

The instability of downside risk measures

I Varga-Haszonits^{1,2} and I Kondor^{1,3,4}

¹ Department of Physics of Complex Systems, Eötvös University, Pázmány Péter sétány 1/A, H-1117 Budapest, Hungary

² Analytics Department of Fixed Income Division, Morgan Stanley Hungary Analytics, Deák Ferenc u. 15, H-1052 Budapest, Hungary

³ Collegium Budapest – Institute for Advanced Study, Szentháromság u. 2, H-1014 Budapest, Hungary

⁴ Parmenides Center for the Study of Thinking, Kardinal Faulhaber Strasse 14a, Munich, D-80333, Germany

E-mail: Istvan.Varga-Haszonits@morganstanley.com, kondor@colbud.hu

Abstract. We study the feasibility and noise sensitivity of portfolio optimization under some downside risk measures (Value-at-Risk, Expected Shortfall, and semivariance) when they are estimated by fitting a parametric distribution on a finite sample of asset returns. We find that the existence of the optimum is a probabilistic issue, depending on the particular random sample, in all three cases. At a critical combination of the parameters of these problems we find an algorithmic phase transition, separating the phase where the optimization is feasible from the one where it is not. This transition is similar to the one discovered earlier for Expected Shortfall based on historical time series. We employ the replica method to compute the phase diagram, as well as to obtain the critical exponent of the estimation error that diverges at the critical point. The analytical results are corroborated by Monte Carlo simulations.

Keywords: Replica Method, Critical Phenomena, Portfolio Optimization, Expected Shortfall, Value-at-Risk, Semivariance

PACS numbers: 89.65.Gh

Submitted to: *JSTAT*

1. Introduction

Portfolio optimization is one of the fundamental problems of financial theory. The first treatment of the topic appeared in the famous work by Markowitz [1], who measured risk by the standard deviation of asset price fluctuations. In this context, portfolio optimization consists in minimizing the variance of the portfolio return given the expected return and the budget constraint. Although this defines a straightforward mathematical problem, the statistical properties of the solution turn out to be non-trivial when the covariances of the asset returns are estimated from a finite sample.

An extensive investigation of the noise sensitivity of the Markowitz portfolio optimization problem [2, 3, 4] revealed that for normally distributed asset returns the expected value of the ratio q_0 of the risk of the estimated optimum and that of the true optimum is proportional to $(1 - N/T)^{-1/2}$, where N is the number of assets in the portfolio and T is the sample size (number of observation periods). In other words, the estimation error diverges as $T \rightarrow N$, and, in order to reduce the estimation error to a reasonable level, one needs a fairly large sample. Moreover, the estimated optimal portfolio weights exhibit dramatic fluctuations from one sample to another, and these fluctuations decay very slowly with increasing sample size. Covariance matrix filtering techniques based on Bayesian Shrinkage [5, 6, 7] and Random Matrix Theory [8, 9, 10, 11, 12] were shown to effectively reduce q_0 [3], however, these techniques do not generally suppress the large fluctuations of the estimated portfolio weights.

In addition to the noise sensitivity of the classical standard deviation, Kondor et al [13] also examined the sensitivity of portfolio optimization under a few alternative risk measures, such as Mean Absolute Deviation [14], Maximal Loss [15] and Expected Shortfall [16, 17]. All of these were found to be even more susceptible to sample fluctuations than standard deviation, and in addition, Expected Shortfall (and Maximal Loss as its special case) displayed an additional instability in that the very existence of the optimum turned out to depend on the sample and the probability of the existence of an optimum was found to be less than one for any finite sample size. In other words, even if Expected Shortfall has a well defined minimum for a given asset return distribution, it may not have an optimum on a finite sample generated by that distribution.

Expected Shortfall is perhaps the simplest and intuitively most appealing example of the celebrated Coherent Risk Measures [18, 19], which were introduced in response to the widespread use of ad-hoc risk measures (including Value-at-Risk) with poor theoretical foundation and well-known shortcomings. However, the instability discovered on the example of Expected Shortfall raised the suspicion that Coherent Risk Measures, all their axiomatic beauty notwithstanding, may be highly susceptible to sampling error in general. Indeed, this conjecture has been proved to be true by showing that no Coherent Measure of Risk has a minimum, if there exists a portfolio that produces positive returns for all observations on the given sample [20].

The studies mentioned above were based on non-parametric estimators of the risk measures in consideration, without any a priori assumption about the sample generating process. However, estimators based on historical time series are notoriously unstable, so it is legitimate to ask whether parametric estimation could suppress the instability. Moreover, Value-at-Risk is often measured in practice by parametric estimation using some assumption about the probability distribution of the asset returns [21]. Since in practice VaR is the most important measure in use today,

and, furthermore, its parametric estimation is analogous to that of ES, it is a natural idea to study the stability of portfolio optimization under VaR and ES by fitting a multivariate Gaussian distribution on the sample of asset returns. The main objective of this paper is to decide whether the instability of historical ES (and VaR) estimation can be circumvented by parametric estimation. For the sake of simplicity we are also going to assume that the data generating process is itself Gaussian. It will turn out that, although parametric fitting reduces the instability, it does not eliminate it.

It should be noted that this paper, as well as the earlier studies mentioned above, investigate the noise sensitivity of the global risk minimization, without imposing any constraint on the expected return. This is a special case of the practically more relevant risk-reward optimization problem. It is clear, however, that adding a linear constraint to the global minimum risk problem does not change essentially the noise sensitivity characteristics. Focusing on the simpler problem makes it easier to understand and identify the effects and consequences of sampling error, while at the same time leaves open the possibility of revisiting the more general problem later.

The rest of the paper is organized as follows. Section 2 is a brief overview of earlier results on the instability of the minimization of Expected Shortfall with non-parametric estimation. In Section 3 we solve the ES/VaR minimization problem assuming that asset returns follow a multivariate normal distribution with explicitly known means, variances and covariances, and we derive the condition for the solution to exist. In Section 4 we investigate the feasibility and noise sensitivity of ES/VaR minimization when the parameters of the asset return distribution are estimated from finite samples. This section, which constitutes the backbone of our paper, is divided into several subsections: in 4.1 we introduce some notations and terminology, in 4.2 we use the replica method to characterize the critical behavior of the finite sample instability of the optimization problem, in 4.3 we back up our findings with simulation, and finally in 4.4 we apply our results to the special case of semivariance minimization. The paper ends on a brief summary.

2. The noise-sensitivity of Expected Shortfall minimization with non-parametric estimation

To put our discussion in context, we provide a brief overview of the results for the minimization of Expected Shortfall using a non-parametric estimator. Expected Shortfall is the mean value of losses exceeding a high threshold (referred to as the confidence level) specified in probability rather than in money. For instance, at confidence level α the Expected Shortfall (ES_α) of an investment is the average of losses that occur in the $(1 - \alpha)100$ percent of the worst cases.

Historical ES based on a finite sample consisting of T observations can be estimated by sorting these observations into ascending order and computing the average of the $T(1 - \alpha)$ smallest values. Special care must be taken, however, when $T(1 - \alpha)$ is not an integer number: in such a case one of the observations has to be 'split'. (For the precise definition of ES_α see for instance [17].) It was shown in [22] that within this scheme portfolio optimization is equivalent to a convex linear programming problem. This is to be contrasted with the case of VaR, which, as a quantile, has no reason to be convex, and, indeed, is often found to be non-convex when estimated from historical time series. (This is why the problem of the noise sensitivity of VaR was ignored in [13]: in a sense *historical* VaR is always unstable.) The highly desirable property of convexity has made ES very popular with academics,

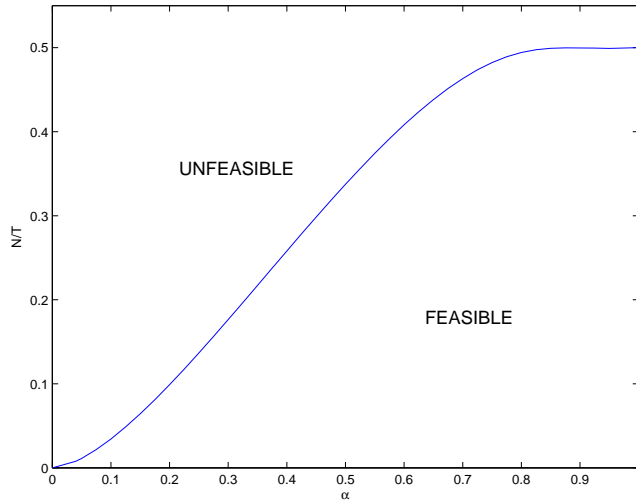


Figure 1. The boundary between the feasible and unfeasible phases of the Expected Shortfall minimization problem on the N/T vs α plane, in the $N \rightarrow \infty$ and $N/T = \text{const}$ limit.

though ES is still very far from replacing VaR in practice or regulation.

As mentioned in the introduction, the noise sensitivity of ES optimization was examined in [13]. That study used a simulation based approach and assumed, for simplicity, iid normal asset returns. The (linear programming based) portfolio optimization algorithm was performed on a large number of such samples and the existence and distribution of the solution was investigated. The main findings of this study are the following:

- ES as a risk measure is much more sensitive to sample to sample fluctuations than the variance.
- On some samples ES does not even have a minimum but diverges to minus infinity.
- The probability of the existence of the optimum depends on the confidence level α , as well as on the ratio between the number of assets N and the number of observations T .
- In the limit where $N \rightarrow \infty$ and N/T is held constant the probability of the existence of the optimum tends either to 1 or to 0. On the N/T vs α plane the zero probability (unfeasible) and unit probability (feasible) regions are separated by a well defined curve (the phase diagram), which was first determined by simulations [13], then computed analytically by the replica method [23] (see Figure 1).

In practical applications the confidence level is typically $\alpha > 0.9$, and as shown by Figure 1 in that region the critical N/T ratio is very close to $1/2$. This means that in the practically relevant cases one must have at least twice as many observations as the number of assets in order to ensure even the mere existence of an optimal portfolio. (And, of course, a much larger sample is needed to make the estimation error reasonably low.) Moreover, the critical value of the ratio N/T decreases for

decreasing confidence level, which implies that ES optimization becomes more and more unstable, requiring larger and larger samples to give a meaningful result.

3. The minimization of parametric ES and VaR for Gaussian asset returns

The sensitivity of Expected Shortfall to sample fluctuations casts a shadow of doubt on its practical applicability in portfolio selection. However, one may wonder whether this instability is not due to the use of raw data in historical estimation and whether a parametric method might be more robust against sample to sample fluctuations.‡ In order to decide the question, we are going to look into the noise sensitivity of portfolio selection in the simplest setting, that is when the underlying process is iid normal and when the risk is estimated by fitting a normal distribution to the sample. This is a standard procedure for VaR estimation [21], but the ES and VaR estimators are so closely related that we can examine them together.

When the return of an asset X is normally distributed with mean μ and standard deviation σ , then both its VaR and ES can be written in the form

$$\mathcal{R}(X) = \phi(\alpha)\sigma - \mu. \quad (1)$$

The particular form of the function $\phi(\alpha)$ depends on whether we are computing VaR or ES:

$$\phi(\alpha) = \begin{cases} \Phi^{-1}(\alpha) & \text{for VaR,} \\ -\frac{1}{1-\alpha} \int_0^{1-\alpha} \Phi^{-1}(p) dp = \frac{e^{-\frac{1}{2}[\Phi^{-1}(\alpha)]^2}}{(1-\alpha)\sqrt{2\pi}} & \text{for ES.} \end{cases} \quad (2)$$

where $\Phi^{-1}(x)$ denotes the inverse of the standard normal cumulative distribution function (or error function):

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy. \quad (3)$$

We assume that $\phi(\alpha)$ is nonnegative and invertible in its domain§, and we will often omit its dependence on α in the notation. All the relevant quantities depend on α only through $\phi(\alpha)$.

Let us now assume that we have N assets in the portfolio and their returns x_i follow a multivariate normal distribution with means μ_i and variances/covariances σ_{ij} (where $i, j = 1, 2, \dots, N$). A portfolio is simply a vector with components w_i representing the amount invested in asset i . Then the expected value and the variance of the portfolio return will be $\sum_{i=1}^N w_i \mu_i$ and $\sum_{i,j=1}^N \sigma_{ij} w_i w_j$, respectively. According to (1), ES and VaR can then be written as:

$$\mathcal{R}_\phi(\{w_i\}) = \phi \sqrt{\sum_{i=1}^N \sum_{j=1}^N \sigma_{ij} w_i w_j} - \sum_{i=1}^N \mu_i w_i. \quad (4)$$

The optimal portfolio can be found by minimizing $\mathcal{R}(\{w_i\})$ subject to the budget constraint

$$\sum_{i=1}^N w_i = 1. \quad (5)$$

‡ We are obliged to M. Gordy for a stimulating discussion on this point.

§ This means that for VaR we only allow confidence levels between 0.5 and 1. This is, however, not a real restriction, since VaR does not make sense as a risk measure for $\alpha < 0.5$.

It is easy to see that this optimization problem is equivalent to minimizing the following Lagrangean:

$$\mathcal{L}(\{w_i\}, z, \lambda, \eta) = \phi\sqrt{z} - \sum_i w_i\mu_i + \lambda \left(\sum_i w_i - 1 \right) + \eta \left(\sum_{ij} w_i\sigma_{ij}w_j - z \right). \quad (6)$$

where λ is used to enforce the budget constraint while z and η have been introduced to make the objective function quadratic in the portfolio weights. The minimization of \mathcal{L} is a routine task, and it turns out that the optimum exists if and only if the covariance matrix σ_{ij} is non-singular and

$$B^2 - AC + A\phi^2 > 0, \quad (7)$$

where we introduced the notations $A = \sum_{ij} \sigma_{ij}^{-1}$, $B = \sum_{ij} \sigma_{ij}^{-1}\mu_j$ and $C = \sum_{ij} \mu_i\sigma_{ij}^{-1}\mu_j$. As long as these conditions are satisfied, the solution is given by

$$w_i^* = \frac{1}{2\eta^*} \sum_j \sigma_{ij}^{-1}(\mu_j - \lambda^*), \quad (8)$$

$$\lambda^* = \frac{B}{A} - \left[\left(\frac{B}{A} \right)^2 - \frac{C + \phi^2}{A} \right]^{1/2}, \quad (9)$$

$$\eta^* = \frac{1}{2} \left[\left(\frac{B}{A} \right)^2 - \frac{C + \phi^2}{A} \right]^{1/2}. \quad (10)$$

Condition (7) makes it clear that the existence of an optimal portfolio is not automatically guaranteed, but depends on the parameters of the underlying distribution (specifically on the expected values and covariances of the asset returns). When these parameters are estimated from a random sample, the fulfillment or violation of (7) (i.e. the feasibility of the optimization problem) will also be a random event.

4. The stability of parametric ES and VaR optimization on finite samples

4.1. The characterization of noise sensitivity

Let us now assume the position of an investor who knows that the returns are Gaussian, but does not know the parameters (i.e. the means, variances and covariances) of the distribution, so she has to estimate them from a finite sample. Let us assume she makes T independent observations, each consisting of a vector of N realized asset returns. This sample can be represented by an $N \times T$ matrix with elements x_{it} equal to the realized return of asset i over time period t ($i = 1, 2, \dots, N$ and $t = 1, 2, \dots, T$). The means μ_i and covariances σ_{ij} can be estimated by the unbiased estimators

$$\hat{\mu}_i = \frac{1}{T} \sum_{t=1}^T x_{it}, \quad (11)$$

$$\hat{\sigma}_{ij} = \frac{1}{T-1} \sum_{t=1}^T (x_{it} - \hat{\mu}_i)(x_{jt} - \hat{\mu}_j). \quad (12)$$

Then the risk of portfolio $\{w_i\}$ can be estimated by substituting $\hat{\mu}_i$ and $\hat{\sigma}_{ij}$ into (4). Let us denote this estimated risk by $\hat{\mathcal{R}}_\phi(\{w_i\})$. Now we can ask two fundamental questions:

- (i) Does $\hat{\mathcal{R}}_\phi(\{w_i\})$ have a minimum?
- (ii) If it does, how far is this minimum from the real optimum?

Question (i) can be answered by checking whether condition (7) is fulfilled by $\hat{\mu}_i$ and $\hat{\sigma}_{ij}$. As for Question (ii), first we need to specify how to measure the distance from the real optimum. To this end, we use the generalization of the measure q_0 introduced in [3], which in the present case is defined as follows.

If we know the parameters μ_i and σ_{ij} of the data generating distribution – for instance, in a simulation study like the ones in [3] and [13] – we explicitly know the true risk function $\mathcal{R}_\phi(\{w_i\})$. Let us assume that the data generating process is such that \mathcal{R}_ϕ has a minimum under the budget constraint, and let us denote the corresponding optimal weights by w_i^* . Our hypothetical investor, however, only knows the estimators $\hat{\mu}_i$ and $\hat{\sigma}_{ij}$, so she will minimize the estimated risk function $\hat{\mathcal{R}}_\phi(\{w_i\})$. Assuming that it exists, let this estimated optimum be \hat{w}_i^* . Although the investor might have the impression that this portfolio has risk $\hat{\mathcal{R}}_\phi(\{\hat{w}_i^*\})$ we know that its real risk is $\mathcal{R}_\phi(\{\hat{w}_i^*\})$ which, by definition, is greater than the risk in the true optimum $\mathcal{R}_\phi(\{w_i^*\})$. Therefore, the quantity

$$q_0 = \frac{\mathcal{R}_\phi(\{\hat{w}_i^*\})}{\mathcal{R}_\phi(\{w_i^*\})} \quad (13)$$

is a natural dimensionless measure of the distance of the estimated optimum from the true optimum. Moreover, the number $q_0 - 1$ has a straightforward interpretation: it is the percentage increase in the optimal risk the investor has to face due to the sampling error.

The properties of q_0 have been extensively studied both numerically and analytically for the case of global variance optimization [3, 4, 24]. Let us briefly summarize the main findings of these investigations:

- q_0 is a random variable which fluctuates from sample to sample, and its distribution depends on N and T .
- For large N and T and their ratio kept constant, $\mathbb{E}q_0^2 = (1 - N/T)^{-1}$ (\mathbb{E} denotes the average over sample fluctuations).
- In the same limit (N/T is held constant and $N \rightarrow \infty$) the variance of q_0^2 vanishes.

In other words, the estimation error q_0 is a self-averaging quantity, and for large N and T its average only depends on the ratio $r = N/T$. The divergence of q_0 in the limit $r \rightarrow 1$ can be regarded as the manifestation of an algorithmic phase transition, with a critical point $r_c = 1$ and a critical exponent $-1/2$ for the estimation error $q_0 \sim (r_c - r)^{-1/2}$.

Further studies of the noise sensitivity of portfolio optimization led to the conclusion that the critical behavior of the estimation error is similar to the above for a number of other risk measures (e.g. mean absolute deviation, maximal loss, non-parametric Expected Shortfall [13]) and data generating processes (e.g. GARCH [25]). As we shall see in the following section, parametric ES and VaR also belong to the same universality class.

4.2. The replica approach

Averaging over samples is the same as what is called quenched averaging (see e.g. [26]) in the statistical physics of disordered systems. Therefore, the heuristic replica method

that has been so successful in that field can also be used effectively to investigate the noise sensitivity of portfolio optimization [23, 27]. In this section, we are going to employ the replica approach 1) to determine under what circumstances the optimum exist, and 2) to compute q_0^2 provided that there is an optimum. The computations will be performed in the 'thermodynamic' limit, that is when $N \rightarrow \infty$ while $r = N/T$ is finite and fixed.

For the sake of simplicity we are going to assume that the data generating distribution is iid standard normal, in other words, the elements x_{it} of the sample matrix are identically distributed and mutually independent standard normal random variables. (It would be possible to relax the assumptions of zero means and zero correlations and still perform the computations, but this would make our argument less straightforward while the main qualitative conclusions would remain the same.) Since in this case $\mu_i = 0$ and $\sigma_{ij} = \delta_{ij}$, the true risk of a portfolio $\{w_i\}$ will be

$$\mathcal{R}_\phi(\{w_i\}) = \phi \sqrt{\sum_{i=1}^N w_i^2}. \quad (14)$$

For later convenience, we are going to use a modified form of the budget constraint:

$$\sum_{i=1}^N w_i = N, \quad (15)$$

which obviously does not change the nature of the optimization problem (it only rescales the result by a factor of N). Thus, the minimum of (14) subject to (15) will be the portfolio with weights $w_1^* = w_2^* = \dots = w_N^* = 1$, and the minimal risk will be $\mathcal{R}_\phi^* = \phi\sqrt{N}$. Hence, for a standard normal data generating distribution the distance of a portfolio $\{w_i\}$ from the true optimum is given by

$$q_0^2 = \frac{1}{N} \sum_{i=1}^N w_i^2. \quad (16)$$

(It is worth noting that in the special case of iid standard normal returns, we get exactly the same formula, if we measure the risk by standard deviation.)

It is clear that VaR/ES optimization based on a sample $\{x_{it}\}$ can be regarded as a statistical physics problem. Combining equations (6), (11) and (12) the Hamiltonian of the problem can be written as

$$\mathcal{H}(\{w_i\}, z, \eta) = N\phi\sqrt{z} - \frac{N}{T} \sum_{i=1}^N \sum_{t=1}^T w_i x_{it} + \eta \left[\sum_{t=1}^T \left(\sum_{i=1}^N w_i x_{it} \right)^2 - Tz \right], \quad (17)$$

where we replaced the factor $1/(T-1)$ by $1/T$ in equation (12), which makes no difference in the thermodynamic limit. (The budget constraint is not explicitly included in the Hamiltonian, but it will be taken into account soon.) We are interested in finding the ground state of this system. It is expedient, however, first to introduce a fictitious inverse temperature β and work out the partition function Z for finite temperature. The partition function is a functional of $\phi(\alpha)$ and the sample x_{it} :

$$\begin{aligned} Z_\beta[\phi; \{x_{it}\}] &= \int_{-\infty}^{\infty} \prod_{i=1}^N dw_i \int_0^{\infty} dz \int_0^{\infty} d\eta \delta \left(\sum_{i=1}^N w_i - N \right) e^{-\beta \mathcal{H}(\{w_i\}, z, \eta)} = \\ &= \int_{-\infty}^{\infty} \prod_{i=1}^N dw_i \int_{-\infty}^{\infty} d\lambda e^{i\lambda(\sum_{i=1}^N w_i - N) - \beta \frac{N}{T} \sum_{i=1}^N \sum_{t=1}^T w_i x_{it}} \times \end{aligned} \quad (18)$$

$$\times \int_0^\infty dz \int_0^\infty d\eta e^{N\beta\phi\sqrt{z} + \beta\eta \left[\sum_{t=1}^T (\sum_{i=1}^N w_i x_{it})^2 - Tz \right]}.$$

Then the risk at the optimum, estimated from sample $\{x_{it}\}$, is computed as:

$$\hat{\mathcal{R}}_\phi^* = - \lim_{\beta \rightarrow \infty} \frac{1}{\beta N} \log Z_\beta [\phi; \{x_{it}\}]. \quad (19)$$

This is nothing but the free energy density at zero temperature (i.e. the ground state energy density).

The free energy and all the "thermal averages" one can derive from it depend on the random sample. In general, one is interested in computing averages over the sample fluctuations (e.g. $\mathbb{E}q_0^2$), so we have to average the free energy over the random samples. To obtain $\mathbb{E}\hat{\mathcal{R}}_\phi^*$ we have to compute $\mathbb{E} \log Z_\beta [\phi; \{x_{it}\}]$. Averaging the logarithm of a random variable is a hard task. The replica method (see e.g. [26]) was invented to circumvent this difficulty by the use of the identity

$$\log Z = \lim_{n \rightarrow 0} \frac{Z^n - 1}{n}, \quad (20)$$

and computing $\mathbb{E}Z^n$ for positive integer n , which is a relatively simple task. In order to be able to take the $n \rightarrow 0$ limit, ultimately one has to analytically continue to real n . The name of the method derives from the fact that Z^n is the partition function of a system that consists of n identical copies (replicas) of the original problem. The Achilles heel of the method is the analytic continuation whose uniqueness usually cannot be guaranteed; we will justify its use ex post by the simulation results to be presented in the next section.

The sample elements x_{it} are independent and identically distributed random variables, so assuming a variance of $1/N$ their joint probability distribution function is

$$p(\{x_{it}\}) = \left(\frac{N}{2\pi}\right)^{NT/2} \exp\left(-\frac{N}{2} \sum_{i=1}^N \sum_{t=1}^T x_{it}^2\right). \quad (21)$$

We can compute $\mathbb{E}Z^n$, by expressing Z^n as the product of n independent, identical integrals over the replicated variables w_i^a , z^a and η^a ($a = 1, 2, \dots, n$), and then taking its average with respect to the density function (21). After computing several Gaussian integrals we arrive at the expression

$$\mathbb{E}Z_\beta^n [\phi] \propto \int_{-\infty}^\infty dQ^{ab} \int_{-i\infty}^{i\infty} d\hat{Q}^{ab} \int_0^\infty dz \int_0^\infty d\eta e^{NG_\beta(\{Q^{ab}\}, \{\hat{Q}^{ab}\}, \{z^a\}, \{\eta^a\})} \quad (22)$$

where we omitted the normalizing factor and used the notations

$$\begin{aligned} G_\beta(\{Q^{ab}\}, \{\hat{Q}^{ab}\}, \{z^a\}, \{\eta^a\}) &= \\ &= \sum_{a,b=1}^n \hat{Q}^{ab} (Q^{ab} - 1) - \frac{1}{2} \text{Tr} \log \hat{\mathbf{Q}} - \frac{1}{2r} \text{Tr} \log \mathbf{Q} - \\ &- \beta \sum_{a=1}^n \phi \sqrt{z^a} + \frac{\beta}{r} \sum_{a=1}^n \eta^a z^a + \frac{1}{N} \log A_\beta(\{Q^{ab}\}, \{\eta^a\}), \end{aligned} \quad (23)$$

and

$$A_\beta(\{Q^{ab}\}, \{\eta^a\}) = \int_{-\infty}^\infty du_t^a \exp \left[-\frac{1}{2} \sum_{a,b=1}^n (\mathbf{Q}^{-1})^{ab} \sum_{t=1}^T u_t^a u_t^b + \beta r \sum_{a=1}^n \sum_{t=1}^T u_t^a \right] \times$$

$$\times \exp \left[-\beta \sum_{a=1}^n \sum_{t=1}^T \eta^a u_t^{a2} + \beta \frac{r}{N} \sum_{a=1}^n \eta^a \left(\sum_{t=1}^T u_t^a \right)^2 \right]. \quad (24)$$

Here we introduced the so called overlap matrix

$$Q^{ab} = \frac{1}{N} \sum_{i=1}^n w_i^a w_i^b. \quad (25)$$

and its conjugate \hat{Q}^{ab} , which is a Lagrange multiplier to enforce the equality above. As we are interested in the $N \rightarrow \infty$ limit, we can use the saddle point method to compute the integral (22). Since we are dealing with a convex optimization problem, we expect that the saddle point is replica symmetric, that is we assume that $Q^{ab} = q + \Delta q \delta^{ab}$, $\hat{Q}^{ab} = \hat{q} + \Delta \hat{q} \delta^{ab}$, $\eta^a = \eta$ and $z^a = z$. After eliminating \hat{q} and $\Delta \hat{q}$ by partial extremization, we get $G_\beta(q, \Delta q, z, \eta) = n[g_0 + \beta g_\beta(q, \Delta q, z, \eta)] + O(n^2)$, where g_0 is some constant and

$$g_\beta(q, \Delta q, z, \eta) = -\frac{1}{2\beta\Delta q} - \frac{1-r}{2\beta r} \left(\log \Delta q + \frac{q}{\Delta q} \right) - \phi\sqrt{z} + \frac{1}{r}z\eta + \frac{1}{\beta N n} \log A(q, \Delta q, \eta) \quad (26)$$

$$A_\beta(q_0, \Delta q, \eta) = \int_{-\infty}^{\infty} du_t^a \exp \left[\frac{q}{2\Delta q^2} \sum_{t=1}^T \left(\sum_{a=1}^n u_t^a \right)^2 - \frac{1}{2\Delta q} \sum_{a=1}^n \sum_{t=1}^T u_t^{a2} \right] \times \exp \left[-\beta\eta \sum_{a=1}^n \sum_{t=1}^T u_t^{a2} + \beta r \sum_{a,t} u_t^a + \beta\eta \frac{r}{N} \sum_{a=1}^n \left(\sum_{t=1}^T u_t^a \right)^2 \right] \quad (27)$$

In the thermodynamic limit, the optimum can be obtained by minimizing the free energy density, which works out to be

$$f_\beta(q, \Delta q, z, \eta) = -\frac{1}{\beta} \lim_{N \rightarrow \infty} \frac{1}{N} \lim_{n \rightarrow 0} g_\beta(q, \Delta q, z, \eta). \quad (28)$$

In this limit $\log A_\beta(q_0, \Delta q, \eta)/Nn$ can be computed explicitly by performing the Hubbard-Stratonovich transformation twice to linearize quadratic terms in the exponent of the integrand, then computing a few more Gaussian integrals and approximating the logarithm function by its series expansion around 1. Finally we get

$$f_\beta(q, \Delta, z, \eta) = \frac{1}{2\Delta} + \frac{1-r}{2r} \left(\frac{1}{\beta} \log \frac{\Delta}{\beta} + \frac{q}{\Delta} \right) + \phi\sqrt{z} - \frac{1}{r}z\eta + \frac{1}{2r} \left[\frac{1}{\beta} \log \left(\frac{2\pi\Delta}{1+2\eta\Delta} \right) + \frac{q}{\Delta + 2\eta\Delta^2} + \Delta r^2 \right] \quad (29)$$

where we introduced the variable $\Delta = \beta\Delta q$. It is clear that in the zero temperature ($\beta \rightarrow \infty$) limit the free energy density is finite only if Δ remains a non-zero, finite constant. (In other words, the difference between the diagonal and non-diagonal elements of the replica matrix is proportional to β^{-1} , therefore, it vanishes in the zero temperature limit.)

Introducing the new variables $\eta' = 2\eta\Delta$ and $q' = q/\Delta^2$ we obtain the zero temperature free energy density in the form:

$$f_0(q', \Delta, z, \eta') = \phi\sqrt{z} - \frac{z\eta'}{2r\Delta} + \frac{1}{2\Delta} + \frac{\Delta}{2r} \left[\left(1 + r + \frac{1}{1+\eta'} \right) q' + r^2 \right]. \quad (30)$$

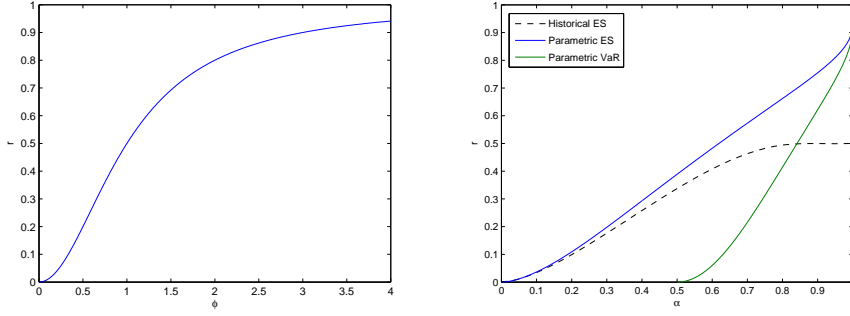


Figure 2. *Left panel:* The curve of critical N/T values as a function of ϕ . *Right panel:* The curves of critical N/T values as a function of α for VaR and ES. For comparison purposes, the critical curve of the historical ES optimization problem is also plotted with a black dashed line.

The saddle point conditions now read

$$\frac{\partial f}{\partial q'} = \frac{\partial f}{\partial \Delta} = \frac{\partial f}{\partial z} = \frac{\partial f}{\partial \eta'} = 0. \quad (31)$$

which implies that the solution is

$$q'^* = \phi^2 \quad (32)$$

$$\Delta^* = [(1-r)\phi^2 - r]^{-1/2} \quad (33)$$

$$\eta'^* = \frac{r}{1-r} \quad (34)$$

$$z^* = \frac{(1-r)^2}{4} \phi^2 \quad (35)$$

From (33) it is clear that the saddle point method is only meaningful, if $(1-r)\phi^2 - r > 0$. That is, in the thermodynamic limit, for each value of ϕ there is a critical value r_c of $r = N/T$ so that the optimization problem is not feasible unless $r < r_c$. (This stability condition corresponds to (7) in the thermodynamic limit.) Equation (33) implies that the critical values r_c are on the curve

$$r_c(\phi) = \frac{\phi^2}{\phi^2 + 1}, \quad (36)$$

which divides the r vs ϕ plane into two distinct phases: one in which the optimization is feasible and another one in which it is not. The implied phase diagrams can be seen in Figure 2. The left panel shows the phase boundary in the r vs ϕ plane. It is interesting to take a look at the asymptotic behavior of $r_c(\phi)$: as it increases in a strictly monotonous manner and $\lim_{\phi \rightarrow \infty} r_c(\phi) = 1$, it is clear that $r_c(\phi) < 1$ for any finite ϕ . In other words, for any confidence level $\alpha < 1$ (whether we are dealing with VaR or ES) the minimal length of the time series that ensures the existence of the optimum must be greater than N .

Substituting the formulas in (2) into (36) we get the phase boundaries of VaR and ES, respectively, in the r vs α plane (right hand side of Figure 2). It can be seen that parametric VaR optimization is more unstable than the parametric optimization of ES, although for practically relevant values of α (that is in the $\alpha > .9$ range) the

difference is not very significant. (For instance, for $\alpha = 99\%$ the critical value r_c is about 0.844 and 0.877 for VaR and ES respectively.) An interesting feature of both phase diagrams is that close to $\alpha = 1$ they tend to $r = 1$ with infinite derivatives.

The right panel of Figure 2 also shows the phase boundary of historical ES, so we can easily compare it to the critical curve of parametric ES. It is clear that the non-parametric phase curve is below the parametric one for any confidence level α , therefore the parametric estimation is more stable. In other words, a shorter time series is enough to ensure the feasibility of portfolio optimization, if parametric ES estimation is used. This was to be expected, but it is important to stress that although parametric fitting reduces the chance that there is no optimum for a given sample (especially for larger values of α), it fails to completely eliminate the feasibility problem originally encountered in historical estimation [13].

Let us now derive the sample average of the noise sensitivity measure q_0^2 in the thermodynamic limit, provided the optimum exists. Let us denote this conditional sample average by $\tilde{\mathbb{E}}$. From (16) and the replica symmetric ansatz it follows that $q_0^2 = q + \Delta/\beta$. Therefore, in the $\beta \rightarrow \infty$ limit we find that the conditional average of the estimation error of the optimal portfolio is

$$\tilde{\mathbb{E}}q_0^2 = q^* \cdot \Delta^{*2} = \frac{\phi^2(\alpha)}{(1-r)\phi^2(\alpha) - r} = \frac{r_c(\alpha)}{r_c(\alpha) - r}. \quad (37)$$

That is, $q_0 \sim (r_c - r)^{-1/2}$, so the estimation error of the parametric VaR and ES optimization displays the same critical behavior as the minimization of variance, mean absolute deviation, maximal loss and non-parametric ES. More generally, it is very probable that the parametric ES and VaR belong to the same universality class as the aforementioned risk measures, which would imply that q_0^2 is self-averaging (that is its variance vanishes in the thermodynamic limit) also here. This is clearly supported by numerical simulations and should be possible to confirm by a (very hard) replica calculation which is, however, beyond the scope of this paper.

4.3. Numerical study

In view of the heuristic character of the previous computation, we feel it is useful to provide numerical evidence to support its results. In order to do this, we generated independent samples from a multivariate standard normal distribution ($\mu_i = 0$ and $\sigma_{ij} = \delta_{ij}$), and attempted to find the minimum of $\mathcal{R}_\phi(\{x_{it}\})$ in each sample. For the sake of simplicity, rather than controlling the value of α , we controlled ϕ directly. To measure the probability of the existence of a minimum for a given combination of N , T and ϕ , we used the following algorithm:

- (i) Generate an $N \times T$ sample matrix $\{x_{it}\}$.
- (ii) Estimate the means and the covariances from $\{x_{it}\}$ using equations (11) and (12).
- (iii) Use the condition (7) to check if the portfolio optimization problem is feasible on the sample $\{x_{it}\}$.
- (iv) Repeat steps (i) to (iii) K times, and count how many times the optimum exists. Let this number be L . Then the estimated probability of feasibility will be $\hat{p}(N, T, \phi) = L/K$.

Clearly, the larger K the more accurate the measurement will be.

The left panel of Figure 3 exhibits simulation results for $\phi = 2$, which corresponds to confidence levels of $\alpha = 0.9772$ for VaR and $\alpha = 0.9420$ for ES. The number of

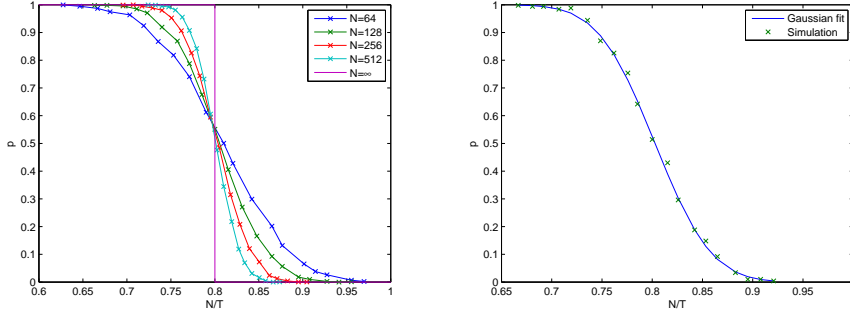


Figure 3. *Left panel:* The estimated probability p of the existence of an optimum as a function of N/T for $\phi = 2$ and different values of N . The curve labeled by $N = \infty$ corresponds to the thermodynamic limit (as computed by the replica method). *Right panel:* Gaussian curve fitting to the measured probabilities for $N = 128$ and $\phi = 2$.

iterations was $K = 2000$ and the p vs ϕ curve was measured for different values of N (64, 128, 256 and 512). Based on the previous section, the critical value of N/T is $r_c = 0.8$, that is, in the thermodynamic limit the optimum exists with probability 1 if $N/T < 0.8$ and it abruptly drops to 0 at the critical value (this is represented by the curve labeled by $N = \infty$ in the figure). The diagram shows that for finite values of N and T the probability of the existence of the optimal portfolio decreases from 1 to 0 continuously. At the same time, as N increases (that is, as we approach the thermodynamic limit) the fall of the probability from 1 to 0 becomes sharper and sharper, as expected. The probability curves belonging to different values intersect one another at the same point, therefore, this point must correspond to the critical value r_c . As shown by the figure, the intersection is, indeed, very close to $r = 0.8$, in excellent agreement with the analytical results.

We also observed that the probability curves fit very well to the function $g_{\mu,\sigma}(x) = 1 - \Phi((x - \mu)/\sigma)$ where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution, and μ and σ are parameters to be determined (e.g. via maximum likelihood estimation). The right hand panel shows simulated data points for $N = 128$ and $\phi = 2$ along with the fitted curve (where $\mu = 0.8028$ and $\sigma = 0.0446$). It is clear that $g_{\mu,\sigma}(x)$ cannot be the exact model for the p vs N/T curve, since for $N/T > 1$ we have $p = 0$. This fact, however, gradually loses its significance as N increases, and σ gets smaller and smaller. As a result, fitting $g_{\mu,\sigma}(x)$ to the numerically computed data points makes it possible to estimate p as a function of ϕ and N/T with a high accuracy, even if the number of iterations K is low; this way simulations can be speeded up by a factor ranging from 10 to 100.

Our numerical study showed that around the critical value $r_c(\phi)$ the probability $p(N, T, \phi)$ follows the behavior displayed in Figure 3 for any value of ϕ , but the steepness of the decline from 1 to 0 varies with ϕ . To demonstrate this, we numerically computed the contour lines of constant p on the N/T vs ϕ plane for $p = 0.1, 0.3, 0.5, 0.7$ and 0.9 with $N = 128$ (the number of iterations was set to $K = 100$, and we fitted $g_{\mu,\sigma}(x)$ to the simulated data points). The results are shown on the left hand side of Figure 4. Comparing this diagram to the left panel of Figure 2 it is evident that the contour lines are arranged around, and have a similar shape to, the theoretical

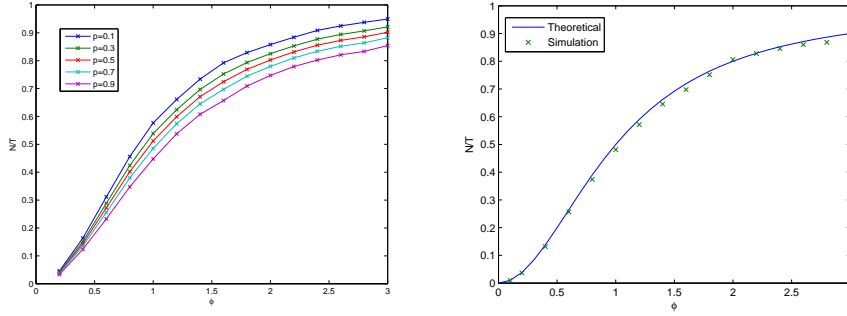


Figure 4. *Left panel:* Contour lines of fixed p for $N = 128$, on the N/T vs ϕ plane. *Right panel:* Phase boundary in the $N \rightarrow \infty$ and N/T finite limit.

phase boundary. As mentioned above, the critical points can be estimated as the intersections of the p vs N/T curves for different values of N . The green points on the right hand panel were numerically computed by fitting $g_{\mu,\sigma}(x)$ to simulated data with $N = 64$ and $N = 128$, and then calculating the intersection of the two fitted curves (the number of iterations was $K = 100$). The estimated critical points (in green) and the computed phase boundary (in blue) line up very well, which confirms the validity of the results obtained through the replica method.

4.4. A note on semivariance

Semivariance is one of the oldest downside risk measures. As we shall see, the results obtained in the previous sections can be directly applied to characterize the stability of portfolio optimization under semivariance, when it is estimated by parametric fitting.

The definition of semivariance is

$$\nu^2(X) = \mathbb{E}[\max\{0, X - \mathbb{E}(X)\}]^2, \quad (38)$$

where X is a random variable representing the return of some security. The measure ν is simply called semi standard deviation, and this quantity can be used to define the following, VaR/ES-like risk measure (which is sometimes called semivariance too, leading to some confusion):

$$\rho(X) = \nu(X) - \mathbb{E}(X). \quad (39)$$

When the variable X is normally distributed with mean μ and standard deviation σ the semi standard deviation is simply $\nu = \sigma/\sqrt{2}$, so the risk measure ρ can be written as

$$\rho(X) = \frac{1}{\sqrt{2}}\sigma - \mu, \quad (40)$$

which is exactly of the same form as (1) with $\phi = 1/\sqrt{2} \approx 0.71$.

This implies immediately that in the case of semivariance minimization the critical value of N/T is $r_c = 1/3$, that is, for large N (i.e. close to the thermodynamic limit), we need a time series that is at least three times as long as the number of assets in the portfolio, in order to have a meaningful optimization problem. Moreover, the conditional average of q_0^2 will be $\tilde{\mathbb{E}}q_0^2 = (1/3 - N/T)^{-1}/3$.

5. Summary

We studied the feasibility and noise sensitivity of portfolio optimization under Value-at-Risk, Expected Shortfall and semivariance in the case when these risk measures are estimated from finite samples using parametric fitting. Similarly to earlier studies based on non-parametric estimation [13, 23] we assumed independent standard normal asset returns. We found that the probability that the optimum exists on a given finite sample is smaller than unity, and this probability is a function of the portfolio size, the sample size and the confidence level of VaR/ES. In the thermodynamic limit (where the portfolio size N goes to infinity but its ratio to the sample size T is held constant), this probability converges to either 0 or 1 depending on N/T and the confidence level α . We employed the replica method to compute the equation of the curve separating the feasible and unfeasible regions on the N/T vs α plane, and also tested and supported the result by numerical simulation. The replica approach also enabled us to compute the average of the measure of noise sensitivity q_0^2 , contingent on the feasibility of the optimization problem. It is highly probable that the parametric ES, VaR and semivariance optimization problems belong to the same universality class as the optimization of many other risk measures (standard deviation, mean absolute deviation, maximal loss, non-parametric ES): we found that the estimation error blows up with a critical exponent $-1/2$ as we approach the phase boundary.

Our results make it possible to compare the parametric and historical estimator of ES. It is clear that parametric estimation does not eliminate the instability of the historical estimator, but it does improve on it, in that the phase diagram of parametric ES runs above the historical curve. This means that for a given confidence level and a given portfolio size we need more data (longer time series) in the historical estimation than in the parametric one, in order to have a meaningful solution to the optimization problem. It seems as if we had some additional source of information in the parametric case. (The effect is even more pronounced in the case of VaR, where the historical estimate cannot be guaranteed to be convex for any confidence level and any length of the time series, whereas the parametric estimate has been shown here to have an optimum at least in a certain region of parameter space.) One may wonder where this additional information may have come from. The answer is simple: in the historical estimation we do not make any assumption about the nature of the underlying distribution, we are just using raw data as they are produced by the data generating process. In contrast, in the parametric case we assume that the process is Gaussian and fit the data to this assumption. This way we are projecting a nontrivial piece of information into the estimation. For technical reasons we have indeed chosen a Gaussian underlying process in the context of this work, but in a real market return fluctuations are neither Gaussian, nor even stationary. To project an arbitrary distribution into real, parsimonious data may produce apparently more stable estimates, but the gain may well turn out to be completely illusory and the results misleading.

We would also like to draw attention to the fact that the critical value of the N/T ratio depends on the risk measure and on the (historical or parametric) method of estimation. This critical ratio is never larger than 1, and, depending on the risk measure and on the confidence level, it may be significantly smaller; e.g., as we have just seen, for the semivariance e.g. it is as low as $1/3$. This means that, depending on the risk measure, we need time series longer than two or three times the size of the portfolio, in order to have a solution at all, and much longer, in order to have a reliable

estimate. In the context of portfolio selection, where, by the very nature of the task, the sampling frequency cannot be higher than once a week or even once a month, this condition is not easy to satisfy. Therefore, in practice the typical N/T ratio may be fairly close to the phase boundary where the estimation error diverges. The knowledge of the phase boundary and the position of our working point (confidence level and N/T ratio) relative to it is highly important if we wish to take sample to sample fluctuations properly into account.

This work has presented further evidence for the instability of widely used risk measures against sample fluctuations. The instability of parametric VaR, easily the most popular risk estimate, is particularly notable. We find it remarkable how powerful the concepts and methods imported from the statistical physics of random systems prove to be in the analysis of these important phenomena.

Acknowledgments

This work has been supported by the "Cooperative Center for Communication Networks Data Analysis", a NAP project sponsored by the National Office of Research and Technology.

References

- [1] H. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. J. Wiley and Sons, New York, 1959.
- [2] S. Pafka and I. Kondor. Noisy covariance matrices and portfolio optimization. *European Physical Journal B*, 27:277–280, 2002.
- [3] S. Pafka and I. Kondor. Noisy covariance matrices and portfolio optimization II. *Physica A*, 319:487–494, 2003.
- [4] S. Pafka and I. Kondor. Estimated correlation matrices and portfolio optimization. *Physica A*, 343:623–634, 2004.
- [5] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, Feb 2004.
- [6] O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621, Dec 2003.
- [7] O. Ledoit and M. Wolf. Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*, 31(1), 2004.
- [8] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters. Noise dressing of financial correlation matrices. *Phys. Rev. Lett.*, 83:1467, 1999.
- [9] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters. Random matrix theory and financial correlations. *Int. J. Theor. Appl. Finance*, 3:391, 2000.
- [10] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, and H.E. Stanley. Universal and non-universal properties of cross-correlations in financial time series. *Phys. Rev. Lett.*, 83:1471, 1999.
- [11] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley. A random matrix approach to cross-correlations in financial data. *Phys Rev E*, 65:066136, 2002.
- [12] Z. Burda, A. Goerlich, A. Jarosz, and J. Jurkiewicz. Signal and noise in correlation matrix. *Physica A*, 343:295, 2004.
- [13] I. Kondor, S. Pafka, and G. Nagy. Noise Sensitivity of Portfolio Selection under Various Risk Measures. *Journal of Banking and Finance*, 31:1545–1573, 2007.
- [14] H. Konno. Portfolio Optimization Using L_1 Risk Function. Technical Report IHSS 88-9, Institute of Human and Social Sciences, Tokyo Institute of Technology, 1988.
- [15] M.R. Young. A Minimax Portfolio Selection Rule with Linear Programming Solution. *Management Science*, 44:673–683, 1998.
- [16] C. Acerbi, C. Nardio, and C. Sirtori. Expected Shortfall as a Tool for Financial Risk Management. Working paper, 2001. (<http://www.gloriamundi.org/detailpopup.asp?ID=453055940>).

- [17] C. Acerbi and D. Tasche. Expected Shortfall: a Natural Coherent Alternative to Value at Risk. *Economic Notes*, 31(2):379–388, 2002.
- [18] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Thinking Coherently. *Risk*, 10(11):68–71, 1997.
- [19] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent Measures of Risk. *Math. Fin.*, 9(3):203–228, 1999.
- [20] I. Kondor and I. Varga-Haszonits. Feasibility of portfolio optimization under coherent risk measures. arXiv:0803.2283v3 [physics.soc-ph], 2008.
- [21] Ph. Jorion. *VaR: The New Benchmark for Managing Financial Risk*. McGraw-Hill, New York, 2000.
- [22] R. T. Rockafellar and S. Uryasev. Optimization of Conditional Value-at-Risk. *Journal of Risk*, 2(3):21–41, 2000.
- [23] S. Ciliberti, I. Kondor, and M. Mézard. On the Feasibility of Portfolio Optimization under Expected Shortfall. *Quantitative Finance*, 4:389–396, 2007.
- [24] Z. Burda, J. Jurkiewicz, and M. A. Nowak. Is econophysics a solid science? *Acta Physica Polonica B*, 34:87–132, 2003.
- [25] I. Varga-Haszonits and I. Kondor. Noise Sensitivity of Portfolio Selection in Constant Conditional Correlation GARCH Models. *Physica A*, 385:307–318, 2007.
- [26] M. Mézard, G. Parisi, and M.A. Virasoro. *Spin Glass Theory and Beyond*, volume 9 of *World Scientific Lecture Notes in Physics, Singapore*. 1987.
- [27] S. Ciliberti, , and M. Mézard. Risk Minimization through Portfolio Replication. *European Physical Journal B*, 57:175–180, 2007.