proposed quantile representation rules out a priori outliers (see Bun & Knowles (2016) for an extended discussion on this topic). Equation (7), on the other hand, is very appealing from a theoretical point of view. The optimality of (7) stands in the following sense. Suppose that we knew $\boldsymbol{C}$ exactly, but constrain ourselves to estimate it in the eigenbasis $\{\boldsymbol{u}_k\}$ of $\boldsymbol{E}$. The optimal estimator of the eigenvalues of $\boldsymbol{C}$ would then read:

$$\xi_k^{\text{ora.}} := \langle \boldsymbol{u}_k, \boldsymbol{C}\boldsymbol{u}_k \rangle \tag{8}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. Note that the superscript 'ora.' stands for oracle, which is the optimal estimator that one would obtain if one knew the true correlation matrix $\boldsymbol{C}$ (which is of course not the case, since we are actually attempting to estimate $\boldsymbol{C}$!). But, as was shown in Bun & Knowles (2016), (7) satisfies, roughly speaking:

$$|\xi_k^{\text{RIE}} - \langle \boldsymbol{u}_k, \boldsymbol{C}\boldsymbol{u}_k \rangle| = O(T^{-1/2}) \quad \forall k \tag{9}$$

for $q = O(1)$. This means that the estimator ξ_k^{RIE} converges, for large enough T, to the optimal 'oracle' estimator (8).

For practical applications, however, (7) contains a parameter η that must be chosen to be small, but at the same time such that $N\eta \gg 1$. A good trade-off is to set $\eta = N^{-1/2}$ (see Bun & Knowles 2016). Unfortunately, in cases where N is not extremely large (say $N = 400$), this corresponds to $\eta = 0.05$, which is not very small, and this induces a systematic downward bias in the estimator of small eigenvalues (see Bun, Bouchaud & Potters (2016) for an extended discussion of this point). This bias can be explicitly calculated in simple cases and suggests the following heuristic correction for any k:[3]

$$\hat{\xi}_k := \xi_k^{\text{RIE}} \times \max(1, \Gamma_k) \tag{10}$$

[1] There are of course many other estimators in the literature, but we only keep estimators that seem relevant for this problem, at least to our eyes.

[2] This assumption is certainly not true of the market mode itself, and possibly of all the spikes corresponding to sectors. However, this is the simplest assumption that should be reasonable for the bulk of $\boldsymbol{C}$ and that can be improved upon (see, for example, Monasson & Villamaina 2015).

[3] This correction is fit for the case of stock markets on which we focus in this note. In the case of futures markets, the presence of very strongly correlated contracts (ie, two different maturities for the same underlying) leads to very small true eigenvalues of the correlation matrix, and requires a more sophisticated regularisation (see Bun, Bouchaud & Potters 2016).

where $z_k = \lambda_k - \mathrm{i}/\sqrt{N}$ is a complex number and Γ_k is the correction factor given in (17) in Box 1. Note that $\hat{\xi}_k$ is always greater than or equal to ξ_k^{RIE}: this allows one to correct nearly perfectly a systematic downward bias in the trace of the RIE correlation matrix Ξ^{RIE}.

Test of the RIE on financial data

Interestingly, the oracle estimator (8) can be estimated empirically and used to directly test the accuracy of the debiased RIE (10). The trick is to remark that the oracle eigenvalues (8) can be interpreted as the 'true' risk associated to a portfolio whose weights are given by the ith eigenvector. Hence, assuming that the data generating process is stationary, we estimate the oracle estimator through the realised risk associated to such eigen-portfolios (Pafka & Kondor 2003). More precisely, we divide the total length of our time series T_{tot} into n consecutive, non-overlapping samples of length T_{out}. The 'training' period has length T, so n is given by:

$$n := \left\lfloor \frac{T_{\mathrm{tot}} - T - 1}{T_{\mathrm{out}}} \right\rfloor \tag{11}$$

The oracle estimator (8) is then computed as:

$$\xi_i^{\mathrm{ora.}} \approx \frac{1}{n} \sum_{j=0}^{n-1} \mathcal{R}^2(t_j, \boldsymbol{u}_i), \quad i = 1, \dots, N \tag{12}$$

for $t_j = T + j \times T_{\mathrm{out}} + 1$ and $\mathcal{R}^2(t, \boldsymbol{w})$ denotes the out-of-sample variance of the returns of portfolio $\boldsymbol{w}$ built at time t, that is to say:

$$\mathcal{R}^2(t, \boldsymbol{w}) := \frac{1}{T_{\mathrm{out}}} \sum_{\tau=t+1}^{t+T_{\mathrm{out}}} \left(\sum_{i=1}^{N} \boldsymbol{w}_i X_{i\tau} \right)^2 \tag{13}$$

where $X_{i\tau}$ denotes the rescaled realised returns defined as in the second section.[4] This implies that $\sum_{i=1}^{N} \mathcal{R}^2(t, \boldsymbol{u}_i) = N$ for any time t.

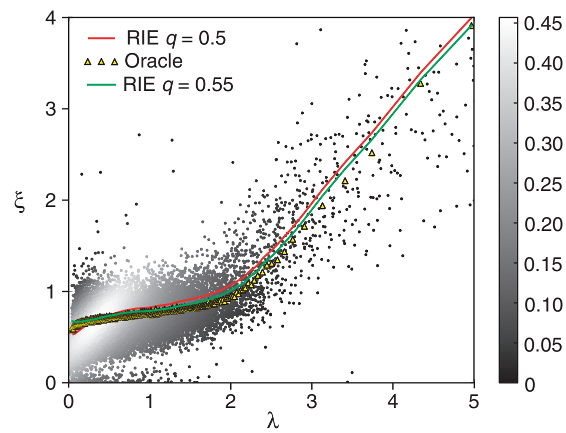
For our simulations, we consider an international pools of stocks with daily data.

- US: 500 most liquid stocks during the training period of the S&P 500 from 1966 until 2012.
- Japan: 500 most liquid stocks during the training period of the all-shares TOPIX index from 1993 until 2015.
- Europe: 500 most liquid stocks during the training period of the Bloomberg European 500 index from 1996 until 2015.

We chose $T = 1{,}000$ (4 years) for the training period, ie, $q = 0.5$, and $T_{\mathrm{out}} = 60$ (three months) for the out-of-sample test period. We plot our results for the US data in figure 1. The results are, we believe, quite remarkable: the RIE formula (10) (red dashed line) tracks very closely the average realised risk (the yellow triangles), specially in the region where there are a lot of eigenvalues. A similar agreement is found in the other pools of stocks as well (see Bun, Bouchaud & Potters 2016). Interestingly, one can choose an effective observation ratio $q_{\mathrm{eff}} > q$ for which the dressed RIE and the oracle estimate nearly coincide (the green line). This effect may be understood by the presence of autocorrelations in the stock returns that are not taken into account in the model of $\boldsymbol{E}$. The presence of autocorrelations has been shown to widen the spectrum of the sample matrix $\boldsymbol{E}$ (Burda, Jurkiewicz & Wacław 2005). Since the agreement reached with the naive value $q = N/T$ is already very good, we leave the problem of calibrating q_{eff} on empirical data for future research.

[4] Again, as we are primarily interested in estimating correlations and not volatilities, both our in-sample and out-of-sample returns are made approximately stationary and normalised.

1 Comparison of the dressed RIE (20) with the proxy (12) using 500 US stocks from 1970 to 2012



The points represent the density map of each realisation of (12) and the colour code indicated the density of data points. The average dressed RIE is plotted with the red dashed line and the average realised risk in yellow. We also provide the prediction of the dressed RIE with an effective observation ratio q_{eff} which is slightly bigger than q (green plain line). The agreement between the green line and the average oracle estimator (the yellow triangles) is quite remarkable

Optimal shrinkage and portfolio optimisation

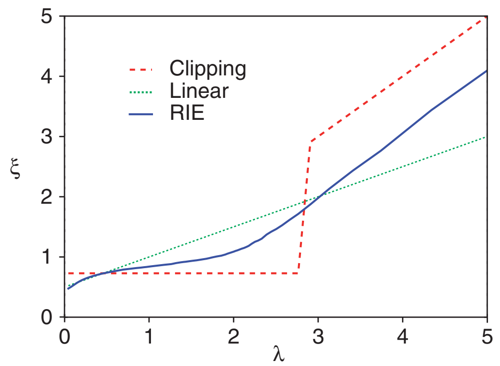
It is interesting to compare the optimal shrinkage function that maps the empirical eigenvalue λ_i onto its 'cleaned' version ξ_i. We show these functions in figure 2 for the three schemes we retained, ie, linear shrinkage (1), clipping (3) and RIE (5), using the same data set as in figure 1. This figure clearly reveals the difference between the three schemes. For clipping (the red dashed line), the intermediate eigenvalues are quite well estimated but the convex shape of the optimal shrinkage function for larger λ_is is not captured. Furthermore, the larger eigenvalues are systematically overestimated. For the linear shrinkage (the green dotted line), it is immediate from figure 2 why this method is not optimal for any shrinkage parameters $\alpha \in [0, 1]$ (that fixes the slope of the line).

We now turn to optimal portfolio construction using the above three cleaning schemes, with the aim of comparing the (average) realised risk of optimal Markowitz portfolios constructed as:

$$\boldsymbol{w} := \frac{\hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{g}}{\boldsymbol{g}^* \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{g}} \tag{14}$$

where $\boldsymbol{g}$ is a vector of predictions and $\hat{\boldsymbol{\Sigma}}$ is the cleaned covariance matrix $\hat{\Sigma}_{ij} := \hat{\sigma}_i \hat{\sigma}_j \hat{\Xi}_{ij}$ for any $i, j \in [\![1, N]\!]$. Note again that we consider here returns normalised by an estimator of their volatility: $\tilde{r}_{it} = r_{it}/\hat{\sigma}_{it}$. This means that our tests are immune to an overall increase or decrease of the the volatility in the out-of-period, and are

2 Comparison of the debiased RIE (10) (the blue line) with clipping at the edge of the Marčenko-Pastur (the red dashed line) and the linear shrinkage with $\alpha = 0.5$ (green dotted line)



We use the same data set as in figure 1

only sensitive to the quality of the estimator of the correlation matrix itself.

In order to ascertain the robustness of our results in different market situations, we consider the following four families of predictors $\boldsymbol{g}$.

1. The minimum variance portfolio, corresponding to $g_i = 1\ \forall i \in [\![1, N]\!]$.

2. The omniscient case, ie, when we know exactly the realised returns on the next out-of-sample period for each stock. This is given by $g_i = \mathcal{N}\tilde{r}_{i,t}(T_{\text{out}})$, where $r_{i,t}(\tau) = (P_{i,t+\tau} - P_{i,t})/P_{i,t}$ with $P_{i,t}$ the price of the ith asset at time t and $\tilde{r}_{it} = r_{it}/\hat{\sigma}_{it}$.

3. Mean reversion on the return of the last day: $g_i = -\mathcal{N}\tilde{r}_{it}$ $\forall i \in [\![1, N]\!]$.

4. Random long-short predictors where $\boldsymbol{g} = \mathcal{N}\boldsymbol{v}$, where $\boldsymbol{v}$ is a random vector uniformly distributed on the unit sphere.

The normalisation factor $\mathcal{N} := \sqrt{N}$ is chosen to ensure $\boldsymbol{w}_i \sim O(N^{-1})$ for all i. The out-of-sample risk $\mathcal{R}^2$ is obtained from (13) by replacing the matrix $\boldsymbol{X}$ by the normalised return matrix $\tilde{\boldsymbol{R}}$ defined by $\tilde{\boldsymbol{R}} := (\tilde{r}_{it}) \in \mathbb{R}^{N\times T}$. We report the average out-of-sample risk for these various portfolios in table A, for the three above cleaning schemes and the three geographical zones, keeping the same value of T (the learning period) and T_{out} (the out-of-sample period) as above. The linear shrinkage estimator uses a shrinkage intensity α estimated from the data following Ledoit & Wolf (2003) (LW). The eigenvalues clipping procedure uses the position of the Marčenko-Pastur edge, $(1+\sqrt{q})^2$, to discriminate between meaningful and noisy eigenvalues. The second to last line gives the result obtained by taking the identity matrix (total shrinkage, $\alpha = 0$) and the last one is obtained by taking the uncleaned, in-sample correlation matrix ($\alpha = 1$).

These tables reveal that (i) it is always better to use a cleaned correlation matrix: the out-of-sample risk without cleaning is, as expected, always higher than with any of the cleaning schemes, even with four years of data; and (ii) in all cases but one (minimum risk portfolio in Japan, where the LW linear shrinkage outperforms), our debiased RIE is providing the lowest out-of-sample risk, independently of the type of predictor used. Note that these results are statistically significant everywhere, except perhaps for the minimum variance strategy with Japanese stocks: see the standard errors that are given between parenthesis in table A. Moreover, the result is robust in the dimension N as indicated in table B for the mean reversion strategy. For the other strategies, some fluctuations can be observed for $N = 100$ but the results are identical to those of table A for $N \geqslant 200$ (see Bun, Bouchaud & Potters 2016).

A. Annualised out-of-sample average volatility (in %) of the different strategies (standard errors are given in brackets)

	US	Japan	Europe
Minimum variance portfolio			
RIE	*10.4* (0.12)	30.0 (2.9)	*13.2* (0.12)
Clipping MP	10.6 (0.12)	30.4 (2.9)	13.6 (0.12)
Linear LW	10.5 (0.12)	*29.5* (2.9)	13.2 (0.13)
Identity	15.0 (0.25)	31.6 (2.92)	20.1 (0.25)
In sample	11.6 (0.13)	32.3 (2.95)	14.6 (0.2)
Omniscient predictor			
RIE	*10.9* (0.15)	*12.1* (0.18)	*9.38* (0.18)
Clipping MP	11.1 (0.15)	12.5 (0.2)	11.1 (0.21)
Linear LW	11.1 (0.16)	12.2 (0.18)	11.1 (0.22)
Identity	17.3 (0.24)	19.4 (0.31)	17.7 (0.34)
In sample	13.4 (0.25)	14.9 (0.28)	12.1 (0.28)
Mean reversion predictor			
RIE	*7.97* (0.14)	*11.2* (0.20)	*7.85* (0.06)
Clipping MP	8.11 (0.14)	11.3 (0.21)	9.35 (0.09)
Linear LW	8.13 (0.14)	11.3 (0.20)	9.26 (0.09)
Identity	17.7 (0.23)	24.0 (0.4)	23.5 (0.2)
In sample	9.75 (0.28)	15.4 (0.3)	9.65 (0.11)
Uniform predictor			
RIE	*1.30* (8e-4)	*1.50* (1e-3)	*1.23* (1e-3)
Clipping MP	1.31 (8e-4)	1.55 (1e-3)	1.32 (1e-3)
Linear LW	1.32 (8e-4)	1.61 (1e-3)	1.27 (1e-3)
Identity	1.56 (2e-3)	1.86 (2e-3)	1.69 (2e-3)
In sample	1.69 (1e-3)	2.00 (2e-3)	2.7 (0.01)

B. Annualised out-of-sample average volatility (in %) of the mean reversion strategy as a function of N with $q = 0.5$

	N = 100	200	300	400
US				
RIE	*21.9*	*11.7*	*10.0*	*8.51*
Clipping MP	22.0	11.9	10.1	8.62
Linear LW	22.6	12.1	10.3	8.74
Identity $\alpha = 0$	43.2	27.3	21.1	19.3
In sample $\alpha = 1$	30.0	15.7	13.5	11.4
Japan				
RIE	*24.5*	*13.8*	*12.5*	*10.5*
Clipping MP	24.4	14.1	13.0	10.9
Linear LW	25.5	14.1	12.7	10.7
Identity $\alpha = 0$	64.0	43.9	41.3	33.6
In sample $\alpha = 1$	31.7	18.5	15.8	13.0
Europe				
RIE	*26.3*	*15.4*	*10.1*	*7.1*
Clipping MP	27.1	15.6	10.1	7.5
Linear LW	27.2	15.9	10.2	7.5
Identity $\alpha = 0$	65.8	41.4	31.9	28.0
In sample $\alpha = 1$	34.7	20.0	11.3	8.0

Summary

To summarise, we have briefly reviewed several cleaning schemes for noisy financial correlation matrices that have been proposed in the last 20 years. The latest one, proposed very recently in Bun & Knowles (2016) as an extension of previous work by Ledoit & Péché, is remarkably good at predicting the out-of-sample risk of individual eigenportfolios (ie, portfolios with weights given by the eigenvectors of the correlation matrices). This method outperforms all other methods to date at minimising the out-of-sample (realised) risk of various types of Markowitz optimal portfolios (once volatilities have been factored out). We are confident that this method will become a standard in the coming years; it also suggests various routes for further improvements, including modelling a genuine evolution of the correlation matrix itself, beyond measurement noise. ■

Joël Bun is a PhD candidate at the Université Paris-Saclay and at the Léonard de Vinci Pôle Universitaire, and an analyst at Capital Fund Management in Paris. Jean-Philippe Bouchaud is chairman and chief scientist and Marc Potters is co-CEO and head of research of Capital Fund Management. **We are indebted to R. Benichou, A. Beveratos, R. Chicheportiche, A. Rej, E. Sérié and G. Simon for many insightful discussions and help in the preparation of this work. We also acknowledge useful conversations with R. Couillet on robust estimators. Finally, we warmly thank A. Knowles for his contribution to this topic that led to the paper Bun & Knowles (2016).**
Email: joel.bun@gmail.com,
jean-philippe.bouchaud@cfm.fr,
marc.potters@cfm.fr.

Box 1. Implementation of the debiased RIE

We present the complete procedure to construct our optimal estimator of $\boldsymbol{E}$ mostly based on the optimal shrinkage function (7). This algorithm can easily be implemented in any computer language that can handle complex numbers.

Given the eigenvalues $[\lambda_k]_{k=1}^N$ of the sample correlation matrix $\boldsymbol{E}$ and $q = N/T$, we define the complex variable $z_k = \lambda_k - \mathrm{i}/\sqrt{N}$ for any $k \in [\![1, N]\!]$. Then, the cleaning scheme reads:

- 1. We compute for any $k \in [\![1, N]\!]$:

$$\xi_k^{\mathrm{RIE}} = \frac{\lambda_k}{|1 - q + q z_k s_k(z_k)|^2} \tag{15}$$

with:

$$s_k(z_k) = \frac{1}{N} \sum_{\substack{j=1 \\ j \neq k}}^{N} \frac{1}{z_k - \lambda_j} \tag{16}$$

- 2. We evaluate at the same time for any $k \in [\![1, N]\!]$:

$$\Gamma_k = \sigma^2 \frac{|1 - q + q z_k g_{mp}(z_k)|^2}{\lambda_k} \tag{17}$$

where $g_{mp}(z)$ is the Stieltjes transform of the (rescaled) Marčenko-Pastur density given by:

$$g_{mp}(z) = \frac{z + \sigma^2(q-1) - \sqrt{z - \lambda_N}\sqrt{z - \lambda_+}}{2qz\sigma^2} \tag{18}$$

with:

$$\lambda_+ = \lambda_N \left(\frac{1+\sqrt{q}}{1-\sqrt{q}}\right)^2, \qquad \sigma^2 = \frac{\lambda_N}{(1-\sqrt{q})^2} \tag{19}$$

where λ_N is the smallest empirical eigenvalue.

- 3. Our debiased optimal RIE for any $k \in [\![1, N]\!]$ reads:

$$\hat{\xi}_k = \begin{cases} \Gamma_k \xi_k^{\mathrm{RIE}} & \text{if } \Gamma_k > 1 \\ \xi_k^{\mathrm{RIE}} & \text{otherwise} \end{cases} \tag{20}$$

REFERENCES

Bouchaud J-P and M Potters, 2003
Theory of financial risk and derivative pricing: from statistical physics to risk management
Cambridge University Press

Bouchaud J-P and M Potters, 2011
Financial applications of random matrix theory: a short review
In *The Oxford Handbook of Random Matrix Theory*
Oxford University Press

Bun J, J-P Bouchaud and M Potters, 2016
Cleaning large correlation matrices: tools from random matrix theory
In preparation

Bun J and A Knowles, 2016
An optimal rotational invariant estimator for general covariance matrices
In preparation

Burda Z, J Jurkiewicz and B Wacław, 2005
Spectral moments of correlated Wishart matrices
Physical Review E 71(2), 026111

Couillet R, A Kammoun and F Pascal, 2016
Second order statistics of robust estimators of scatter: application to GLRT detection for elliptical signals
Journal of Multivariate Analysis 143, pages 249–274

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

Haff L, 1980
Empirical Bayes estimation of the multivariate normal covariance matrix
Annals of Statistics 8, pages 586–597

Ledoit O and S Péché, 2011
Eigenvectors of some large sample covariance matrix ensembles
Probability Theory and Related Fields 151(1-2), pages 233–264

Ledoit O and M Wolf, 2003
Improved estimation of the covariance matrix of stock returns with an application to portfolio selection
Journal of Empirical Finance 10(5), pages 603–621

Ledoit O and M Wolf, 2014
Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks
Available at SSRN 2383361

Markowitz HM, 1968
Portfolio selection: efficient diversification of investments, volume 16
Yale University Press

Marčenko VA and LA Pastur, 1967
Distribution of eigenvalues for some sets of random matrices
Matematicheskii Sbornik 114(4), pages 507–536

Monasson R and D Villamaina, 2015
Estimating the principal components of correlation matrices from all their empirical eigenvectors
Europhysics Letters 112(5), 50001

Pafka S and I Kondor, 2003
Noisy covariance matrices and portfolio optimization II
Physica A: Statistical Mechanics and its Applications 319, pages 487–494