Estimated correlation matrices and portfolio optimization

Szilárd Pafka [a,b,1], Imre Kondor [a,c,2]

[a] *Department of Physics of Complex Systems, Eötvös University,*
Pázmány P. sétány 1/a, H-1117 Budapest, Hungary

[b] *Risk Management Department, CIB Bank,*
Medve u. 4–14., H–1027 Budapest, Hungary

[c] *Collegium Budapest - Institute for Advanced Study,*
Szentháromság u. 2., H–1014 Budapest, Hungary

Abstract

Correlations of returns on various assets play a central role in financial theory and also in many practical applications. From a theoretical point of view, the main interest lies in the proper description of the structure and dynamics of correlations, whereas for the practitioner the emphasis is on the ability of the models to provide adequate inputs for the numerous portfolio and risk management procedures used in the financial industry. The theory of portfolios, initiated by Markowitz, has suffered from the "curse of dimensions" from the very outset. Over the past decades a large number of different techniques have been developed to tackle this problem and reduce the effective dimension of large bank portfolios, but the efficiency and reliability of these procedures are extremely hard to assess or compare. In this paper we propose a model (simulation)-based approach which can be used for the systematical testing of all these dimensional reduction techniques. To illustrate the usefulness of our framework, we develop several toy models that display some of the main characteristic features of empirical correlations and generate artificial time series from them. Then, we regard these time series as empirical data and reconstruct the corresponding correlation matrices which will inevitably contain a certain amount of noise, due to the finiteness of the time series. Next we apply several correlation matrix estimators and dimension reduction techniques introduced in the literature and/or applied in practice. As in our artificial world the only source of error is the finite length of the time series and, in addition, the "true" model, hence also the "true" correlation matrix, are precisely known, therefore in sharp contrast with empirical studies, we can precisely compare the performance of the various noise reduction techniques. One of our recurrent observations is that the recently introduced filtering technique based on random matrix theory performs consistently well in all the investigated cases. Based on this experience, we believe that our simulation-based approach can also be useful for the systematic investigation of several related problems of current interest in finance.

1 Introduction

Correlation matrices of financial returns play a crucial role in several branches of modern finance such as investment theory, capital allocation and risk management. For example, financial correlation matrices are the key input parameters to Markowitz's classical portfolio optimization problem [1], which aims at providing a recipe for the selection of a portfolio of assets so that risk (quantified by the standard deviation of the portfolio's return) is minimized for a given level of expected return. For any practical use of the theory it would therefore be necessary to have reliable estimates for the correlations of returns (of the assets making up the portfolio), which are usually obtained from historical return series data. However, if one estimates a $n \times n$ correlation matrix from n time series of length T each, with T bounded for evident practical reasons, one inevitably introduces estimation error, which for large n can become so overwhelming that the whole applicability of the theory becomes questionable.

This difficulty has been well known by economists for a long time (see e.g. [2] and the numerous references therein). Several aspects of the effect of noise (in the correlation matrices determined from empirical data) on the classical portfolio selection problem have been investigated e.g. in refs. [3]. One way to cope with the problem of noise is to impose some structure on the correlation matrix, which may certainly introduce some bias in the estimation, but by effectively reducing the dimensionality of the problem, could, in fact, be expected to improve the overall performance. The best-known such structure is that imposed by the single-index (or market) model, which has stimulated strong interest in the academic literature (see e.g. [2] for an overview and references) and has also become widely used in the financial industry (the coefficient "beta", relating the returns of an asset to the returns of the corresponding wide market index, has long been a widespread tool in the financial community). On economic or statistical grounds, several other correlation structures have been experimented with in the academic literature and financial industry, for example multi-index models, grouping by industry sectors, macroeconomic factor models, models based on principal component analysis

---

[1] syl@complex.elte.hu
[2] kondor@colbud.hu

etc. Several studies (see e.g. refs. [4]) attempt to compare the performance of these correlation estimation procedures as input providers for the portfolio selection problem, although all these studies have been restricted to the use of given specific empirical samples. More recently, additional procedures to impose some structure on correlations (e.g. Bayesian shrinkage estimators) or bounds directly on the portfolio weights (e.g. no short selling) have been explored, see e.g. refs. [5]. The general conclusion of all these studies is that reducing the dimensionality of the problem by imposing some structure on the correlation matrix may be of great help for the selection of portfolios with better risk–return characteristics.

The problem of estimation noise in financial correlation matrices has been put in a new light by the application of results from random matrix theory [6–8]. These studies have shown that empirical correlation matrices deduced from financial return series contain such a high amount of noise that, apart from a few large eigenvalues and the corresponding eigenvectors, their structure can essentially be regarded as random. In [7], e.g., it is reported that about 94% of the spectrum of correlation matrices determined from return series of the S&P 500 stocks can be fitted by that of a random matrix. Furthermore, two subsequent studies [9,10] have shown that the risk–return characteristics of optimized portfolios could be improved, if prior to optimization one filtered out the lower part of the eigenvalue spectrum of the correlation matrix in an attempt to remove (at least partially) the noise, a procedure similar to principal component analysis. Other approaches inspired by physics and aimed at extracting information from noisy correlation data have been introduced in [11,12]. It is important to note that all the above studies have used (given) empirical datasets, which in addition to the noise due to the finite length of the time series, also contain several other sources of error (caused by non-stationarity, market microstructure etc.).

The motivation of our previous study [13] came from this context. In order to get rid of these additional sources of error, we based our analysis on data artificially generated from some toy models. This procedure offers a major advantage in that the "true" parameters of the underlying stochastic process, hence also the correlation matrix is exactly known. The key observation of [13] is that the effect of noise, e.g. in the context of a simple portfolio optimization framework, strongly depends on the ratio T/n, where n is the size of the portfolio while T is the length of the available time series. Moreover, in the limit $n \to \infty$, $T \to \infty$ but $T/n = $ const. the suboptimality of the portfolio optimized using the "noisy" correlation matrix (with respect to the portfolio obtained using the "true" matrix) is $(1 - n/T)^{-1/2}$ exactly. Therefore, since the length of the time series T is limited in any practical application, any bound one would like to impose on the effect of noise translates, in fact, into a constraint on the portfolio size n.

3

The aim of this paper (besides extending the analysis of the previous study) is to introduce a *model (simulation)-based approach* that can be generally used for systematically testing and comparing the various noise (or dimension) reduction techniques that have been introduced in the literature and applied in practice. As an illustration of the usefulness of this approach, we introduce several toy models with the goal to progressively incorporate the relevant features of real-life financial correlations and, in the world of these models, we study the effect of noise (due here solely to the estimation error caused by the finiteness of the surrogate time series generated by the models) on a very simple form of the classical portfolio optimization problem. More precisely, we compare the performance of different correlation matrix "estimation" methods (the filtering procedure introduced in [9,10] among them) in providing inputs for the selection of portfolios with optimal risk characteristics. Our findings reconfirm the usefulness of techniques that effectively reduce the dimensionality of the correlation matrix for portfolio optimization. The approach we adopt here is, in fact, very common in physics, where one starts with some bare model and progressively adds finer and finer details in order to study the behavior of the "world" embodied by the model by comparing it to real-life (experimental) results. We believe that our model-based approach can be useful for the *systematic* study of several other problems in which financial correlation matrices play a crucial role.

2   Results and Discussion

We keep to the following simplified version of the classical portfolio optimization problem used already in [13]: the portfolio variance $\sum_{i,j=1}^{n} w_i \, \sigma_{ij} \, w_j$ is minimized under the budget constraint $\sum_{i=1}^{n} w_i = 1$, where w_i denotes the weight of asset i in the portfolio and σ_{ij} the covariance matrix of returns. That is, we look for the minimal risk portfolio, thereby eliminating the additional uncertainty arising from the return constraint [3] (or any other constraint that may be present). This simplified form provides the most convenient laboratory for testing the effect of noise in correlations. The weights of the minimal risk portfolio are:

$$w_i^* = \frac{\sum_{j=1}^{n} \sigma_{ij}^{-1}}{\sum_{j,k=1}^{n} \sigma_{jk}^{-1}}. \tag{1}$$

---

[3] The portfolio optimization problem without contraints on expected returns (i.e. finding the portfolio with minimal risk) is in fact meaningful even in its own, for example the problem of replicating (tracking) a benchmark with given instruments can be exactly mapped into it by considering the excess returns of the instruments over the benchmark (see e.g. Chan *et al* from Ref. [4]).

Starting from a given "true" covariance matrix $\sigma_{ij}^{(0)}$ ($n \times n$) we generate surrogate time series y_{it} (of finite length T), $y_{it} = \sum_{j=1}^{n} L_{ij} x_{jt}$, with $x_{jt} \sim$ i.i.d. N(0, 1) and L_{ij} the Cholesky decomposition of the matrix $\sigma_{ij}^{(0)}$. In this way we obtain "return series" y_{it} that have a distribution characterized by the "true" covariance matrix $\sigma_{ij}^{(0)}$. In order to mimic real-life situations (where the true covariance matrix is not known), we calculate different "estimates" $\sigma_{ij}^{(1)}$ for the covariance matrix based on several competing procedures and then use these estimates in our portfolio optimization. Finally, we compare the performance of these procedures using measures related to the risk (standard deviation) of the "optimal" portfolios constructed on the basis of the corresponding estimates. The main advantage of this simulation-based approach is that the "true" covariance matrix can be incorporated in the evaluation, which is certainly much cleaner than using, as in empirical studies, some proxy for it (which will inevitably introduce an additional source of noise).

In our previous study [13] we used a very simple structure ("model") for $\sigma_{ij}^{(0)}$ (namely the identity matrix) and we studied the effect of noise when the "estimated" matrix $\sigma_{ij}^{(1)}$ is the sample (or historical) covariance matrix. In this paper we introduce several other "models" (proposals for the structure of $\sigma_{ij}^{(0)}$) which are intended to incorporate progressively the most relevant characteristics of real-life financial correlations (the models are given in terms of the corresponding correlation matrix $\rho_{ij}^{(0)}$):

(1) "Single-index", "market" or "average correlation" model. The correlation matrix has 1's in the diagonal and a constant ρ_0 ($0 < \rho_0 < 1$) off-diagonal (all correlations are the same, hence the name of "average correlation" model). The eigenstructure of such a matrix consists of one large $(O(n))$ [4] eigenvalue with the corresponding eigenvector in the direction of $(1, 1, \ldots, 1)$ and a $(n-1)$-fold degenerate small eigenvalue. The eigenvector corresponding to the large eigenvalue can be thought of as describing a broad "index" composed of all the stocks (the "market"), hence the name of "single-index" or "market" model. This model is motivated by a similar feature (namely the presence of a single dominant eigenvalue) of stock market correlations found by numerous research studies (see e.g. [2] for references).

(2) "Market+sectors" model. A very simple structure intended to incorporate this much debated [5] feature of real-life financial correlations can be based on a correlation matrix composed of $n_1 \times n_1$ blocks (with 1 in the diagonal and ρ_1 off-diagonal) and ρ_0 outside the blocks ($0 < \rho_0 < \rho_1 < 1$ and $\frac{n}{n_1}$ integer). In this model there is still a strong influence of the "market",

---

[4] $\lambda_1 = 1 + (n-1)\rho_0$, which for the usual values of the parameters is large compared to $\lambda_2 = \lambda_3 = \ldots = \lambda_n = 1 - \rho_0$.
[5] See e.g. refs. [14].

but stocks from the same block ("industrial sector") display additional common correlations. On the other hand, the eigenspectrum of such a matrix[6] is closer to the eigenspectrum of real-life financial correlation matrices as described e.g. in [10]. This correlation structure also fits better the findings of [11,12], which, using a hierarchial tree approach, found also that stocks tend to be coupled according to their belonging to the same industrial sector.

(3) "Semi-empirical" (bootstrapped) model. Starting from a large set of empirical financial data[7] for each portfolio size n, we select randomly (bootstrap) n time series from the set of empirical return data and an $n \times n$ covariance matrix is calculated using the full length of the available series. This matrix is then used as $\sigma_{ij}^{(0)}$ in the simulations (to generate the surrogate data). In order to examine the sensitivity of our results with respect to the choice of the n time series, we repeat the simulations several times (with different bootstrapped empirical series) and we compare the results. The correlation structure of this model is hoped to be the closest to real-world financial correlations, although the disadvantage of its use is that, similar to empirical studies, it is based on a given set of empirical data which may be representative in certain situations, and less so in others.

In the framework of each of the models introduced above, we investigate the performance of three alternative choices for the "estimated" covariance matrix $\sigma_{ij}^{(1)}$:

(1) Sample (historical) covariance matrix.
(2) "Single-index" covariance matrix, i.e. the matrix obtained from the sample covariance matrix by a simplified filtering procedure similar to the one described below, but considering only the largest eigenvalue (and the corresponding eigenvector), which is believed to correspond to a broad market index covering all stocks, see e.g. [10].
(3) Filtered covariance matrix using the procedure based on random matrix theory [9,10]. For this, one starts with the sample correlation matrix and keeps only the eigenvalues and the corresponding eigenvectors reflecting deviations from random matrix theory predictions (those outside the random matrix noise-band) and then constructs a "cleaned" correlation matrix such that the trace of the matrix is preserved. The intuition behind

---

[6] The eigenstructure is formed of a large eigenvalue $\lambda_1 = 1 + (n_1 - 1)\rho_1 + (n - n_1)\rho_0$, a $\frac{n}{n_1} - 1$-fold degenerated subspace corresponding to medium-size eigenvalues $\lambda_2 = \lambda_3 = \ldots = \lambda_{\frac{n}{n_1}} = 1 + (n_1 - 1)\rho_1 - n_1\rho_0$ and a $n - \frac{n}{n_1}$-fold degenerated subspace with eigenvalues $\lambda_{\frac{n}{n_1}+1} = \lambda_{\frac{n}{n_1}+2} = \ldots = \lambda_n = 1 - \rho_1$.

[7] The same dataset as in [13] has been used (daily return series on 406 major US stocks during the period 1991–1996, 1308 observations for each stock). We thank again J.-P. Bouchaud and L. Laloux for making their data [7,9] available to us.

this procedure is that deviations from random matrix theory predictions should correspond to "information" and describe genuine correlations in the system, while the eigenstates corresponding to random matrix theory predictions should be manifestations of purely random "noise". The filtered covariance matrix is then obtained from the filtered correlation matrix and sample standard deviations. This procedure is very much reminiscent of principal component analysis, although classical multivariate analysis generally gives no hints about how many components (factors) are to be included in the matrix constructed using the principal components (see e.g. [15]). The filtering procedure based on random matrix theory can therefore be thought of as a theoretically sound indication for the number of principal components to be included in the analysis.

To study the effect of noise on the portfolio optimization problem we use metrics based on the following quantities:

(1) $\sum_{i,j=1}^{n} w_i^{(0)*} \sigma_{ij}^{(0)} w_j^{(0)*}$, the "true" risk of the optimal portfolio without noise, where $w_i^{(0)*}$ denotes the solution to the optimization problem with $\sigma_{ij}^{(0)}$;

(2) $\sum_{i,j=1}^{n} w_i^{(1)*} \sigma_{ij}^{(0)} w_j^{(1)*}$, the "true" risk of the optimal portfolio determined in the case of noise, where $w_i^{(1)*}$ denotes the solution to the optimization problem with $\sigma_{ij}^{(1)}$;

(3) $\sum_{i,j=1}^{n} w_i^{(1)*} \sigma_{ij}^{(1)} w_j^{(1)*}$, the "predicted" risk (cf. [9,10,13]), that is the risk that *can be* observed when the optimization is based on the "empirical" series;

(4) $\sum_{i,j=1}^{n} w_i^{(1)*} \sigma_{ij}^{(2)} w_j^{(1)*}$, the "realized" risk (cf. [9,10,13]), that is the risk that *would be* observed if the portfolio were held one more "period", where $\sigma_{ij}^{(2)}$ is the covariance matrix calculated from the returns in this second period.

To facilitate the comparison, we calculate the ratios of the square roots of the three latter quantities to the first one, and denote these by q_0, q_1 and q_2, respectively. That is q_0, q_1 and q_2 represent the "true", the "predicted" resp. the "realized" risk, expressed in units of the "true" risk in the absence of noise. In other words, q_0 directly describes the ability of a given estimation procedure to provide the correct input for portfolio optimization, q_1 describes the bias one makes if one uses the estimated matrix for the calculation of the risk of the optimal portfolio, while q_2 is the risk measured if one waits in time and uses the information from the new series for risk measurement (see also [13]).

We start with presenting the simulation results when the series have been generated using the "*market*" *model* (for $\sigma_{ij}^{(0)}$). Since the main feature of the correlation structure (one outstanding large eigenvalue) is, at least for the

parameter values used in our simulations, preserved also in the correlation matrix obtained from the generated series ($\sigma_{ij}^{(1)}$), the results for the filtering based on the largest eigenvalue and on random matrix theory are in fact the same. Therefore, we proceed with comparing the performance of the historical and filtered estimation procedures for different values of the model parameters n, T and ρ_0 using the evaluation metrics q_0, q_1, q_2 and q_2/q_1. A summary of our simulation results is presented in Table 1.

Table 1
Optimal portfolio risk and performance indicators for the historical (h) and market (m) correlation matrix estimators for different values of the parameters of the model ($\sigma_{ij}^{(0)}$).

ρ_0	n	T	T/n	$q_0^{(h)}$	$q_0^{(m)}$	$q_1^{(h)}$	$q_1^{(m)}$	$q_2^{(h)}$	$q_2^{(m)}$	$q_2/q_1^{(h)}$	$q_2/q_1^{(m)}$
0.2	200	300	1.5	1.77	1.11	0.56	0.78	1.77	1.13	3.16	1.46
0.2	1000	1500	1.5	1.73	1.12	0.59	0.78	1.71	1.11	2.96	1.42
0.6	1000	1500	1.5	1.75	1.11	0.58	0.77	1.75	1.12	3.01	1.45
0.2	1000	2000	2	1.42	1.11	0.71	0.82	1.43	1.11	2.00	1.35
0.2	1000	5000	5	1.11	1.07	0.89	0.91	1.12	1.07	1.26	1.18
0.2	1000	500	0.5	-	1.12	-	0.57	-	1.12	-	1.92

It turns out that, for sufficiently large n and T, the value of the q's depends strongly only on T/n (and, interestingly, does not seem to depend on ρ_0). This can be seen also from the results presented in Table 1 (the variation in the first 3 rows is in fact within the usual standard deviation bounds). This is not very surprising in view of the results for the historical matrix, which has been studied in our previous paper [13]. The strong dependence on T/n seems to be valid, however, also when the filtered matrix is used. One important difference to note is, however, the significant improvement in the risk characteristics of the optimal portfolio when the filtering procedure is used for estimation: e.g. for $T/n = 2$ instead of obtaining a portfolio with risk more than 40% larger than the truly optimal one (see q_0), using the filtering procedure one can get portfolios with only 10% larger risk. Furthermore, as it can also be seen from the table, using the filtered matrix one can obtain portfolios close to the optimal one even for $T \leq n$ when the sample (historical) matrix is singular and completely useless for the optimization. This improvement in performance is not difficult to understand, since with the filtering procedure one implicitly incorporates into the "estimation" the additional information about the structure of the correlation matrix. Note also that q_2 is very close to q_0 for all parameter values, therefore the risk measured in the second "period" seems to be a good proxy for the "true" risk of the optimal portfolio.

We now present the results when the series are generated with the "market+sectors" model, for different values of the parameters n, T, n_1, ρ_0 and

Table 2
Optimal portfolio risk and performance indicators for the historical (h), market (m) and random matrix theory (r) correlation matrix estimators for different values of the parameters of the model ($\sigma_{ij}^{(0)}$).

| ρ_0	ρ_1	n_1	n	T	$q_0^{(h)}$	$q_0^{(m)}$	$q_0^{(r)}$	$q_1^{(h)}$	$q_1^{(m)}$	$q_1^{(r)}$	$q_2/q_1^{(h)}$	$q_2/q_1^{(m)}$	$q_2/q_1^{(r)}$
0.2	0.4	25	200	300	1.71	1.27	1.13	0.58	0.77	0.76	2.93	1.65	1.47
0.2	0.4	25	1000	1500	1.75	1.28	1.13	0.58	0.77	0.76	3.07	1.63	1.46
0.2	0.6	25	1000	1500	1.74	1.64	1.13	0.59	0.78	0.76	2.94	2.09	1.47
0.4	0.6	25	1000	1500	1.73	1.36	1.13	0.58	0.77	0.76	2.96	1.77	1.49
0.2	0.4	50	1000	1500	1.71	1.42	1.12	0.58	0.77	0.77	2.96	1.84	1.46
0.2	0.4	25	1000	2000	1.42	1.24	1.12	0.70	0.82	0.81	1.99	1.50	1.37
0.2	0.4	25	1000	5000	1.11	1.16	1.07	0.89	0.91	0.90	1.24	1.27	1.17
0.2	0.4	25	1000	500	-	1.24	1.19	-	0.58	0.55	-	2.14	2.17
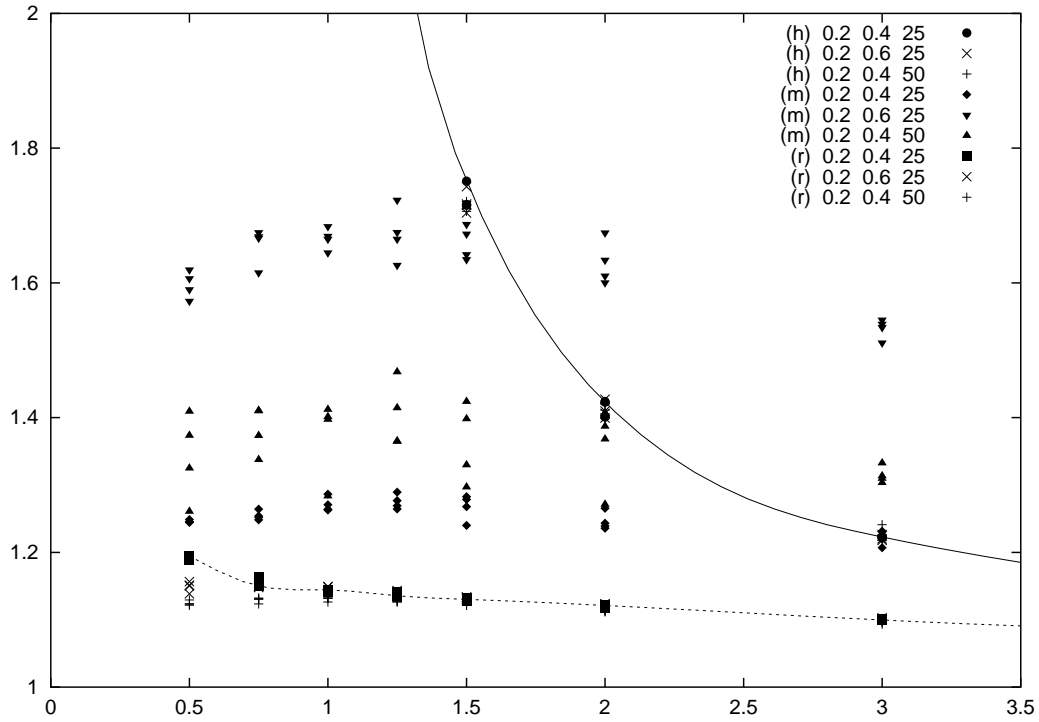
Fig. 1. q_0 as a function of T/n for different values of the parameters ρ_0, ρ_1 and n_1 and different values of n and T. In the case of the historical and random matrix theory estimator (h and r, resp.) the points line up approximately on a line (solid and dotted, resp.). For the market estimator (m), however, the dependence on virtually all the parameters is clear from the figure (e.g. the increase in either ρ_1 or n_1 leads to the increase of q_0).

9

ρ_1. Our results are summarized in Table 2. The values for q_2's are again very close to q_0 and therefore have been left out from the table. We have found that the value of the q's in the case of the historical and random matrix theory-based estimators, again, depends strongly on T/n and not on the value of the other parameters, while this is not true for the estimator based on the largest eigenvalue only. This is illustrated in Fig. 1, where q_0 in the case of the three estimators is represented as a function of T/n for different value of the parameters n, T, n_1, ρ_0 and ρ_1. The dependence of q_0 for the "single-index" estimator on the parameters ρ_0, ρ_1 and n_1 can be easily understood, since either the increase of ρ_1 or n_1, or the decrease of ρ_0 can be thought of as the increase in the relative strength of "inter-sector" correlations (relative to the overall correlation corresponding to the "market") and therefore an estimator taking into account only the "market" component of correlations (and ignoring the "sector" component) is of course expected to perform worse in this case. Another important point to note is that, in most cases, the random-matrix-theory-based filtering outperforms the single-index estimator which in turn outperforms the historical estimator. Moreover, the first two estimators can be used even when the latter one provides a singular matrix totally inappropriate for input to the portfolio optimization (for $T \leq n$).

Finally, we analyze the performance of the three correlation matrix estimators in the case of the *"semi-empirical" model* for $\sigma_{ij}^{(0)}$ (the matrix is bootstrapped from the empirical matrix of a given large set of financial series). More precisely, for each value of the parameter n, we select at random n series from the available dataset and we calculate the historical matrix which is then used as $\sigma_{ij}^{(0)}$ in our simulations [8]. Our results are summarized in Table 3 (the values for q_2's have been again left out of the table.) In this case, the q's for the two filtering matrix estimations do not depend so strongly on T/n, some dependence on n (and T) can also be observed (see Fig. 2). It can be said again that, in general, the filtering procedures outperform significantly the historical matrix estimation, with the filtering based on the random matrix theory approach performing best.

In conclusion, our simulation study provides a more general argument for the usefulness of techniques for "massaging" empirical correlation matrices before using them as inputs for portfolio optimization as suggested e.g. by [4,5,9,10]. In particular, it confirms the fruitfulness of the random matrix theory-based filtering procedure for portfolio selection applications.

---

[8] Since most of the values for the length T of the time series used in our simulations is small compared to the lengths of the original dataset from which $\sigma_{ij}^{(0)}$ is computed, the noise due to the "measurement error" of $\sigma_{ij}^{(0)}$ can be hoped to be small compared to the noise (deliberately) introduced by the finiteness of T.

Table 3
Optimal portfolio risk and performance indicators for the historical (h), market (m) and random matrix theory (r) correlation matrix estimators for different values of the parameters of the model ($\sigma_{ij}^{(0)}$).

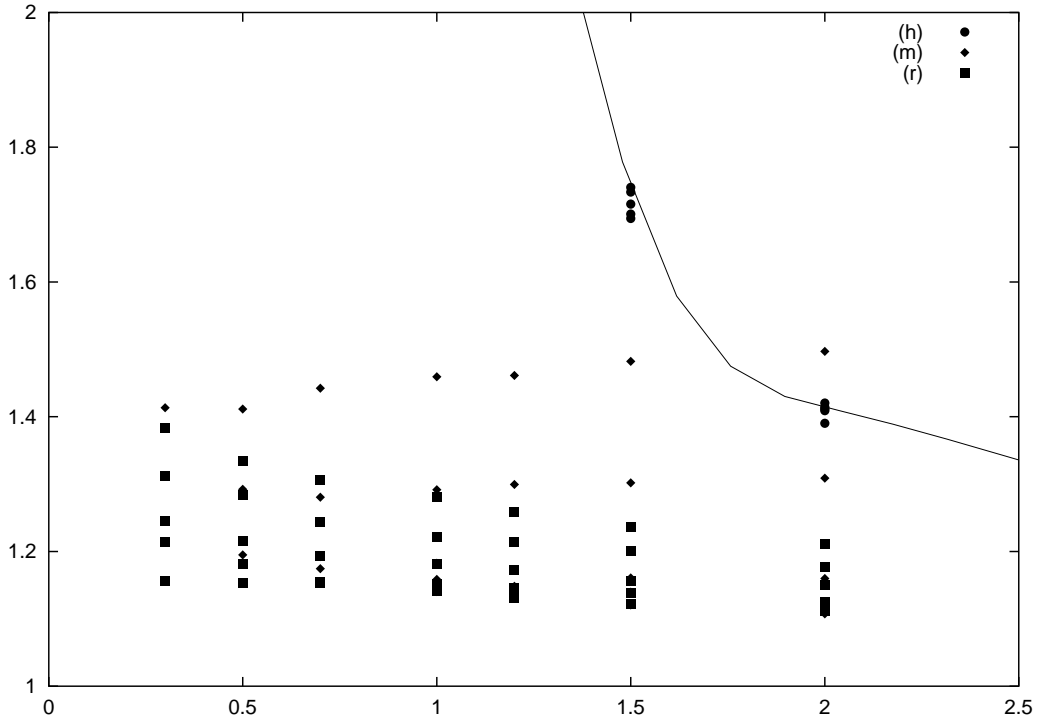| n	T	T/n	$q_0^{(h)}$	$q_0^{(m)}$	$q_0^{(r)}$	$q_1^{(h)}$	$q_1^{(m)}$	$q_1^{(r)}$	$q_2/q_1^{(h)}$	$q_2/q_1^{(m)}$	$q_2/q_1^{(r)}$
200	300	1.5	1.70	1.30	1.20	0.58	0.78	0.83	3.03	1.67	1.44
300	450	1.5	1.74	1.48	1.24	0.58	0.76	0.84	2.99	1.94	1.45
300	600	2	1.41	1.50	1.21	0.71	0.77	0.90	2.02	1.95	1.35
300	1500	2	1.12	1.53	1.15	0.89	0.80	0.96	1.26	1.92	1.21
300	150	0.5	-	1.41	1.33	-	0.76	0.73	-	2.02	1.85

Fig. 2. q_0 as a function of T/n for different values of n and T. In the case of the historical estimator (h) the points line approximately on a line. For the market and random matrix theory estimator (m and r, resp.), however, the dependence on n and T is clear from the figure.

There are several possibilities to extend the analysis of this paper. First, "models" that incorporate more subtle features of financial returns (such as the dynamics of volatility and correlations or the non-Gaussian nature of returns) could be used for generating the surrogate time series data. Second, our model-based approach can be used for evaluating the performance of several other correlation matrix "estimators" introduced in the literature and/or used in

practice (e.g. the Bayesian shrinkage estimator of Ledoit *et al* of Ref. [5], or the estimator of RiskMetrics [16], where the covariance matrix is obtained by exponentially weighting the square returns so that returns further in the past get lower weights [9]). Third, our simulation-based approach can also be applied when a more sophisticated risk measure (e.g. expected shortfall/Conditional Value-at-Risk) is used. For example, when expected shortfall is chosen as risk measure, the target function becomes piecewise linear in the weights of the assets and the optimization problem can be reduced to linear programming, see e.g. [18]. However, the parameters of the target function have to be estimated from empirical financial data (return series) again, therefore they will be noisy and one can expect that in the case of a large number of assets the effect of noise may be significant. The study of all these problems listed above remains the subject of future work.

The implications of successful noise filtering of correlation matrices used for portfolio optimization are enormous. Correlation matrices are not only at the heart of modern finance and investment theory, but they also appear in most practical risk management and asset allocation procedures used in the financial industry. In particular, most implementations of practical risk–return portfolio optimization or benchmark tracking involve either correlation matrices or "scenarios" usually generated using correlation matrices, see e.g. [19]. A short overview on the techniques used by practitioners for reducing noise and estimation error in correlation matrices can be found in [20]. On the other hand, from a purely academic point of view, understanding the structure and dynamics of correlations in financial markets is of central interest in finance and related fields, therefore any study that makes it possible to reveal finer and finer details of this structure could be of significant importance.

3 Conclusion

In this paper we described a model (simulation)-based approach which can be used for a systematic investigation of the performance of various noise reduction procedures applied in portfolio selection and risk management. To demonstrate the usefulness of this approach we developed several toy models for the structure of financial correlations and, by considering only the noise arising from the finite length of the model-generated time series, we analyzed the performance of several correlation matrix estimation procedures in a simple portfolio optimization context. Our results agree well with the findings of previous empirical studies. The effect of noise in correlation matrices deter-

---

[9] This technique is widely utilized in practical risk management for assesing the market risk of portfolios, for example using the parametric Value-at-Risk (VaR) method.

mined from financial series can indeed be large. However, most practitioners use techniques that, by generally reducing the effective dimensionality of the problem, can very efficiently suppress the effect of noise. We found that the filtering based on random matrix theory is particularly powerful in this respect. The success of dimensional reduction procedures goes a long way to explain how correlation matrices that contain a huge amount of noise can nevertheless remain useful in practice.

Acknowledgements

References

[1] H. Markowitz, J. Finance **7**, 91 (1952); H. Markowitz, *Portfolio Selection: Efficient Diversification of Investments* (J. Wiley and Sons, New York, 1959).

[2] E. J. Elton and M. J. Gruber, *Modern Portfolio Theory and Investment Analysis* (J. Wiley and Sons, New York, 1995).

[3] G. Frankfurter, H. Phillips and J. Seagle, J. Fin. Quant. Anal. **6**, 1251 (1971); J. Dickinson, J. Fin. Quant. Anal. **9**, 447 (1974); J. Jobson and B. Korkie, J. Am. Stat. Assoc. **75**, 544 (1980); R. Michaud, Fin. Anal. Journal **45**, 31 (1989); V. Chopra and W. T. Ziemba, J. Portf. Managem. **19**, 6 (1993).

[4] E. J. Elton and M. J. Gruber, J. Finance **28**, 1203 (1973); E. J. Elton, M. J. Gruber and T. Urich, J. Finance **33**, 1375 (1978); C. Eun and B. Resnick, J. Finance **39**, 1311 (1984); L. Chan, J. Karceski and J. Lakonishok, Rev. Fin. Stud. **12**, 937 (1999); N. Amenc and L. Martellini, J. Altern. Inv. **5**, 7 (2002); C. Bengtsson and L. Holst, Lund University working paper (2002).

[5] P. Jorion, J. Fin. Quant. Anal. **21**, 544 (1986); P. Frost and J. Savarino, J. Fin. Quant. Anal. **21**, 293 (1986); P. Frost and J. Savarino, J. Portf. Managem. **14**, 29 (1988); V. Chopra, C. Hensel and A. Turner, Managem. Sci. **39**, 845 (1993); O. Ledoit and M. Wolf, J. Empir. Fin., forthcoming (2003); R. Jagannathan and T. Ma, NBER working paper No. 8922 (2002).

[6] G. Galluccio, J.-P. Bouchaud and M. Potters, Physica A **259**, 449 (1998).

[7] L. Laloux, P. Cizeau, J.-P. Bouchaud and M. Potters, Phys. Rev. Lett. **83**, 1467 (1999); L. Laloux, P. Cizeau, J.-P. Bouchaud and M. Potters, Risk **12**, No. 3, 69 (1999).

[8] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral and H. E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).

[9] L. Laloux, P. Cizeau, J.-P. Bouchaud and M. Potters, Int. J. Theor. Appl. Finance **3**, 391 (2000).

[10] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr and H. E. Stanley, e-print cond-mat/0108023; B. Rosenow, V. Plerou, P. Gopikrishnan and H. E. Stanley, e-print cond-mat/0111537.

[11] G. Bonanno, F. Lillo and R. N. Mantegna, Physica A **299**, 16 (2001), e-print cond-mat/0104369; G. Bonanno, G. Caldarelli, F. Lillo and R. N. Mantegna, e-print cond-mat/0211546.

[12] L. Kullmann, J. Kertesz and R. N. Mantegna, Physica A **287**, 412 (2000); J. P. Onnela, A. Chakraborti, K. Kaski and J. Kertesz, e-print cond-mat/0208131.

[13] S. Pafka and I. Kondor, Physica A **319**, 487 (2003), e-print cond-mat/0205119.

[14] B. King, J. Business **39**, 139 (1966); S. Meyers, J. Finance **28**, 695 (1973); J. Farrell, J. Business **47**, 186 (1974); M. Livingston, J. Finance **32**, 861 (1977); Y. Malevergne and D. Sornette, e-print cond-mat/0210115.

[15] W. R. Krzanowski, *Principles of Multivariate Analysis*, (Clarendon Press, Oxford, 1998).

[16] J. P. Morgan and Reuters, RiskMetrics – Technical Document (New York, 1996).

[17] R. Engle, V. Ng and M. Rotschild, J. Econometrics **45**, 213 (1990); C. Alexander, "Orthogonal GARCH" in *Mastering Risk*, Vol. 2, ed. by C. Alexander (FT–Prentice Hall, 2001).

[18] R. T. Rockafellar and S. Uryasev, Univ. of Florida working paper (1999); R. T. Rockafellar and S. Uryasev, Risk, **2**, No. 3, 21 (2000); C. Acerbi and P. Simonetti, working paper Abaxbank (2002), e-print www.gloriamundi.org.

[19] Barra, portfolio optimization documentation, www.barra.com; Algorithmics, scenario and risk–reward optimization documentation, www.algorithmics.com; R. Litterman and K. Winkelmann, Goldman Sachs Risk Management Series paper (1998); APT, portfolio optimization documentation, www.apt.com.

[20] D. diBartolomeo, "Risk of Equity Securities and Portfolios", Northfield Information Services Inc. paper, www.northinfo.com.