Analytic solution to variance optimization with no short positions

Imre Kondor[1,2,3], Gábor Papp[4], Fabio Caccioli[5,6]
1-Parmenides Foundation, Pullach, Germany
2- London Mathematical Laboratory, London, UK
3- Complexity Science Hub, Vienna, Austria
4- Eötvös Loránd University, Institute for Physics, Budapest, Hungary
5- University College London, Department of Computer Science, London, WC1E 6BT, UK
6- Systemic Risk Centre, London School of Economics and Political Sciences, London, UK

May 28, 2017

Abstract

A portfolio of independent, but not identically distributed, returns is optimized under the variance risk measure with a ban on short positions, in the high-dimensional limit where the number N of the different assets in the portfolio and the sample size T are assumed large with their ratio $r = N/T$ kept finite. To the best of our knowledge, this is the first time such a constrained optimization is carried out analytically, which is made possible by the application of methods borrowed from the theory of disordered systems. The no-short-selling constraint acts as an asymmetric ℓ_1 regularizer, setting some of the portfolio weights to zero and keeping the out-of-sample estimator for the variance bounded, avoiding the divergence present in the non-regularized case. However, the ban on short positions does not prevent the phase transition in the optimization problem, only shifts the critical point from its non-regularized value of $r = 1$ to 2, and changes its character: at $r = 2$ the out-of-sample estimator for the portfolio variance stays finite and the estimated in-sample variance vanishes, while another critical parameter, related to the estimated portfolio weights and the condensate density, diverges at the critical value $r = 2$. We have performed numerical simulations to support the analytic results and found perfect agreement for $N/T < 2$ in the large N limit. Numerical experiments on finite size samples of symmetrically distributed returns show that above $r = 1$ solutions with zero in-sample variance start to sporadically arise, their probability of appearance increasing as r approaches 2, steeply rising around the critical point, and becoming nearly one beyond $r = 2$. A closed formula obtained for this probability shows that in the large N limit the transition becomes sharp. The zero in-sample variance solutions are not legitimate solutions of the optimization problem, as they are infinitely sensitive to any change in the input parameters, in particular they will wildly fluctuate from sample to sample. With some narrative license we may say that the no-short constraint, with prohibiting large compensating positions, takes care of the longitudinal (length) fluctuations of the optimal weight vector, but does not eliminate the divergent transverse fluctuations corresponding to a rearrangement of the composition of the portfolio. We also calculate the distribution of the optimal weights over the random samples and show that the regularizer preferentially removes the assets with large variances, in accord with one's natural expectation.

1 Introduction

Institutional portfolios are often optimized under a ban on short positions. If the distribution of the returns on the securities making up the portfolio is exactly known, the optimization is straightforward to carry out. In practice, this distribution is never known, but has to be inferred from observations in the market. If the available data is finite, the optimal estimated portfolio weights will be different from their true values, and the resulting portfolio will suffer from estimation error. This error will be particularly large if the dimension N of the portfolio (the number of different assets) is not small relative to the sample size (the length of available time series) T. This problem has been approached by various numerical methods, see e.g. [1] for an overview. In real life context of risk management or asset management a purely numerical approach may, however, be very computationally demanding and, as will be discussed below, may in addition be also misleading, especially if one lacks a full control over the optimization algorithm implemented in the risk management package and a good understanding of the structure of the problem.

Such an understanding can come from an analytic approach. Analytic calculations of the optimal estimated portfolio have been performed by various groups under the assumption that the underlying statistical distribution is normal, the objective function is the variance and the optimization is subject to the budget constraint and, in some cases, an ℓ_2 regularizer [2–15]. The most recent, nonlinear realization of ℓ_2 shrinkage [16–18] has turned out to be particularly effective in suppressing sample fluctuations. A special approach to portfolio optimization [19–26] rests on the replica method borrowed from the statistical physics of disordered systems [27]. These papers focused on the minimal risk portfolio, but [28] treated the full Markowitz problem [29] including the constraint on the expected return, while in [25, 30] an ℓ_2 constraint has been imposed on the portfolio weights. Such a regularizer can suppress large sample fluctuations that lead to a high degree of estimation error, especially in the high dimensional setting where both the dimension N and the sample size T are large. An alternative motivation for an ℓ_2 constraint is to prevent the over-concentration of the optimal portfolio on a small number of blue chips [30–32], a particularly strong tendency in small markets, and also by taking into account the market impact of a future liquidation of the portfolio already at the stage of its composition [21].

Considerations of transaction costs and the technical difficulty of frequent rebalancing a very large portfolio may make it desirable to reduce the dimension and strive for a sparse portfolio. This can be achieved by borrowing the popular and very successful ℓ_1 regularization from machine learning [33]. Jagannathan and Ma [4] were the first to notice that a ban on short positions improves the stability of estimated optimal portfolios, and it is clear that the exclusion of short positions can be regarded as a special case of ℓ_1 regularization. Subsequently Brodie et al [34] applied an ℓ_1 regularizer on the portfolio weights in an empirical study of real life portfolios in various markets and demonstrated its satisfactory performance compared with the $1/N$ portfolio [14].

To the best of our knowledge, no analytic result exists in the literature for portfolio optimization under an ℓ_1 constraint. The purpose of the present paper is to perform such an analytic optimization of the variance as the risk measure supplemented with a special case of the ℓ_1 constraint, a ban on negative portfolio weights. The method that makes this possible is again the replica method. In its simplest form that we apply here it assumes that return samples of size T are drawn from an N-dimensional normal distribution. These samples are then regarded as if arising from empirical observations, and the various quantities of interest are averaged over the ensemble of the samples. The averaging can be explicitly carried out in the high-dimensional limit where N and T go to infinity with their ratio $r = N/T$ kept finite. For simplicity, we will also assume that the expected return of each asset in the portfolio is zero and seek to determine the global minimal risk portfolio, but we allow the assets to have different variances in order to be able to study the effect of the no-short constraint on assets with different volatility. We are considering independent normal variables and assume that the returns are serially independent (zero autocorrelation). Although we always speak about the elements of the portfolio as assets, we do not necessarily mean individual securities; these elements can be viewed as a collection of arbitrary risk factors as long as their statistical properties conform to the above assumptions.

We also analyze the numerical aspects of this problem and find that in the large N limit the simulations precisely follow the theoretical curves up to the critical point $N/T = 2$. Above this critical point a continuum of zero variance solutions appear. While these solutions are clearly meaningless, numerical work in this region requires special care: some solvers (e.g. Matlab's fmincon) modify the problem in order to make sure a stable solution exists even when the covariance matrix is less than full rank. Without a careful study of the algorithm's description and without anticipating the instability, it is easy to overlook this phase transition.

To better understand the $r = 2$ transition, we also performed numerical experiments on finite N, finite T samples. These studies demonstrated that the zero in-sample variance solutions start to appear already above $r = 1$, but initially the probability of their appearance is very small. As we approach $r = 2$ from below, this probability starts to increase, steeply rising to values close to one as we cross the transition point. We display a closed combinatorial formula for this probability, and support it by extensive numerical simulations. This probability law is universal, independent of the nature of the probability distribution of the returns (as long as it is continuous and symmetric), and shows how the continuous transition goes over into a sharp, step-like one in the limit $N \to \infty$. Accordingly, the critical value $r_c = 2$ is also universal, independent of the replica method or the Gaussian assumption about the distribution of returns. The transition at $r = 2$ is in several respects similar to the one in the minimax risk measure described in [35].

The plan of the rest of the paper is as follows. For the sake of establishing a basis for later comparison and introducing some notation, in Sec. 2 we address the trivial problem of optimizing the variance assuming we have complete information, as if having an infinitely large sample. In Sec. 3 we consider the case of variance optimization without the no-short constraint, but now for $r = N/T$ finite. Some of the results here reproduce those known previously, but the distribution of weights is new, as is also the discussion of the geometry of the phase transition (that in the unconstrained case takes place at $r = 1$). Sec. 4 is the central part of the paper. Here, we perform the optimization of variance with a constraint forbidding short positions, and derive results for the estimator for the out-of-sample and the in-sample estimator for the portfolio variance, along with results for the distribution of weights over the random samples. This constitutes a complete solution of the no-short constrained problem, the first instance such a solution has been achieved by analytic means. Our formulae illustrate how a ban on short selling removes some of the assets from the portfolio, and how an asset's volatility affects the probability of its elimination. We identify the phase transition at $r = 2$ mentioned above, which is different in nature from the unconstrained one at $r = 1$ in that a new critical parameter diverges, but the estimation error stays finite here. Sec. 5 is a summary of the results. Technical details are relegated to two appendices. Appendix A presents the replica derivation of the free energy functional for the optimization of the variance supplemented by a generic constraint, while Appendix B derives the saddle point equations and the distribution of the weights.

2 Optimizing the variance with complete information, $r = 0$

In this section we present an analytic treatment of the optimization of the variance of a portfolio composed of N securities with zero expected returns and a diagonal covariance matrix with given elements σ_i^2 along the diagonal, $i = 1, 2, \ldots, N$. The risk σ_p^2 of the portfolio measured in terms of the variance is

$$\sigma_p^2 = \sum_i \sigma_i^2 w_i^2 \tag{2.1}$$

to be minimized under the budget constraint

$$\sum_i w_i = N, \tag{2.2}$$

where, instead of the usual 1, we normalized the portfolio weights w_i to N, in order to keep them of order unity. (In the following we will consider the dimension N of the portfolio as a large number, letting it go to infinity when the calculations so demand.) As the assets are assumed to have zero expected returns, we do not stipulate a constraint on the expected return of the portfolio, and seek the global minimum pof the risk.

The optimization problem (2.1), (2.2) is trivial to solve by the method of Lagrange multipliers. The minimum of

$$\sum_i \sigma_i^2 w_i^2 - \lambda(\sum_i w_i - N) \tag{2.3}$$

is at $w_i = \lambda/2\sigma_i^2$, and the budget constraint fixes the Lagrange multiplier to be

$$\lambda = \frac{2N}{\sum_i \frac{1}{\sigma_i^2}}. \tag{2.4}$$

The optimal portfolio weights are then obtained as

$$w_i^* = \frac{1}{\sigma_i^2} \frac{N}{\sum_j \frac{1}{\sigma_j^2}} \tag{2.5}$$

and the minimal risk is

$$\sigma_p^{*2} = \frac{N}{\frac{1}{N}\sum_j \frac{1}{\sigma_j^2}}. \tag{2.6}$$

For later convenience we define

$$F = \frac{T\sigma_p^{*2}}{2N} = \frac{1}{2r} \frac{N}{\frac{1}{N}\sum_j \frac{1}{\sigma_j^2}}, \tag{2.7}$$

and we will refer to this as the "free energy" or the cost function. The factor $1/(2r)$, where $r = N/T$, will then appear also in the Lagrange multiplier λ. If we define λ' as the Lagrange multiplier associated with the minimization of the free energy (2.7), we have that

$$\lambda' = \frac{1}{2r}\lambda = \frac{1}{2r} \frac{2}{\frac{1}{N}\sum_i \frac{1}{\sigma_i^2}}. \tag{2.8}$$

In the following we will always use λ' everywhere, and will omit the prime with no risk of confusion.

Note that due to the normalization of the weights F is of order N. In the following it will be convenient to consider the free energy per asset

$$f = \frac{1}{2r} \frac{1}{\frac{1}{N}\sum_j \frac{1}{\sigma_j^2}} = \frac{1}{2}\lambda \tag{2.9}$$

As already evident from (2.3), the Lagrange multiplier associated with the budget constraint must be positive; a negative value would correspond to no security in the portfolio at all. Thus λ plays a role analogous to the chemical potential, the quantity that governs the number of particles in a physical system, and, for brevity, we will refer to λ as the chemical potential in the following. The positivity of λ is completely trivial at this point, but it will acquire significance in the computations later: its vanishing will herald the phase transition.

The optimal weights are the larger the smaller their variance, in particular, if one of the securities is riskless, its weight takes up the full weight N. Also, if there is a riskless security in the portfolio, the whole portfolio becomes riskless and σ_p^* vanishes.

Note also that the no-short-selling condition did not have to be stipulated in this preliminary instance: the weights have worked out to be positive automatically. This will not remain true when the parameters of the model are estimated on the basis of finite samples.

The optimization problem as laid out above assumes that we have complete knowledge about the probability distribution of the returns: in particular we know the (zero) values of the expected returns and the values of the variances σ_i. In reality, we never have complete information. What we may have are samples of size T drawn from the joint distribution of returns, which in our setting is

$$P(\{x_{it}\}) = \prod_i \left(\sqrt{\frac{N}{2\pi\sigma_i^2}} e^{-Nx_{it}^2/2\sigma_i^2} \right) . \tag{2.10}$$

An important parameter of the problem is the ratio $r = N/T$. The larger the sample size T relative to the dimension N, the better the estimates we can make for the optimal weights and the optimal value of the risk. We expect, therefore, that in the limit $r \to 0$ we can retrieve the "true" values of the weights as given in (2.5), and the "true" value of the optimal risk, (2.6).

Present day institutional portfolios are large, with N's in the range of hundreds or thousands, while sample sizes are limited by stationarity considerations to below 1000 (four years worth of daily data) at most, but often much less. Therefore, the value of r is never really small in practice. This leads to large sample fluctuations, so large indeed that at a critical value of r the estimation error becomes infinite and the optimization meaningless. In the case of unregularized variance as risk measure, this critical value is $r_c = 1$, which is where the estimated covariance matrix loses its positive definiteness and the first zero eigenvalue appears.

Difficulties of a similar nature appear in countless problems in modern statistics and machine learning [36]. The remedy is to introduce regularizers, i.e. terms added to the cost function with the purpose of suppressing the large sample fluctuations. Of course, regularization will also introduce bias, but the hope is that a reasonable balance can be struck between bias and fluctuations.

Perhaps the most popular regularizer today is the one based on the ℓ_1 norm [37]. Its appeal was greatly enhanced by the proof by Candès et al. [38] that ℓ_1 can successfully imitate ℓ_0, the straight weeding out of the superfluous, irrelevant variables, thereby strongly reducing the dimension of the problem. In the portfolio context this would mean reducing the dimensionality by setting the weights to zero of the securities that are deemed irrelevant, presumably those with the largest volatilities. In the following, we are going to demonstrate the action of ℓ_1 regularization in the special case corresponding to a no-short selling constraint. Before addressing that problem, however, we wish to present the optimization of variance without the no-short constraint.

3 Unconstrained variance optimization

By "unconstrained" we mean dropping the no-short condition; the budget constraint will of course be upheld.

The relevant free energy functional is obtained from (B.9) by setting $\eta_1 = \eta_2 = 0$ and making use of the identity

$$W(x) + W(-x) = \frac{x^2+1}{2}, \tag{3.1}$$

satisfied by the transcendental function W appearing in (B.9) in Appendix B. Then f works out to be

$$f = \lambda - \Delta\hat{q}_0 - \hat{\Delta}q_0 + \frac{1}{2r}\frac{q_0}{1+\Delta} + \frac{\hat{q}_0}{2\hat{\Delta}} - \frac{\lambda^2}{4\hat{\Delta}}\frac{1}{N}\sum_i \frac{1}{\sigma_i^2}. \tag{3.2}$$

Setting the derivatives of f with respect to the "order parameters" λ, q_0, Δ, $\hat{\Delta}$ and $\hat{q}_0$ to zero gives the following saddle-point or stationarity conditions:

$$\lambda = 2\hat{\Delta}\left(\frac{1}{N}\sum_i \frac{1}{\sigma_i^2}\right)^{-1}, \tag{3.3}$$

$$\hat{\Delta} = \frac{1}{2r}\frac{1}{1+\Delta}, \tag{3.4}$$

$$\hat{q}_0 = -\frac{q_0}{2r(1+\Delta)^2}, \tag{3.5}$$

$$q_0 = -\frac{\hat{q}_0}{2\hat{\Delta}^2} + \frac{\lambda^2}{4\hat{\Delta}^2}\frac{1}{N}\sum_i \frac{1}{\sigma_i^2}, \tag{3.6}$$

$$\Delta = \frac{1}{2\hat{\Delta}}. \tag{3.7}$$

Combining (3.3)–(3.7) one can easily see that the cost function f at the saddle point is equal to

$$f = \frac{\lambda}{2}. \tag{3.8}$$

The solution of the saddle point equations is straightforward:

$$\lambda = \frac{1-r}{r}\frac{1}{\frac{1}{N}\sum_i \frac{1}{\sigma_i^2}}, \tag{3.9}$$

$$\Delta = \frac{r}{1-r}, \tag{3.10}$$

$$q_0 = \frac{1}{1-r}\frac{1}{\frac{1}{N}\sum_i \frac{1}{\sigma_i^2}}, \tag{3.11}$$

$$\hat{q}_0 = -\frac{1-r}{2r}\frac{1}{\frac{1}{N}\sum_i \frac{1}{\sigma_i^2}}, \tag{3.12}$$

$$\hat{\Delta} = \frac{1-r}{2r}, \tag{3.13}$$

and the free energy per asset is

$$f = \frac{1-r}{2r} \frac{1}{\frac{1}{N}\sum_i \frac{1}{\sigma_i^2}}. \tag{3.14}$$

Turning to the distribution of weights, we see from (B.11) and (B.12) that for $\eta_1 = \eta_2 = 0$ $w_i^{(1)} = w_i^{(2)}$, so the first term (the δ-peak of the zero weights) in (B.17) vanishes, while the second term becomes

$$p(w) = \frac{1}{N} \sum_i \frac{1}{\sigma_w^{(i)} \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{w - w_0^{(i)}}{\sigma_w^{(i)}} \right)^2 \right), \tag{3.15}$$

where

$$w_0^{(i)} = \frac{\lambda}{2\sigma_i^2 \hat{\Delta}} = \frac{\lambda r(1+\Delta)}{\sigma_i^2} \tag{3.16}$$

and

$$\sigma_w^{(i)} = \frac{\sqrt{q_0 r}}{\sigma_i}. \tag{3.17}$$

From (3.9) and (3.10) it follows that

$$w_0^{(i)} = \frac{1}{\sigma_i^2} \frac{N}{\sum_j \frac{1}{\sigma_j^2}}, \tag{3.18}$$

the same as in (2.5). Therefore, in the unconstrained optimization case the estimated weights fluctuate about their true values. This does not remain so once regularization is switched on.

The order parameters λ, Δ and q_0 have a direct meaning. As already seen in Section 2, λ is the "chemical potential", the Lagrange multiplier associated with the budget constraint. As such, it must be positive, and its vanishing signals an instability. In the present unconstained case, the quantity Δ is inversely proportional to the in-sample estimate for the free-energy. It is non-negative by definition, and its divergence is another signal of the instability that sets in for $\lambda = 0$. Finally, q_0 is related to the out-of-sample estimator of the variance. In [22] it was shown for the special case $\sigma_i = 1$, for all i, that $\sqrt{q_0} - 1$ is the relative estimation error. When the variances of returns in the portfolio are different, q_0 has to be normalized as [28]

$$\tilde{q}_0 = q_0 \frac{1}{N} \sum_i \frac{1}{\sigma_i^2} = \frac{1}{1-r}, \tag{3.19}$$

in order to make $\tilde{q}_0$ equal to the ratio between the optimal out-of-sample estimator for the risk of the portfolio (with weights $\hat{w}_i^*$) and the risk of the true optimal portfolio (with weights w_i^*)

$$\tilde{q}_0 = \frac{\sum_{ij} \sigma_{ij} \hat{w}_i^* \hat{w}_j^*}{\sum_{ij} \sigma_{ij} w_i^* w_j^*} \tag{3.20}$$

so that $\sqrt{\tilde{q}_0} - 1$ becomes the relative error associated with the estimation of risk. Because of the simple proportionality between q_0 and $\tilde{q}_0$, we will speak about q_0 as (the measure of) the out-of-sample estimation error. The divergence of q_0 or $\tilde{q}_0$ at $r = 1$ is pointing to the same instability as that of Δ or the vanishing of λ.

3.1 The limit of complete information

When $r \to 0$, the sample size T is much larger than the dimension N, so we have complete information and should be able to recover the results in Section 2.

This is indeed so: for $r \to 0$ (3.9) and (3.14) duly reproduce (2.4) and (2.7), respectively. From (3.11) and (3.19) we also see that $\tilde{q}_0 = 1$, that is the estimation error vanishes. Furthermore, (3.10) implies that Δ vanishes with r. Then from (3.9) and (3.16) it follows that

$$w^{(i)} = \frac{1}{\sigma_i^2} \frac{1}{\frac{1}{N}\sum_j \frac{1}{\sigma_j^2}} \tag{3.21}$$

is the weight of asset i in the optimal portfolio, in agreement with (2.5).

The width $\sigma_w^{(i)}$ of the Gaussian distribution of the weights over the samples goes to zero with r, so the distribution (3.15) becomes a series of δ-spikes

$$p(w) = \frac{1}{N}\sum_i \delta\left(w - w^{(i)}\right), \tag{3.22}$$

where δ is the Dirac δ-distribution.

3.2 The high-dimensional case and the instability

If r is not very small, N and T become comparable and we are in the high-dimensional setting. From (3.9)-(3.11) we see that with increasing r the chemical potential λ decreases, the estimation error q_0 increases, while the cost function f decreases. As a result of averaging over the samples, the sharp peaks in the distribution of weights in (3.15) broaden into Gaussians.

As we approach $r = 1$, Δ and the relative estimation error q_0 grow without bound, and the width of the Gaussian in (3.15) also diverges, so the different assets are not resolvable anymore. All these are signatures of an instability, divergent fluctuations from sample to sample, which we can rightly call a phase transition.

Note that in the same limit $r \to 1$ the chemical potential λ and the free energy f, the in-sample estimation of the cost, vanish.

The nature of this phase transition has been analyzed in detail in [28], where it was found that the eigenvalues of the Hessian (the matrix of the second derivatives of the replica functional) all vanish at the critical point $r = 1$. It is then clear that the results of the present section cannot be continued beyond this point, because the replica method, relying on a saddle point approximation (see Appendiy A), is bound to break down where the stability matrix becomes indefinite.

On the other hand, there is nothing to prevent us from considering large dimensions and relatively small samples, that is a situation when $r > 1$. What is happening in this region is the subject of the next subsection.

3.3 Linear algebraic interpretation of the instability at $r = 1$

In the simple case of the variance, the root of the instability at $r = 1$ is quite obvious; nevertheless it deserves a brief discussion here, especially because similar instabilities appear in several other risk measures including the Expected Shortfall [19], mean absolute deviation [20], the minimax problem [35], even in a GARCH-based non-stationary process [39], where they are considerably more difficult to explain. Moreover, we shall encounter a somewhat similar instability later when we introduce a constraint on short positions.

Let us consider the minimization of the empirical portfolio variance $\hat{\sigma}_p^2$ with the matrix of observed returns x. The empirical covariance matrix C is given by

$$C_{ij} = \frac{1}{T} \sum_t x_{it} x_{jt}$$

and the empirical variance of the portfolio by

$$\hat{\sigma}_p^2 = \frac{1}{T} \sum_{ijt} w_i x_{it} x_{jt} w_j = \frac{1}{T} \sum_{t=1}^{T} \left(\sum_i w_i x_{it} \right)^2 . \tag{3.23}$$

This is to be minimized over the weights w_i subject to the budget constraint $\sum_i w_i = N$.

The rank of the covariance matrix C is the smaller of N and T with probability one. The minimization of $\hat{\sigma}_p^2$ gives us N equations which determine the solution as long as $N \leq T$. When N is larger than T, only T of these equations are independent, so we have more unknowns than equations. For $N \geq T + 1$ any weight vector selected from the null-space of C will be a solution of the minimization problem, with $\hat{\sigma}_p^2 = 0$ as the minimal value of the cost function.

An alternative way to describe the situation is that with N larger than T the cost function will be flat along the directions lying in the null space of the covariance matrix and the solution can run away along these flat directions to an arbitrary distance from the origin. This means that arbitrarily large compensating positive and negative weights can arise, without violating the budget constraint and still keeping the in-sample estimated portfolio variance at zero.

Arbitrarily large leverage combined with a zero value of the risk measure is a prescription for disaster. The first author to point out this dangerous feature of the variance was Jorion [40]. A similar apparent arbitrage in Expected Shortfall and other downside risk measures was analyzed and identified as the root of instability in [21, 24, 41, 42].

It must be clear from the foregoing that this instability has nothing to do with the replica method, or the Gaussian distribution of returns, or the averaging over the samples. The root of this instability is purely geometrical, it arises in every single sample and for any underlying distribution of the returns, and it always takes place at the same critical ratio $r = N/T = 1$. The universality of the critical value $r_c = 1$ of the unconstrained variance optimization was demonstrated in [28] and is a special case of the universality discussed by [43] and [44].

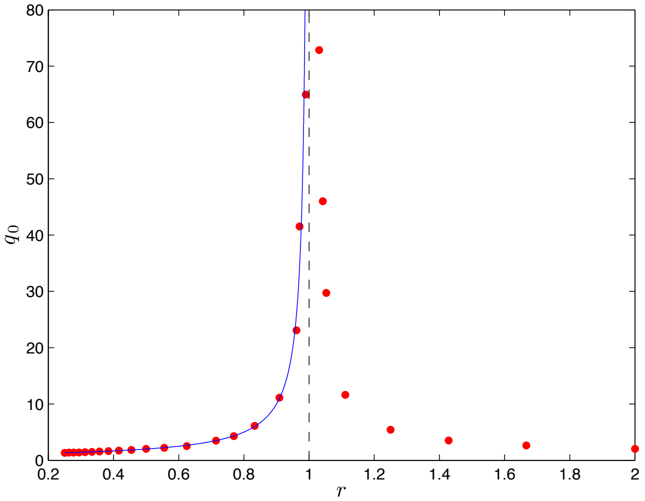

Figure 1: *Estimation error as a function of r. The solid blue line represents the analytical solution obtained with the replica method. The dashed black line indicates the position of the critical point $r_c = 1$. Red dots represent results of numerical simulations averaged over 1000 simulations for a system with $N = 100$. Numerical simulations have been performed with Matlab using the function "fmincon" and the active-set algorithm. The match between numerical and analytical result is very good in the allowed region $r < 1$. Due to a built-in regularizer in the solver, numerical solutions can be found also in the forbidden region. This could create the illusion that it is possible to find reliable solutions to the optimization problem also with very few data points.*

To conclude this subsection, let us point out the significance of this instability for empirical work. Without additional constraints the instability must show up for any empirical sample with $N > T$. To check this, we generated synthetic time series of length T for various values of N. For simplicity, we considered a set of assets with the same variance $\sigma_i = 1$ for all i, and determined the optimal cost and the estimation error q_0 for r values ranging from zero up to 2. The result of this numerical experiment performed with the Matlab solver "fmincon" is shown in Fig. 1. The surprising feature is that after a strong increase on approaching $r = 1$, q_0 starts to decrease above $r = 1$ again, as if the estimator became restabilized. Thus the program produces a stable result even in the region where we know that a continuum of equivalent solutions exist. The resolution of this puzzle lies in the fact that some of the numerically optimized solvers contain what effectively amounts to a regularizer that does not influence the result as long as there is a meaningful one, but kicks in when a singular covariance matrix is encountered, and distributes the solution evenly across the zero modes. Of course, this is properly indicated in the description of the solver, but easily overlooked by the user. This should be a warning to users against the blind application of ready-make programs without understanding their details and without a grasp of the main feature of the expected solution already before the numerical study.

The instability of the unconstrained variance has been pointed out several times earlier, and it is also easy to notice in empirical work from the ever-increasing sample fluctuations. This is not the case for the instability of the no-short-constrained variance optimization to which we turn now.

4 Optimization with no short positions

Portfolio optimization is often subject to constraints or an outright ban on short positions. Optimizing the variance under such conditions is a problem in quadratic programming that is routinely solved numerically. In this section we give what we believe to be the first analytic treatment of portfolio optimization with no short positions allowed.

The starting point is (B.1) and (B.2). If we want to exclude negative weights, we impose infinite penalty on them by letting $\eta_2 \to \infty$ in (B.2). Positive positions will not be penalized, so we set $\eta_1 = 0$. According to (B.9)–(B.16), this leads to the free energy and stationarity conditions as follows:

$$f = \lambda - \Delta\hat{q}_0 - \hat{\Delta}q_0 + \frac{1}{2r}\frac{q_0}{1+\Delta} + \frac{\hat{q}_0}{\hat{\Delta}}\frac{1}{N}\sum_i W\left(\frac{\lambda}{\sigma_i\sqrt{-2\hat{q}_0}}\right) \tag{4.1}$$

$$\frac{1}{\sqrt{q_0 r}} = \frac{1}{N}\sum_i \frac{1}{\sigma_i}\Psi\left(\frac{\lambda}{\sigma_i\sqrt{-2\hat{q}_0}}\right) \tag{4.2}$$

$$\Delta = \frac{1}{2\hat{\Delta}}\frac{1}{N}\sum_i \Phi\left(\frac{\lambda}{\sigma_i\sqrt{-2\hat{q}_0}}\right) \tag{4.3}$$

$$\frac{1}{2r} = \frac{1}{N}\sum_i W\left(\frac{\lambda}{\sigma_i\sqrt{-2\hat{q}_0}}\right) \tag{4.4}$$

and (B.4) and (B.5) remain unchanged:

$$\hat{\Delta} = \frac{1}{2r(1+\Delta)} \tag{4.5}$$

$$\hat{q}_0 = -\frac{q_0}{2r(1+\Delta)^2} \tag{4.6}$$

In (4.2) - (4.4) we used the fact that Φ, Ψ, and W all go to zero as their argument tends to minus infinity.

Using the identity $W(x) = \frac{1}{2}x\Psi(x) + \frac{1}{2}\Phi(x)$ and the stationarity conditions above we can transform (4.4) into

$$\lambda = \frac{q_0}{r(1+\Delta)^2}, \tag{4.7}$$

but by (4.6) this is also equal to

$$\lambda = -2\hat{q}_0. \tag{4.8}$$

Then the arguments of the functions Ψ, Φ and W in (4.2)–(4.4) simplify as $\sqrt{\lambda}/\sigma_i$. (Note that here, as elsewhere in the paper, the choice of the sign of the square root is dictated by the meaning of the quantity in question.). Eq. (4.4) becomes

$$\frac{1}{2r} = \frac{1}{N}\sum_i W\left(\frac{\sqrt{\lambda}}{\sigma_i}\right), \tag{4.9}$$

and (4.3) and (4.5) combine to give

$$\Delta = \frac{r\frac{1}{N}\sum_i \Phi\left(\frac{\sqrt{\lambda}}{\sigma_i}\right)}{1 - r\frac{1}{N}\sum_i \Phi\left(\frac{\sqrt{\lambda}}{\sigma_i}\right)}. \tag{4.10}$$

Finally, for the relative estimation error, which apart form a normalizing factor is the out-of-sample estimator for the optimal value of risk, we find

$$q_0 = \lambda r(1+\Delta)^2. \tag{4.11}$$

Eq. (4.9) is straightforward to solve on a machine to obtain λ as a function of the parameters r, N, and σ_i. Once λ is known, Δ and q_0 can be determined from (4.10) and (4.11). Furthermore, by the help of the stationarity conditions we can derive the expression for the free energy

$$f = \frac{\lambda}{2} \tag{4.12}$$

as in section 3, so the knowledge of λ will also provide the free energy as a function of r, N, and σ_i.

As for the distribution of the optimal estimated weights, by (B.17) and (B.18) we have

$$p(w) = n_0\delta(w) + \theta(w)\frac{1}{N}\sum_i \frac{1}{\sigma_w^{(i)}\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{w - w_0^{(i)}}{\sigma_w^{(i)}}\right)^2\right] \tag{4.13}$$

where θ is the Heaviside function that ensures only non-negative weights appear in the distribution. The first term is the density of the weights set to zero by the no-short constraint:

$$n_0 = \frac{1}{N}\sum_i \Phi\left(-\frac{w_0^{(i)}}{\sigma_w^{(i)}}\right). \tag{4.14}$$

The Gaussian density of the i-th weight is centered at $w_0^{(i)}$, which by (B.11) and (4.7) is equal to

$$w_0^{(i)} = \frac{q_0}{(1+\Delta)}\frac{1}{\sigma_i^2}, \tag{4.15}$$

with standard deviation

$$\sigma_w^{(i)} = \frac{\sqrt{q_0 r}}{\sigma_i}. \tag{4.16}$$

With this we have determined the expected positions of the estimated optimal weights and their distribution, as well as the in-sample estimated cost and the out-of-sample estimator related to the relative estimation error – that is we have solved the optimization of variance with a no-short-position constraint.

The limit $r \to 0$ again corresponds to $\lambda \to \infty$, and it can easily be worked out to recover the results in Subsection 3.1, and Section 2.

4.1 The high-dimensional regime and the critical point at $r = 2$

When r is finite, we are in the high-dimensional regime where N and T are of the same order of magnitude. As W is positive and monotonic increasing, it follows from (4.9) that with r increasing $\sqrt{\lambda}$ must decrease. However, it cannot decrease below zero, and here $W(0) = 1/4$, so r has a maximal value $r_c = 2$ beyond which it cannot grow. It seems therefore that for a given size T of the samples there is an upper bound $N = 2T$ beyond which we cannot consistently continue this theory.

What is happening at $r_c = 2$? First, we realize that because of the proportionality between f and λ, Eq (4.12), f itself also has to vanish at $r = 2$. But f is proportional to the in-sample estimate of the portfolio variance ${\sigma_p^*}^2$, eq (2.7), so f is by definition non-negative and we run into a natural bound at $r = 2$.

Let us now consider the behavior Δ. Expanding (4.9), (4.10) and (4.11) around $r = 2$ we find

$$\Delta = \frac{4}{2-r}, \; r \to 2^-. \tag{4.17}$$

This reveals the meaning of the special value $r = 2$: at this critical value a phase transition is taking place and Δ becomes infinitely large. This transition may seem analogous to the one we found in the unconstrained case, but the critical value of r has been shifted by the no-short constraint to $r_c = 2$ from the unconstrained $r_c = 1$.

There is a further difference: eq. (4.11) tells us that the behavior of q_0 at the phase transition is determined by the limit of $\lambda\Delta^2$ as $r \to 2$. It can be seen that

$$\lim_{r\to 2} q_0 = \lim_{r \to 2} 2\lambda\Delta^2 = \frac{\pi}{\left(\frac{1}{N}\sum_i \frac{1}{\sigma_i}\right)^2} \tag{4.18}$$

which is finite. Therefore, in contrast to the unconstrained phase transition at $r = 1$, the estimation error

$$\tilde{q}_0 = q_0 \frac{1}{N}\sum_i \frac{1}{\sigma_i^2} = \frac{\frac{\pi}{N}\sum_i \frac{1}{\sigma_i^2}}{\left(\frac{1}{N}\sum_i \frac{1}{\sigma_i}\right)^2}, \; r \to 2^- \tag{4.19}$$

remains finite. (Note that $\tilde{q}_0$ is larger or equal to one for any r, as it should, given its meaning as the relative estimation error. In particular, in the limit $r \to 2$ the expression multiplying π in the above formula is larger than equal to one for any distribution of the true variances σ_i, due to the Cauchy inequality.) Thus the phase transition at $r = 2$ displays finite estimation error.

If we picture the portfolio weights as the components of a vector then we can say that the Euclidean norm $\sum_i w_i^2$ of this vector remains finite, but the fluctuations of its direction are infinite. In other words, the longitudinal fluctuations of the weight vector have been reined in by the no-short-selling constraint, however this constraint is unable to suppress the transverse fluctuations. This is rather natural if we consider that the ban on short selling constrains the large compensating positions, but does not forbid the reshuffling of the components of the weight vector from sample to sample.

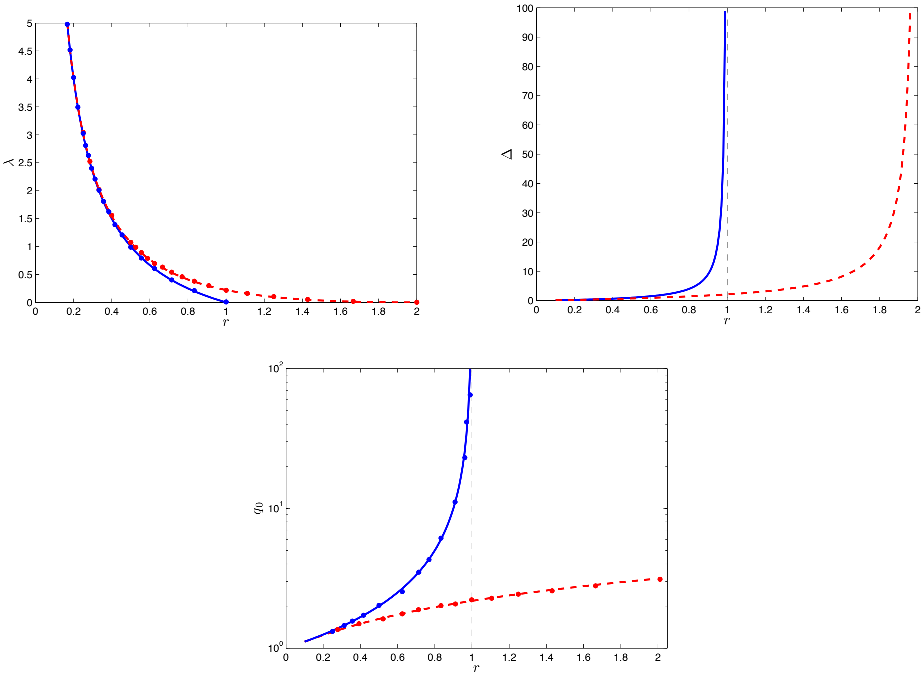

Figure 2: *The three panels show the behavior of λ (top left panel), Δ (top right panel) and q_0 (bottom panel) as a function of r for the cases with (solid lines) and without (dashed lines) short-selling. From the figures it is clear that the no-short selling case displays an instability at $r_c = 2$. This instability is characterized by the divergence of the parameter Δ and the vanishing of λ (proportional to the in-sample estimate of risk), but a finite estimation error. Results of numerical simulations (dots in the figures) are in agreement with the analytical result.*

Fig. 2 compares the results for λ, Δ and q_0 as functions of r for the unconstrained and the no-short-constrained cases, respectively. For simplicity, we show these results for a portfolio with all assets having the same variance $\sigma_i = 1$ for all i.

Let us now consider the distribution of weights when $r \to 2$. Because of the divergence of Δ all the $w_0^{(i)} \to 0$, eq (4.15). This means that the Φ's in the first term all tend to 1/2, so the limiting density of the weights condensed at the origin becomes half of the total weight. At the same time the centers of the Gaussians in the second term will also go to zero, but according to (4.16), their widths remain finite. For a physicist, several features of the phase transition at $r = 2$ may be vaguely reminiscent of Bose condensation.

4.2 Preferential elimination of large volatility assets

The constraint on short positions is a special case of ℓ_1 regularization. As such, it is expected to result in a sparse optimal portfolio, that is to eliminate some of the assets. The build-up of the weight at $w = 0$ is the consequence if this tendency of ℓ_1.

Our results do not refer to a single sample, but to averages over the samples. On average, each asset contributes to the peak of the weight distribution at $w = 0$, i.e. each asset gets eliminated with a certain probability. However, the probability of getting eliminated depends on the variance of the given asset.

The argument of the function Φ in (4.14) is

$$-\frac{w_0^{(i)}}{\sigma_w^{(i)}} = -\left(\frac{q_0}{r}\right)^{1/2} \frac{1}{1+\Delta} \frac{1}{\sigma_i} \tag{4.20}$$

and Φ is monotonic increasing. Accordingly, assets with a large standard deviation (large volatility) become eliminated with larger probability than those with a small volatility. This selection is particularly strong when the coefficient of $1/\sigma_i$ in (4.20) is large, that is r is small, while the distinction between high and low volatility items disappears as we approach $r = 2$, where $\Delta \to \infty$. This is plausible: if we have a lot of information (r small) the regularizer can clearly distinguish between the low and high volatility items, but when fluctuations dominate any possibility of making a difference vanishes.

Note that as $\Phi(0) = 1/2$ and the argument of Φ in (4.14) is always negative, in the limit $r \to 2$ the "condensate density" n_0 approaches its maximal value 1/2 from below: the no-short constraint pushes at most half of the assets into the "condensate". At the same time, because of the divergence of Δ the centers of the Gaussians also shift to the origin, but their standard deviations remain finite.

4.3 The nature of the instability at $r = 2$

The nature of the phase transition taking place at $r = 2$ is somewhat different from the one at $r = 1$. While the latter takes place with probability one even for finite N and T, the transition at $r = 2$ depends on the random samples and in this respect it is a close relative of the transitions in the optimization of the Maximal Loss (or minimax) and the Expected Shortfall risk measures discussed in [35] and [22], respectively. In order to better understand how this instability develops in the large N limit, we performed extended simulations for a large number (going up to 100 000) of finite N and T samples with returns drawn from various symmetric distributions, and determined the number of samples in which the optimal in-sample variance was zero, relative to the total number of simulated samples. In short, we measured the probability of finding zero variance samples. We found that this probability was universal, largely independent of the underlying distribution of returns. Its main features are the following: Below $r = 1$ the probability of finding zero variance samples is identically zero. Between

$r = 1$ and $r = 2$ the probability is small, starting to increase as we approach $r = 2$ and rapidly reaching one for r values exceeding 2. The transition is the faster the larger the dimension and becomes sharp in the limit of high dimensions. In close analogy with the minimax problem, the probability of such a zero variance solution arising is given by the closed formula

$$p(N,T) = \frac{1}{2^{N-1}} \sum_{k=T}^{N-1} \binom{N-1}{k}, \tag{4.21}$$

valid for any continuous and symmetric return distribution. (The condition of continuity is necessary to make the probability of two returns coinciding zero.) As N increases, the transition at $r = 2$ is becoming sharper and sharper, ultimately going over into a step function. Except for the smallest N and T pairs the measured values can be scaled onto the universal curve

$$p(N,T) = \Phi\left(\frac{r - r_c}{r}\sqrt{N}\right). \tag{4.22}$$

which is displayed in Fig. 3.

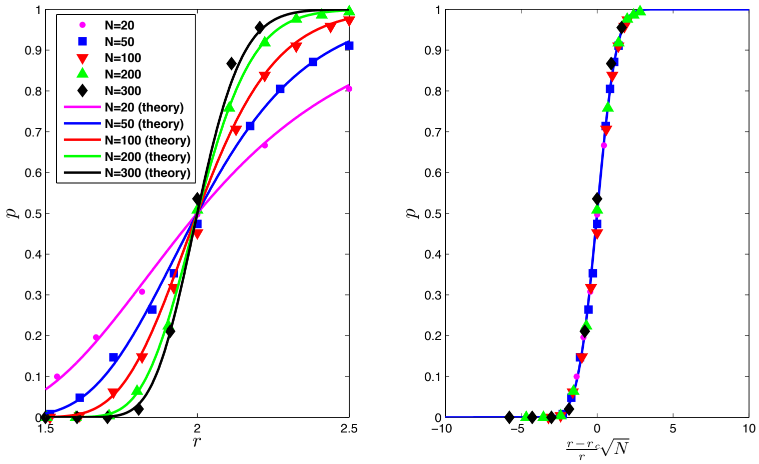

Figure 3: *Left panel: probability of observing a zero variance solution as a function of the ratio $r = N/T$ for different values of N. The solid lines refer to the analytic formula (4.22), while dots refer to numerical simulations. Right panel: collapse of the curves into the same scaling function.*

The zero in-sample variance solutions above $r = 2$ are the natural continuations of the analytic result for the vanishing in-sample estimator at $r = 2$. Thus, in fact, in high dimensions and above the critical ratio $r = 2$, we again have a continuum of solutions to the optimization problem, corresponding to a flat cost landscape for each sample. These solutions are infinitely sensitive to any change in the input data, and jump about the landscape from sample to sample.

Concerning numerical work, it perhaps requires even more care now than around the $r = 1$ phase transition. At variance with that, the instability at $r = 2$ is not accompanied by large fluctuations in the estimated cost, it is more subtle, it corresponds to the fluctuations of the direction of the weight vector. Some of the standard solvers do signal the problem when encountering a singular covariance matrix, others take care of the difficulty by regularizing the problem on their own. It is the obligation of the user to carefully acquaint herself with the details of the solver instead of accepting a seemingly stable answer to a meaningless question.

5 Summary

Let us briefly summarize the main results of this paper. We have considered a portfolio in the high-dimensional limit where the number of different assets N and the sample size T are large, with their ratio $r = N/T$ kept finite. We assumed that the returns on the assets were independent normal variables with zero expected value and different variances. We optimized the variance of the portfolio under the budget constraint with or without a ban on short positions and averaged the results over the random samples by the method of replicas borrowed from the statistical physics of disordered systems.

In the simple case where unlimited short positions were allowed we recovered known results for the out-of-sample estimator and the in-sample average of the portfolio variance. As a new result, we also derived the distribution of optimal weights over the random samples. We found that the originally sharply distinguishable spikes of this distribution broaden with increasing r until in the limit $r \to 1$ any distinction between the different weights gets completely washed away due to the divergent sample fluctuations. In the same limit the estimation error diverges and the in-sample variance of the portfolio vanishes; at $r = 1$ a phase transition is taking place. This is the same point where the first zero eigenvalue of the covariance matrix appears. Beyond this critical value of r the variance cannot be meaningfully optimized: a continuum of solutions appear, since any combination of the zero eigenvectors of the covariance matrix make the variance zero. As argued above, the phenomenon does not depend on the use of the replica method or the assumption about the Gaussian distribution of returns: it is a purely geometric effect, depending solely on the fact that the rank of the covariance matrix is the smaller of N, T in any sample, with probability one.

In order to support and illustrate the theoretical results, we also solved the quadratic programming task of optimizing the variance numerically. While the agreement between the analytic theory and numerics is perfect below the critical point $r = 1$, for $r > 1$ we found that some standard solvers continue to find a stable, unique solution with all the optimal weights the same, the portfolio variance identically zero and the estimation error and susceptibility decreasing with r increasing further. This apparent restabilization is an artifact, due to a built-in stabilizing feature (essentially an ℓ_2 regularizer) in the solvers.

The main result of the paper is the solution of variance optimization under a ban on short positions. This problem, which has a great importance in practice, has not been solved analytically before. The method of replicas allowed us to derive results for the same quantities as in the previous case: we have determined the out-of-sample estimator for the variance, the optimal in-sample variance, and the distribution of optimal portfolio weights again. The constraint on short positions acts as a kind of ℓ_1 regularizer and eliminates some (at most half) of the assets resulting in a sparser portfolio. Accordingly, a sharp peak is built up in the distribution of weights at the origin and the remaining weights are all positive. In agreement with one's natural expectation, assets with larger volatility get eliminated with higher probability than the low volatility items.

It might have been expected that the constraint on short positions would tame the large sample fluctuations. This expectation is borne out only partially: it is true that the optimization can now be performed also above the previous critical value $r = 1$, but at $r = 2$ we discover another phase transition. This time the estimation error stays finite, but another quantity still diverges here. The in-sample estimator for the portfolio variance vanishes, and the distribution of weights is smeared out again.

Numerical work on finite size samples shows that the probability of solutions with zero in-sample variance is zero below $r = 1$, very small between $r = 1$ and $r = 2$, and rapidly goes to one above $r = 2$. Accordingly, we find perfect agreement between the analytic theory and simulations below the $r = 2$ transition already for moderate sized samples, but in the region above $r = 2$, where the instability prevents the analytic theory to penetrate, a continuum of unstable, zero-variance solutions arises, with a flat cost-landscape. To analyze the nature of this transition, we performed numerical work also on small to moderate size samples, and determined the probability law of finding zero-variance solutions, with small probability between $r = 1$ and $r = 2$, and probability one above $r = 2$. We also found a closed, universal formula for this probability, independent of the distribution of returns, as long as it was continuous and symmetric. This guarantees that the transition found by the help of the replica method and Gaussian underlying returns is, in fact, universal, independent of these technicalities.

Concerning the application of solvers from libraries such as R or Matlab, our experience is similar to that around the $r = 1$ transition. Some standard solvers keep finding a stable, unique solution with all the weights the same also above $r = 2$ where we know that a continuum of solutions exist, and the solvers should, in principle, obtain an unstable solution, different in each sample. The explanation of the phenomenon is the same as in the unconstrained case: these solvers are built in such a way as to find the diagonal solution whenever the covariance matrix has zero modes.

The financial content of the instabilities described above is the following. When unlimited short positions are allowed one can assume very large compensating positive and negative positions without violating the budget constraint. As the dimension of the portfolio increases, the (Euclidean) length of the weight vector, hence also the leverage, diverge – a fundamentally risky situation around a point where the estimated portfolio variance vanishes. When we switch on the constraint on short positions, it becomes impossible to build up large compensating positions, and the length of the weight vector, hence also the estimation error, remain finite, but the solution is still unstable with respect to rearrangements, or simply to a reshuffling of the components of the optimal weight vector from sample to sample. This corresponds to divergent transverse fluctuations of the weight vector.

A final remark on the miraculous restabilization of the numerical solutions: in empirical work where one has real life data without the luxury of a large number of samples to average over, one may easily overlook the instability in the no-short-selling case, especially if the software package is a black box for the portfolio manager. We think one should never use a ready-made program without the detailed knowledge of the algorithm implemented in it. Furthermore, one should never trust a purely numerical result without an understanding of the main structural features of the problem, such as the instability described here. Although seldom able to follow it, we agree with Lev Landau's maxim: one should not attempt to solve a problem before knowing the solution in advance.

Appendix A Derivation of the free energy with the replica method

We consider the following problem: given a financial market where N risky assets are traded we want to find the portfolio $\vec{w}$ that minimizes the risk function

$$R(\vec{w}) = \frac{1}{2}\sum_{i,j} w_i C_{ij} w_j, \tag{A.1}$$

under the budget constraint $\sum_{i=1}^{N} w_i = N$. In the above expression w_i represents the position held on asset i, while C_{ij} is the assets covariance matrix. In practice, the true covariance matrix is unknown and one has to rely on estimators based on historical data. If x_{it} represents the return of asset i at time t, the entries of the covariance matrix can be estimated as

$$C_{ij} = \frac{1}{T}\sum_{t=1}^{T} x_{it}x_{jt}. \tag{A.2}$$

Furthermore, we consider adding the following term (an asymmetric ℓ_1 regularizer) to the cost function

$$g(\vec{w}) = \eta_1 \sum_i w_i \theta(w_i) - \eta_2 \sum_i w_i \theta(-w_i), \tag{A.3}$$

so that the optimization problem becomes

$$\min_{\vec{w}} \quad \{\frac{1}{2}\sum_{ij} w_i x_{it} x_{jt} w_j + g(\vec{w})\} \tag{A.4}$$

$$\text{s.t.} \quad \sum_i w_i = N, \tag{A.5}$$

where for later convenience we have multiplied the empirical covariance matrix by a factor T. In the following we assume that the x_{it} are drawn from independent Gaussian distributions of zero mean and variance σ_i^2/N.
Taking advantage of the identity

$$\langle (\log Z)^n \rangle = \Big\langle \frac{Z^n - 1}{n} \Big\rangle, \tag{A.6}$$

valid in the limit $n \to 0$, the typical properties of the solution can be captured by computing the replicated partition function

$$Z_n(\vec{w}) = \Big\langle \int_{-\infty}^{\infty} \prod_{i=1}^{N} \prod_{a=1}^{n} dw_i^a e^{-\gamma\left(\frac{1}{2}\sum_{i,j,t,a} w_i^a x_{it} x_{jt} w_j^a + g(\vec{w})\right)} \prod_a \delta(\sum_i w_i^a - N) \Big\rangle_{\vec{x}_t} \tag{A.7}$$

and then taking the limits

$$\lim_{\gamma\to\infty} \lim_{n\to 0} \frac{1}{\gamma} Z_n(\vec{w}), \tag{A.8}$$

where γ is a fictitious inverse temperature that we introduce to simplify the calculation and $\langle \cdots \rangle$ represents an average over the probability distribution of returns. The above partition function refers to a system of n replicas of the original system, and the index a is introduced to label different replicas, so that w_i^a represents the i-th weight of the a-th replica. Introducing an integral representation for the delta function and performing a Hubbard-Stratonovich transformation the replicated partition function can be written as

$$\begin{aligned} Z_n(\vec{w}) &= \Big\langle \int_{-\infty}^{\infty} \prod_{i,a,t}^{N} dw_i^a d\phi_{at} d\lambda^a \text{exp}\left[-\frac{1}{2}\sum_{a,t}\phi_{at}^2 + i\sqrt{\gamma}\sum_{i,t,a}\phi_t^a w_i^a x_{it} \right] \\ &\times \;\; \text{exp}\left[\sum_a \lambda^a (\sum_i w_i^a - N) - \gamma g(\vec{w}) \right] \Big\rangle_{\vec{x}_t}. \end{aligned}$$

Averaging over the probability distributions of returns gives

$$\begin{aligned} Z_n(\vec{w}) &= \int_{-\infty}^{\infty} \prod_{i,a,b,t} dw_i^a d\hat{Q}_{ab} d\phi_{at} d\lambda^a \exp\left[-\frac{1}{2}\sum_{a,t}\phi_{at}^2 - \frac{\gamma}{2}\sum_{a,b,t}\phi_{at} Q_{ab} \phi_{b,t} \right] \\ &\times \;\; \exp\left[\sum_{a,b} \hat{Q}_{ab} \left(N Q_{ab} - \sum_i \sigma_i^2 w_i^a w_i^b \right) + \sum_a \lambda^a \left(\sum_i w_i^a - N \right) - \gamma g(\vec{w}) \right] \end{aligned}$$

where we have introduced the overlap matrix $Q_{ab} = \frac{1}{N}\sum_i \sigma_i^2 w_i^a w_i^b$ and the conjugate variables $\hat{Q}_{ab}$ to enforce this relation.
We can now integrate over the variables ϕ_{at} to obtain

$$\begin{aligned} Z_n(\vec{w}) &= \int_{-\infty}^{\infty} \prod_{i,a,b,t} dw_i^a d\hat{Q}_{ab} d\lambda^a \exp\left[-\frac{T}{2} \text{tr} \log\left(\delta_{ab} + \gamma Q_{ab}\right) \right] \\ &\times \;\; \exp\left[\sum_{a,b} \hat{Q}_{ab} \left(N Q_{ab} - \sum_i \sigma_i^2 w_i^a w_i^b \right) + \sum_a \lambda^a \left(\sum_i w_i^a - N \right) - \gamma g(\vec{w}) \right] \end{aligned}$$

The convexity of the cost function motivates the choice of the replica symmetric ansatz

$$Q_{ab} = \begin{cases} q_0 + \Delta, & a = b \\ q_0, & a \neq b \end{cases} \tag{A.9}$$

$$\hat{Q}_{ab} = \begin{cases} \hat{q}_0 + \hat{\Delta}, & a = b \\ \hat{q}_0, & a \neq b. \end{cases} \tag{A.10}$$

To leading order in n we have

$$\begin{aligned} -\frac{T}{2}\mathrm{tr}\log(\delta_{ab} + \gamma Q_{ab}) &= -\frac{T}{2}\left[\log\left(1 + \gamma\Delta\right) + \frac{\gamma q_0}{1 + \gamma\Delta}\right] &\text{(A.11)} \\ \sum_{a,b}\hat{Q}_{ab}Q_{ab} &= Nn(\hat{q}_0\Delta + q_0\hat{\Delta} + \Delta\hat{\Delta}), &\text{(A.12)} \end{aligned}$$

while the $\vec{w}$-dependent part of the partition function can be written as

$$\int d\lambda^a d\hat{\Delta} d\hat{q}_0 \exp\left[Nn\left\langle \log\int dw e^{-\hat{\Delta}\sigma^2 w^2 + wz\sigma\sqrt{-2\hat{q}_0} + \lambda w - g(\vec{w})]}\right\rangle_{z\sigma}\right], \tag{A.13}$$

where $\langle\cdots\rangle_{z\sigma}$ denotes averages over the normal variable z and the distribution of asset variances:

$$\langle h(z,\sigma)\rangle_{z\sigma} = \int d\sigma \frac{1}{N}\sum_i \delta(\sigma - \sigma_i)\left(\int_{-\infty}^{\infty}\frac{dz}{\sqrt{2\pi}}h(z,\sigma)e^{-z^2/2}\right). \tag{A.14}$$

If we now write the partition function as

$$Z_n = \int d\lambda dq_0 d\Delta d\hat{q}_0 d\hat{\Delta} e^{-\gamma n N f(\lambda, q_0, \Delta, \hat{q}_0, \hat{\Delta})}, \tag{A.15}$$

we find

$$\begin{aligned} f(\lambda, q_0, \Delta, \hat{q}_0, \hat{\Delta}) &= \frac{1}{2\gamma r}\left[\log(1 + \gamma\Delta) + \frac{\gamma q_0}{1 + \gamma\Delta}\right] + \frac{\lambda}{\gamma} - \frac{1}{\gamma}(\hat{q}_0\Delta + q_0\hat{\Delta} + \Delta\hat{\Delta}) \\ &- \frac{1}{\gamma}\left\langle \log\int dw e^{-\hat{\Delta}\sigma^2 w^2 + wz\sigma\sqrt{-2\hat{q}_0} + \lambda w - g(\vec{w})}\right\rangle_{z\sigma} \end{aligned}$$

Performing the change of variables $\Delta \to \Delta/\gamma$, $\hat{q}_0 \to \gamma^2\hat{q}_0$, $\hat{\Delta} \to \gamma\hat{\Delta}$, $\lambda \to \gamma\lambda$ and taking the limit $\gamma \to \infty$ we finally have

$$f(\lambda, q_0, \Delta, \hat{q}_0, \hat{\Delta}) = \frac{q_0}{2r(1 + \Delta)} - \hat{q}_0\Delta - \hat{\Delta}q_0 + \lambda + \min_{\vec{w}}\left\langle V(\vec{w})\right\rangle_{z\sigma}, \tag{A.16}$$

where

$$V = \hat{\Delta}\sigma^2 w^2 - wz\sigma\sqrt{-2\hat{q}_0} - \lambda w + \eta_1\theta(w) - \eta_2\theta(-w). \tag{A.17}$$

Appendix B The saddle point conditions and the distribution of weights

In Appendix A we derived the free energy functional

$$f(\lambda, q_0, \Delta, \hat{q}_0, \hat{\Delta}) = \frac{q_0}{2r(1 + \Delta)} - \hat{q}_0\Delta - \hat{\Delta}q_0 + \lambda + \min_{\vec{w}}\left\langle V(\vec{w})\right\rangle_{z\sigma}, \tag{B.1}$$

where the "potential" is

$$V = \hat{\Delta}\sigma^2 w^2 - wz\sigma\sqrt{-2\hat{q}_0} - \lambda w + \eta_1\theta(w) - \eta_2\theta(-w). \tag{B.2}$$

The double averaging $\langle \dots \rangle_{\sigma,z}$ means

$$\int_0^\infty d\sigma \frac{1}{N}\sum_i \delta(\sigma - \sigma_i)\int_{-\infty}^\infty \frac{dz}{\sqrt{2\pi}}e^{-z^2/2}\dots \tag{B.3}$$

The potential does not contain q_0 and Δ, therefore the saddle point (or stationarity) conditions can be written up for these variables immediately

$$\frac{\partial f}{\partial q_0} = 0 \Rightarrow \hat{\Delta} = \frac{1}{2r(1+\Delta)}, \tag{B.4}$$

$$\frac{\partial f}{\partial \Delta} = 0 \Rightarrow \hat{q}_0 = -\frac{q_0}{2r(1+\Delta)^2}. \tag{B.5}$$

From these the useful combination

$$\sigma_w = \frac{\sqrt{-2\hat{q}_0}}{2\hat{\Delta}} = \sqrt{q_0 r} \tag{B.6}$$

can be obtained.

Here and in the following we will frequently encounter the integrals of the standard normal distribution:

$$\Phi(x) = \int_{-\infty}^x \frac{dt}{\sqrt{2\pi}}e^{-t^2/2},$$

$$\Psi(x) = \int_{-\infty}^x dt\Phi(t),$$

$$W(x) = \int_{-\infty}^x dt\Psi(t).$$

The minimum of the potential is at

$$w^* = \frac{\sigma z\sqrt{-2\hat{q}_0} + \lambda - \eta_1\theta(w^*) + \eta_2\theta(-w^*)}{2\hat{\Delta}\sigma^2}. \tag{B.7}$$

Substituting this back into (B.2) and performing the double average according to the recipe in (B.3) we find that the last term in (B.1) is

$$\langle V^* \rangle_{z\sigma} = \frac{\hat{q}_0}{\hat{\Delta}}\frac{1}{N}\sum_i \left(W\left(\frac{\lambda - \eta_1}{\sigma_i\sqrt{-2\hat{q}_0}}\right) + W\left(-\frac{\lambda + \eta_2}{\sigma_i\sqrt{-2\hat{q}_0}}\right)\right). \tag{B.8}$$

Then the free energy becomes

$$f = \lambda - \Delta\hat{q}_0 - \hat{\Delta}q_0 + \frac{q_0}{2r(1+\Delta)} + \frac{\hat{q}_0}{\hat{\Delta}}\frac{1}{N}\sum_i \left(W\left(\frac{\lambda - \eta_1}{\sigma_i\sqrt{-2\hat{q}_0}}\right) + W\left(-\frac{\lambda + \eta_2}{\sigma_i\sqrt{-2\hat{q}_0}}\right)\right) \tag{B.9}$$

The remaining three saddle point equations are obtained by taking the derivatives of the above expression with respect to λ, $\hat{\Delta}$ and $\hat{q}_0$ respectively.

$$\frac{\partial f}{\partial \lambda} = 0 \Rightarrow 1 + \frac{\hat{q}_0}{\hat{\Delta}} \frac{1}{N} \sum_i \frac{1}{\sigma_i \sqrt{-2\hat{q}_0}} \left(\Psi\left(\frac{\lambda - \eta_1}{\sigma_i \sqrt{-2\hat{q}_0}} \right) - \Psi\left(- \frac{\lambda + \eta_2}{\sigma_i \sqrt{-2\hat{q}_0}} \right) \right) = 0$$

or, with (B.6),

$$\frac{1}{\sqrt{q_0 r}} = \frac{1}{N} \sum_i \frac{1}{\sigma_i} \left(\Psi\left(\frac{w_1^{(i)}}{\sigma_w^{(i)}} \right) - \Psi\left(- \frac{w_2^{(i)}}{\sigma_w^{(i)}} \right) \right). \tag{B.10}$$

Here the notations

$$w_1^{(i)} = \frac{\lambda - \eta_1}{2\sigma_i^2 \hat{\Delta}} = \frac{(\lambda - \eta_1) r (1 + \Delta)}{\sigma_i^2}, \tag{B.11}$$

$$w_2^{(i)} = \frac{\lambda + \eta_2}{2\sigma_i^2 \hat{\Delta}} = \frac{(\lambda + \eta_2) r (1 + \Delta)}{\sigma_i^2} \tag{B.12}$$

and

$$\sigma_w^{(i)} = \frac{\sigma_w}{\sigma_i} = \frac{\sqrt{q_0 r}}{\sigma_i} \tag{B.13}$$

have been introduced.

$$\frac{\partial f}{\partial \hat{q}_0} = 0 \Rightarrow \Delta = \frac{1}{2\hat{\Delta} N} \sum_i \left(\Phi\left(\frac{w_1^{(i)}}{\sigma_w^{(i)}} \right) + \Phi\left(- \frac{w_2^{(i)}}{\sigma_w^{(i)}} \right) \right). \tag{B.14}$$

where the identity $W(x) = \frac{1}{2} x \Psi(x) + \frac{1}{2} \Phi(x)$ has been used. With (B.4) we can cast (B.14) into the form

$$\Delta = \frac{\frac{r}{N} \sum_i \left(\Phi\left(\frac{w_1^{(i)}}{\sigma_w^{(i)}} \right) + \Phi\left(- \frac{w_2^{(i)}}{\sigma_w^{(i)}} \right) \right)}{1 - \frac{r}{N} \sum_i \left(\Phi\left(\frac{w_1^{(i)}}{\sigma_w^{(i)}} \right) + \Phi\left(- \frac{w_2^{(i)}}{\sigma_w^{(i)}} \right) \right)}. \tag{B.15}$$

Finally

$$\frac{\partial f}{\partial \hat{\Delta}} = 0 \Rightarrow q_0 = - \frac{\hat{q}_0}{\hat{\Delta}^2} \frac{1}{N} \sum_i \left(W\left(\frac{w_1^{(i)}}{\sigma_w^{(i)}} \right) + W\left(- \frac{w_2^{(i)}}{\sigma_w^{(i)}} \right) \right),$$

which can be written by help of (B.4), (B.5) as

$$\frac{1}{2r} = \frac{1}{N} \sum_i \left(W\left(\frac{w_1^{(i)}}{\sigma_w^{(i)}} \right) + W\left(- \frac{w_2^{(i)}}{\sigma_w^{(i)}} \right) \right). \tag{B.16}$$

The distribution of weights can be obtained from

$$p(w) = \langle \delta(w - w^*) \rangle_{z\sigma}$$

and works out to be

$$
\begin{aligned}
p(w) &= \frac{1}{N}\sum_i \left(\Phi\left(\frac{-w_1^{(i)}}{\sigma_w^{(i)}}\right) - \Phi\left(-\frac{w_2^{(i)}}{\sigma_w^{(i)}}\right)\right)\delta(w) \\
&+ \frac{1}{N}\sum_i \frac{1}{\sigma_w^{(i)}\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{w-w_1^{(i)}}{\sigma_w^{(i)}}\right)^2\right)\theta(w) \\
&+ \frac{1}{N}\sum_i \frac{1}{\sigma_w^{(i)}\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{w-w_2^{(i)}}{\sigma_w^{(i)}}\right)^2\right)\theta(-w) \qquad \text{(B.17)}
\end{aligned}
$$

Here, the first term is the density of the zero weights

$$n_0 \equiv \frac{1}{N}\sum_i \left(\Phi\left(\frac{w_2^{(i)}}{\sigma_w^{(i)}}\right) - \Phi\left(\frac{w_1^{(i)}}{\sigma_w^{(i)}}\right)\right), \qquad \text{(B.18)}$$

and

$$n_0^{(i)} = \frac{1}{N}\left(\Phi\left(\frac{w_2^{(i)}}{\sigma_w^{(i)}}\right) - \Phi\left(\frac{w_1^{(i)}}{\sigma_w^{(i)}}\right)\right), \qquad \text{(B.19)}$$

is the contribution of the i-th asset to this "condensate". The appearance of this term is due to the ℓ_1 regularizer.

The distribution of the non-zero weights is given by the second and third terms of (B.17). This formula reveals the meaning of the symbols introduced in (B.11), (B.12) and (B.13): $w_1^{(i)}$ and $w_2^{(i)}$ are the centers of the two Gaussians in (B.17), while $\sigma_w^{(i)}$ their standard deviation.

Acknowledgements

Valuable discussions with Risi Kondor and István Csabai are greatly appreciated. We are also obliged to Barbara Dömötör for calling Jorion's paper to our attention, to A. Eggington for reading the manuscript, and to Joël Bun for pointing out an essential error in a preliminary version of this paper. F.C. acknowledges support of the Economic and Social Research Council (ESRC) in funding the Systemic Risk Centre (ES/K002309/1).

References

[1] B. Scherer and R. D. Martin. *Introduction to Modern Portfolio Optimization With NUOPT and S-PLUS*. Springer, 2005.

[2] J. D. Jobson and B. Korkie. Improved estimation for Markowitz portfolios using James-Stein type estimators. *Proceedings of the American Statistical Association (Business and Economic Statistics)*, 1:279–284, 1979.

[3] P. Jorion. Bayes-stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis*, 21:279–292, 1986.

[4] R. Jagannathan and T. Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 58:1651–1684, 2003.

[5] 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, 2003.

[6] O. Ledoit and M. Wolf. Honey, I shrunk the sample covariance matrix. *J. Portfolio Management*, 31:110, 2004.

[7] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, 88:365–411, 2004.

[8] A. Kempf and C. Memmel. Estimating the global minimum variance portfolio. *Schmalenbach Business Review*, 58:332–348, 2006.

[9] Y. Okhrin and W. Schmid. Distributional properties of portfolio weights. *Journal of Econometrics*, 134:235 – 256, 2006.

[10] V. Golosnoy and Y. Okhrin. Multivariate shrinkage for optimal portfolio weights. *The European Journal of Finance*, 13:441–458, 2007.

[11] G. Frahm. Linear Statistical Inference for Global and Local Minimum Variance Portfolios. *Statistical Papers*, 2008. DOI: 10.1007/s00362-008-0170-z.

[12] G. K. Basak, R. Jagannathan, and T. Ma. A jackknife estimator for tracking error variance of optimal portfolios constructed using estimated inputs. *Management Science*, 55(6):990–1002, 2009.

[13] V. DeMiguel, L. Garlappi, F. J. Nogales, and R. Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55:798–812, 2009.

[14] V. DeMiguel, L. Garlappi, and R. Uppal. Optimal versus naive diversification: how efficient is the 1/n portfolio strategy? *Review of Financial Studies*, 22(22):1915–1953, 2009.

[15] G. Frahm and C. Memmel. Dominating estimators for minimum-variance portfolios. *Journal of Econometrics*, 159(2):289–302, 2010.

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

[17] O. Ledoit and M. Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Institute for Empirical Research in Economics University of Zurich Working Paper*, (515), 2011.

[18] J. Bun, J-P. Bouchaud, and M. Potters. My beautiful laundrette: Cleaning correlation matrices for portfolio optimization. *available at https://www.researchgate.net/publication/302339055*, 2016.

[19] S. Ciliberti, I. Kondor, and M. Mézard. On the feasibility of portfolio optimization under expected shortfall. *Quantitative Finance*, 7:389–396, 2007.

[20] S. Ciliberti and M. Mézard. Risk minimization through portfolio replication. *Eur. Phys. J.*, B 57:175–180, 2007.

[21] F. Caccioli, S. Still, M. Marsili, and I. Kondor. Optimal liquidation strategies regularize portfolio selection. *The European Journal of Finance*, 19(6):554–571, 2013.

[22] F. Caccioli, I. Kondor, and G. Papp. Portfolio optimization under expected shortfall: contour maps of estimation error. *arXiv preprint arXiv:1510.04943*, 2015.

[23] I. Kondor, F. Caccioli, G. Papp, and M. Marsili. Contour map of estimation error for expected shortfall. *Available at http://ssrn.com/abstract=2567876 and http://arxiv.org/abs/1502.0621*, 2015.

[24] F. Caccioli, I. Kondor, M. Marsili, and S. Still. Liquidity risk and instabilities in portfolio optimization. *International Journal of Theoretical and Applied Finance*, 19(05):1650035, 2016.

[25] G. Papp, F. Caccioli, and I. Kondor. Variance-bias trade-off in portfolio optimization under expected shortfall with ℓ_2 regularization. *available at http:// arXiv:1602.08297v1 [q-fin.PM]*, 2016.

[26] T. Shinzato. Replica analysis for the duality of the portfolio optimization problem. *Phys. Rev. E*, 94:052307, 2016.

[27] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin glass theory and beyond.* World Scientific Lecture Notes in Physics Vol. 9, World Scientific, Singapore, 1987.

[28] Istvan Varga-Haszonits, Fabio Caccioli, and Imre Kondor. Replica approach to mean-variance portfolio optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(12):123404, 2016.

[29] H. Markowitz. Portfolio selection. *Journal of Finance*, 7:77–91, 1952.

[30] Takashi Shinzato. Minimal investment risk of a portfolio optimization problem with budget and investment concentration constraints. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(2):023301, 2017.

[31] J.-P. Bouchaud and M. Potters. *Theory of financial risk and derivative pricing.* Cambridge Univ. Press, 2003.

[32] A. Gábor and I. Kondor. Portfolios with nonlinear constraints and spin glasses. *Physica A: Statistical Mechanics and its Applications*, 274(1):222–228, 1999.

[33] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning, data mining, inference, and prediction. Second edition.* Springer series in statistics Springer, Berlin, 2008.

[34] J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Loris. Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Science*, 106(30):12267–12272, 2009.

[35] 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.

[36] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media, 2011.

[37] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[38] Candès, E. J. and Romberg, J. K. and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[39] I. Varga-Haszonits and I. Kondor. Noise sensitivity of portfolio selection in constant conditional correlation GARCH models. *Physica*, A385:307–318, 2007.

[40] P. Jorion. Portfolio optimization in practice. *Financial Analysts Journal*, 48(1):68–74, 1992.

[41] Imre Kondor and István Varga-Haszonits. Instability of portfolio optimization under coherent risk measures. *Advances in Complex Systems*, 13(03):425–437, 2010.

[42] Istvan Varga-Haszonits and Imre Kondor. The instability of downside risk measures. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(12):P12007, 2008.

[43] D. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of The Royal Society A, Mathematical Physical and Engineering Sciences*, 367:4273–93, 2009.

[44] D. Amelunxen, M. Lotz, M. B. McCoy, and Joel A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. *Inform. Inference*, 3(3):224–294, 2013.