2

Sparse Markowitz Portfolios

Christine De Mol
Université Libre de Bruxelles, Belgium

2.1 Markowitz Portfolios

Modern portfolio theory originated from the work of Markowitz (1952), who insisted on the fact that returns should be balanced with risk and established the theoretical basis for portfolio optimization according to this principle. The portfolios are to be composed from a universe of N securities with returns at time t given by $r_{i,t}$, $i = 1, \ldots, N$, and assumed to be stationary. We denote by $\mathbf{E}[\mathbf{r}_t] = \boldsymbol{\mu}$ the $N \times 1$ vector of the expected returns of the different assets, and by $\mathbf{E}[(\mathbf{r}_t - \boldsymbol{\mu})(\mathbf{r}_t - \boldsymbol{\mu})^\top] = \boldsymbol{C}$ the covariance matrix of the returns ($\boldsymbol{\mu}^\top$ is the transpose of $\boldsymbol{\mu}$).

A portfolio is characterized by a $N \times 1$ vector of weights $\mathbf{w} = (w_1, \ldots, w_N)^\top$, where w_i is the amount of capital to be invested in asset number i. Traditionally, it is assumed that a fixed capital, normalized to one, is available and should be fully invested. Hence the weights are required to sum to one: $\sum_{i=1}^{N} w_i = 1$, or else $\mathbf{w}^\top \mathbf{1}_N = 1$, where $\mathbf{1}_N$ denotes the $N \times 1$ vector with all entries equal to 1. For a given portfolio $\mathbf{w}$, the expected return is then equal to $\mathbf{w}^\top \boldsymbol{\mu}$, whereas its variance, which serves as a measure of risk, is given by $\mathbf{w}^\top \boldsymbol{C} \mathbf{w}$. Following Markowitz, the standard paradigm in portfolio optimization is to find a portfolio that has minimal variance for a given expected return $\rho = \mathbf{w}^\top \boldsymbol{\mu}$. More precisely, one seeks $\mathbf{w}_*$ such that:

$$\begin{aligned} \mathbf{w}_* = \arg \min_{\mathbf{w}} \mathbf{w}^\top \boldsymbol{C} \mathbf{w} \\ \text{s. t. } \mathbf{w}^\top \boldsymbol{\mu} = \rho \\ \mathbf{w}^\top \mathbf{1}_N = 1. \end{aligned} \tag{2.1}$$

The constraint that the weights should sum to one can be dropped when including also in the portfolio a risk-free asset, with fixed return r_0, in which one invests a fraction w_0 of the unit capital, so that

$$w_0 + \mathbf{w}^\top \mathbf{1}_N = 1. \tag{2.2}$$

Financial Signal Processing and Machine Learning, First Edition.
Edited by Ali N. Akansu, Sanjeev R. Kulkarni and Dmitry Malioutov.

The return of the combined portfolio is then given by

$$w_0 r_0 + \mathbf{w}^\top \mathbf{r}_t = r_0 + \mathbf{w}^\top (\mathbf{r}_t - r_0 \mathbf{1}_N). \tag{2.3}$$

Hence we can reason in terms of "excess return" of this portfolio, which is given by $\mathbf{w}^\top \widetilde{\mathbf{r}}_t$ where the "excess returns" are defined as $\widetilde{\mathbf{r}}_t = \mathbf{r}_t - r_0 \mathbf{1}_N$. The "excess expected returns" are then $\widetilde{\boldsymbol{\mu}} = \mathbf{E}[\widetilde{\mathbf{r}}_t] = \mathbf{E}[\mathbf{r}_t] - r_0 \mathbf{1}_N = \boldsymbol{\mu} - r_0 \mathbf{1}_N$. The Markowitz optimal portfolio weights in this setting are solving

$$\begin{aligned} \widetilde{\mathbf{w}}_* = \arg \min_{\mathbf{w}} \mathbf{w}^\top \boldsymbol{C} \mathbf{w} \\ \text{s. t. } \mathbf{w}^\top \widetilde{\boldsymbol{\mu}} = \widetilde{\rho} \end{aligned} \tag{2.4}$$

with the same covariance matrix as in (2.1) since the return of the risk-free asset is purely deterministic instead of stochastic. The weight corresponding to the risk-free asset is adjusted as $\widetilde{\omega}_{*,0} = 1 - \widetilde{\mathbf{w}}_*^\top \mathbf{1}_N$ (and is not included in the weight vector $\widetilde{\mathbf{w}}_*$). Introducing a Lagrange parameter and fixing it in order to satisfy the linear constraint, one easily sees that

$$\widetilde{\mathbf{w}}_* = \frac{\widetilde{\rho}}{\widetilde{\boldsymbol{\mu}}^\top \boldsymbol{C}^{-1} \widetilde{\boldsymbol{\mu}}} \ \boldsymbol{C}^{-1} \widetilde{\boldsymbol{\mu}} \tag{2.5}$$

assuming that $\boldsymbol{C}$ is strictly positive definite so that its inverse exists. This means that, whatever the value of the excess target return $\widetilde{\rho}$, the weights of the optimal portfolio are proportional to $\boldsymbol{C}^{-1} \widetilde{\boldsymbol{\mu}}$. The corresponding variance is given by

$$\widetilde{\sigma}^2 = \widetilde{\mathbf{w}}_*^\top \boldsymbol{C} \widetilde{\mathbf{w}}_* = \frac{\widetilde{\rho}^2}{\widetilde{\boldsymbol{\mu}}^\top \boldsymbol{C}^{-1} \widetilde{\boldsymbol{\mu}}} \tag{2.6}$$

which implies that, when varying $\widetilde{\rho}$, the optimal portfolios lie on a straight line in the plane $(\widetilde{\sigma}, \widetilde{\rho})$, called the *capital market line* or *efficient frontier*, the slope of which is referred to as the Sharpe ratio:

$$S = \frac{\widetilde{\rho}}{\widetilde{\sigma}} = \sqrt{\widetilde{\boldsymbol{\mu}}^\top \boldsymbol{C}^{-1} \widetilde{\boldsymbol{\mu}}}. \tag{2.7}$$

We also see that all efficient portfolios (i.e, those lying on the efficient frontier) can be obtained by combining linearly the portfolio containing only the risk-free asset, with weight $\widetilde{\omega}_{*,0} = 1$, and any other efficient portfolio, with weights $\widetilde{\mathbf{w}}_*$. The weights of the efficient portfolio, which contains only risky assets, are then derived by renormalization as $\widetilde{\mathbf{w}}_* / \widetilde{\mathbf{w}}_*^\top \mathbf{1}_N$, with of course $\widetilde{\omega}_{*,0} = 0$. This phenomenon is often referred to as Tobin's two-fund separation theorem. The portfolios on the frontier to the right of this last portfolio require a short position on the risk-free asset $\widetilde{\omega}_{*,0} < 0$, meaning that money is borrowed at the risk-free rate to buy risky assets.

Notice that in the absence of a risk-free asset, the efficient frontier composed by the optimal portfolios satisfying (2.1), with weights required to sum to one, is slightly more complicated: it is a parabola in the variance – return plane (σ^2, ρ) that becomes a "Markowitz bullet" in the plane (σ, ρ). By introducing two Lagrange parameters for the two linear constraints, one can derive the expression of the optimal weights, which are a linear combination of $\boldsymbol{C}^{-1} \boldsymbol{\mu}$ and $\boldsymbol{C}^{-1} \mathbf{1}_N$, generalizing Tobin's theorem in the sense that any portfolio on the efficient frontier can be expressed as a linear combination of two arbitrary ones on the same frontier.

The Markowitz portfolio optimization problem can also be reformulated as a regression problem, as noted by Brodie *et al*. (2009). Indeed, we have $\boldsymbol{C} = \mathbf{E}[\mathbf{r}_t \mathbf{r}_t^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$, so that the minimization problem (2.1) is equivalent to

$$\begin{aligned} \mathbf{w}_* = \arg\min_{\mathbf{w}} \mathbf{E}[|\rho - \mathbf{w}^\top \mathbf{r}_t|^2] \\ \text{s. t. } \mathbf{w}^\top \boldsymbol{\mu} = \rho \\ \mathbf{w}^\top \mathbf{1}_N = 1. \end{aligned} \tag{2.8}$$

Let us remark that when using excess returns, there is no need to implement the constraints since the minimization of $\mathbf{E}[|\widetilde{\rho} - \widetilde{\mathbf{w}}^\top \widetilde{\mathbf{r}}_t|^2]$ (for any constant $\widetilde{\rho}$) is easily shown to deliver weights proportional to $\boldsymbol{C}^{-1}\widetilde{\boldsymbol{\mu}}$, which by renormalization correspond to a portfolio on the capital market line.

In practice, for empirical implementations, one needs to estimate the returns as well as the covariance matrix and to plug in the resulting estimates in all the expressions above. Usually, expectations are replaced by sample averages (i.e., for the returns by $\widehat{\boldsymbol{\mu}} = \frac{1}{T}\sum_{t=1}^T \mathbf{r}_t$ and for the covariance matrix by $\widehat{\boldsymbol{C}} = \frac{1}{T}\sum_{t=1}^T [\mathbf{r}_t \mathbf{r}_t^\top] - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^\top$).

For the regression formulation, we define $\boldsymbol{R}$ to be the $T \times N$ matrix of which row t is given by $\mathbf{r}_t^\top$, namely $\boldsymbol{R}_{t,i} = (\mathbf{r}_t)_i = r_{i,t}$. The optimization problem (2.8) is then replaced by

$$\begin{aligned} \widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{T}\|\rho \mathbf{1}_T - \boldsymbol{R}\mathbf{w}\|_2^2 \\ \text{s. t. } \mathbf{w}^\top \widehat{\boldsymbol{\mu}} = \rho \\ \mathbf{w}^\top \mathbf{1}_N = 1, \end{aligned} \tag{2.9}$$

where $\|\mathbf{a}\|_2^2$ denotes the squared Euclidean norm $\sum_{t=1}^T \mathbf{a}_t^2$ of the vector $\mathbf{a}$ in $\mathbb{R}^T$.

There are many possible variations in the formulation of the Markowitz portfolio optimization problem, but they are not essential for the message we want to convey. Moreover, although lots of papers in the literature on portfolio theory have explored other risk measures, for example more robust ones, we will only consider here the traditional framework where risk is measured by the variance. For a broader picture, see for example the books by Campbell *et al*. (1997) and Ruppert (2004).

2.2 Portfolio Optimization as an Inverse Problem: The Need for Regularization

Despite its elegance, it is well known that the Markowitz theory has to face several difficulties when implemented in practice, as soon as the number of assets N in the portfolio gets large. There has been extensive effort in recent years to explain the origin of such difficulties and to propose remedies. Interestingly, DeMiguel *et al*. (2009a) have assessed several optimization procedures proposed in the literature and shown that, surprisingly, they do not clearly outperform the "naive" (also called "Talmudic") strategy, which consists in attributing equal weights, namely $1/N$, to all assets in the portfolio. The fact that this naive strategy is hard to beat—and therefore constitutes a tough benchmark – is sometimes referred to as the $1/N$ *puzzle*.

A natural explanation for these difficulties comes in mind when noticing, as done by Brodie *et al*. (2009), that the determination of the optimal weights solving problem (2.1) or (2.4) can be viewed as an inverse problem, requiring the inversion of the covariance matrix $\boldsymbol{C}$ or, in practice, of its estimate $\widehat{\boldsymbol{C}}$. In the presence of collinearity between the returns, this matrix is most likely to be "ill-conditioned." The same is true for the regression formulation (2.9) where it is the matrix $\boldsymbol{R}^\top \boldsymbol{R}$ which has to be inverted. Let us recall that the condition number of a matrix is defined as the ratio of the largest to the smallest of its singular values (or eigenvalues when it is symmetric). If this ratio is small, the matrix can be easily inverted, and the corresponding weights can be computed numerically in a stable way. However, when the condition number gets large, the usual numerical inversion procedures will deliver unstable results, due to the amplification of small errors (e.g., rounding errors would be enough) in the eigendirections correponding to the smallest singular or eigenvalues. Since, typically, asset returns tend to be highly correlated, the condition number will be large, leading to numerically unstable, hence unreliable, estimates of the weight vector $\mathbf{w}$. As a consequence, some of the computed weights can take very large values, including large negative values corresponding to short positions.

Contrary to what is often claimed in the literature, let us stress the fact that improving the estimation of the returns and of the covariance matrix will not really solve the problem. Indeed, in inverting a true (population) but large covariance matrix, we would have to face the same kind of ill-conditioning as with empirical estimates, except for very special models such as the identity matrix or a well-conditioned diagonal matrix. Such models, however, cannot be expected to be very realistic.

A standard way to deal with inverse problems in the presence of ill-conditioning of the matrix to be inverted is provided by so-called regularization methods. The idea is to include additional constraints on the solution of the inverse problem (here, the weight vector) that will prevent the error amplification due to ill-conditioning and hence allow one to obtain meaningful, stable estimates of the weights. These constraints are expected, as far as possible, to represent prior knowledge about the solution of the problem under consideration. Alternatively, one can add a penalty to the objective function. It is this strategy that we will adopt here, noticing that most often, equivalence results with a constrained formulation can be established as long as we deal with convex optimization problems. For more details about regularization techniques for inverse problems, we refer to the book by Bertero and Boccacci (1998).

A classical procedure for stabilizing least-squares problems is to use a quadratic penalty, the simplest instance being the squared ℓ_2 norm of the weight vector: $\|\mathbf{w}\|_2^2 = \sum_{i=1}^N |\mathbf{w}_i|^2$. It goes under the name of Tikhonov regularization in inverse problem theory and of ridge regression in statistics. Such a penalty can be added to regularize any of the optimization problems considered in Section 2.1. For example, using a risk-free asset, let us consider problem (2.4) and replace it by

$$\widetilde{\mathbf{w}}_{ridge} = \arg\min_{\mathbf{w}} [\mathbf{w}^\top \boldsymbol{C}\mathbf{w} + \lambda \|\mathbf{w}\|_2^2] \quad \text{s. t. } \mathbf{w}^\top \widetilde{\boldsymbol{\mu}} = \widetilde{\rho} \tag{2.10}$$

where λ is a positive parameter, called the regularization parameter, allowing one to tune the balance between the variance term and the penalty. Using a Lagrange parameter and fixing its value to satisfy the linear constraint, we get the explicit solution

$$\widetilde{\mathbf{w}}_{ridge} = \frac{\widetilde{\rho}}{\widetilde{\boldsymbol{\mu}}^\top (\boldsymbol{C} + \lambda \boldsymbol{I})^{-1} \widetilde{\boldsymbol{\mu}}} (\boldsymbol{C} + \lambda \boldsymbol{I})^{-1} \widetilde{\boldsymbol{\mu}} \tag{2.11}$$

where $\boldsymbol{I}$ denotes the $N \times N$ identity matrix. Hence, the weights of the "ridge" optimal portfolio are proportional to $(\boldsymbol{C} + \lambda \boldsymbol{I})^{-1}\widetilde{\boldsymbol{\mu}}$, whatever the value of the excess target return $\widetilde{\rho}$. The corresponding variance is given by

$$\widetilde{\sigma}^2 = \widetilde{\mathbf{w}}_{ridge}^{\top} \boldsymbol{C}\, \widetilde{\mathbf{w}}_{ridge} = \frac{\widetilde{\rho}^2}{(\widetilde{\boldsymbol{\mu}}^{\top}(\boldsymbol{C} + \lambda \boldsymbol{I})^{-1}\widetilde{\boldsymbol{\mu}})^2}\, \widetilde{\boldsymbol{\mu}}^{\top}(\boldsymbol{C} + \lambda \boldsymbol{I})^{-1}\boldsymbol{C}(\boldsymbol{C} + \lambda \boldsymbol{I})^{-1}\widetilde{\boldsymbol{\mu}} \tag{2.12}$$

which implies that, when λ is fixed, $\widetilde{\sigma}$ is again proportional to $\widetilde{\rho}$ and that the efficient ridge portfolios also lie on a straight line in the plane $(\widetilde{\sigma}, \widetilde{\rho})$, generalizing Tobin's theorem to this setting. Notice that its slope, the Sharpe ratio, does depend on the value of the regularization parameter λ.

Another standard regularization procedure, called *truncated singular value decomposition*, (TSVD), consists of diagonalizing the covariance matrix and using for the inversion only the subspace spanned by the eigenvectors corresponding to the largest eigenvalues (e.g., the K largest). This is also referred to as reduced-rank or principal-components regression. and it corresponds to replacing in the formulas (2.11, 2.12) the regularized inverse $(\boldsymbol{C} + \lambda \boldsymbol{I})^{-1}$ by $\boldsymbol{V}_K \boldsymbol{D}_K^{-1} \boldsymbol{V}_K^{\top}$, where $\boldsymbol{D}_K$ is the diagonal matrix containing the K largest eigenvalues d_k^2 of $\boldsymbol{C}$ and $\boldsymbol{V}_K$ is the $N \times K$ matrix containing the corresponding orthonormalized eigenvectors. Whereas this method implements a sharp (binary) cutoff on the eigenvalue spectrum of the covariance matrix, notice that ridge regression involves instead a smoother filtering of this spectrum where the eigenvalues d_k^2 (positive since $\boldsymbol{C}$ is positive definite) are replaced by $d_k^2 + \lambda$ or, equivalently, in the inversion process, $1/d_k^2$ is replaced by $1/(d_k^2 + \lambda) = \phi_\lambda(d_k^2)/d_k^2$, where $\phi_\lambda(d_k^2) = d_k^2/(d_k^2 + \lambda)$ is a filtering, attenuation, or "shrinkage" factor, comprised between 0 and 1, allowing one to control the instabilities generated by division by the smallest eigenvalues. More general types of filtering factors can be used to regularize the problem. We refer the reader, for example, to the paper by De Mol *et al*. (2008) for a discussion of the link between principal components and ridge regression in the context of forecasting of high-dimensional time series, and to the paper by Carrasco and Noumon (2012) for a broader analysis of linear regularization methods, including an iterative method called Landweber's iteration, in the context of portfolio theory.

Regularized versions of the problems (2.1) and (2.9) can be defined and solved in a similar way as for (2.4). Tikhonov's regularization method has also been applied to the estimation of the covariance matrix by Park and O'Leary (2010). Let us remark that there are many other methods, proposed in the literature to stabilize the construction of Markowitz portfolios, which can be viewed as a form of explicit or implicit regularization, including Bayesian techniques as used for example in the so-called Black–Litterman model. However, they are usually more complicated, and reviewing them would go beyond the scope of this chapter.

2.3 Sparse Portfolios

As discussed in Section 2.2, regularization methods such as rigde regression or TSVD allow one to define and compute stable weights for Markowitz portfolios. The resulting vector of regularized weights generically has all its entries different from zero, even if there may be a lot of small values. This would oblige the investor to buy a certain amount of each security, which is not necessarily a convenient strategy for small investors. Brodie *et al*. (2009) have proposed to use instead a regularization based on a penalty that enforces sparsity of the weight

vector, namely the presence of many zero entries in that vector, corresponding to assets that will not be included in the portfolio. More precisely, they introduce in the optimization problem, formulated as (2.9), a penalty on the ℓ_1 norm of the vector of weights $\mathbf{w}$, defined by $\|\mathbf{w}\|_1 = \sum_{i=1}^{N} |w_i|$. This problem then becomes

$$\begin{aligned} \mathbf{w}_{sparse} = \arg \min_{\mathbf{w}} [\|\rho \mathbf{1}_T - \boldsymbol{R}\mathbf{w}\|_2^2 + \tau \|\mathbf{w}\|_1] \\ \text{s. t. } \mathbf{w}^\top \widehat{\boldsymbol{\mu}} = \rho \\ \mathbf{w}^\top \mathbf{1}_N = 1, \end{aligned} \tag{2.13}$$

where the regularization parameter is denoted by τ. Note that the factor $1/T$ from (2.9) has been absorbed in the parameter τ. When removing the constraints, a problem of this kind is referred to as lasso regression, after Tibshirani (1996). Lasso, an acronym for least absolute shrinkage and selection operator, helps by reminding that it allows for variable (here, asset) selection since it favors the recovery of sparse vectors $\mathbf{w}$ (i.e., vectors containing many zero entries, the position of which, however, is not known in advance). This sparsifying effect is also widely used nowadays in signal and image processing (see, e.g., the review paper by Chen *et al*. (2001) and the references therein).

As argued by Brodie *et al*. (2009), besides its sparsity-enforcing properties, the ℓ_1-norm penalty offers the advantage of being a good model for the transaction costs incurred to compose the portfolio, costs that are not at all taken into account in the Markowitz original framework. Indeed, these can be assumed to be roughly proportional, for a given asset, to the amount of the transaction, whether buying or short-selling, and hence to the absolute value of the portfolio weight w_i. There may be an additional fixed fee, however, which would then be proportional to the number K of assets to include in the portfolio (i.e., proportional to the cardinality of the portfolio, or the number of its nonzero entries, sometimes also called by abuse of language the ℓ_0 "norm" ($\|\mathbf{w}\|_0$) of the weight vector $\mathbf{w}$). Usually, however, such fees can be neglected. Let us remark, moreover, that implementing a cardinality penalty or constraint would render the portfolio optimization problem very cumbersome (i.e., nonconvex and of combinatorial complexity). It has become a standard practice to use the ℓ_1 norm $\|\mathbf{w}\|_1$ as a "convex relaxation" for $\|\mathbf{w}\|_0$. Under appropriate assumptions, there even exist some theoretical guarantees that both penalties will actually deliver the same answer (see, e.g., the book on compressive sensing by Foucart and Rauhut (2013) and the references therein).

Let us remark that, in problem (2.13), it is actually the amount of "shorting" that is regulated; indeed, because of the constraint that the weights should add to one, the objective function can be rewritten as

$$\|\rho \mathbf{1}_T - \boldsymbol{R}\mathbf{w}\|_2^2 + 2\tau \sum_{i \text{ with } w_i<0} |w_i| + \tau, \tag{2.14}$$

in which the last term, being constant, is of course irrelevant for determining the solution. In this setting, we see that the ℓ_1-norm penalty is equivalent to a penalty on the negative weights (i.e., on the short positions), only. In the limit of very large values of the regularization parameter τ, we get, as a special case, a portfolio with only positive weights (i.e., no short positions). Such no-short optimal portfolios had been considered previously in the financial literature by Jagannathan and Ma (2003) and were known for their good performances, but, surprisingly, their sparse character had gone unnoticed. As shown by Brodie *et al*. (2009), these no-short portfolios, obtained for the largest values of τ, are typically also the sparsest in the family

defined by (2.13). When decreasing τ beyond some point, negative weights start to appear, but the ℓ_1-norm penalty allows one to control their size and to ensure numerical stability of the portfolio weights. The regularizing properties of the ℓ_1-norm penalty (or constraint) for high-dimensional regression problems in the presence of collinearity is well known since the paper by Tibshirani (1996), and the fact that the lasso strategy yields a proper regularization method (as is the quadratic Tikhonov regularization method) even in an infinite-dimensional framework has been established by Daubechies *et al*. (2004). Notice that these results were derived in an unconstrained setting, but the presence of additional linear constraints can only reinforce the regularization effect. A paper by Rosenbaum and Tsybakov (2010) investigates the effect of errors on the matrix of the returns.

Compared to more classical linear regularization techniques (e.g., by means of a ℓ_2-norm penalty), the lasso approach not only presents advantages as described above but also has some drawbacks. A first problem is that the ℓ_1-norm penalty enforces a nonlinear shrinkage of the portfolio weights that renders the determination of the efficient frontier much more difficult than in the unpenalized case or in the case of ridge regression. For any given value of τ, such frontier ought to be computed point by point by solving (2.13) for different values of the target return ρ. Another difficulty is that, though still convex, the optimization problem (2.13) is more challenging and, in particular, does not admit a closed-form solution. There are several possibilities to solve numerically the resulting quadratic program. Brodie *et al*. (2009) used the homotopy method developed by Osborne *et al*. (2000a, 2000b), also known as the least-angle regression (LARS) algorithm by Efron *et al*. (2004). This algorithm proceeds by decreasing the value of τ progressively from very large values, exploiting the fact that the dependence of the optimal weight on τ is piecewise linear. It is very fast if the number of active assets (nonzero weights) is small. Because of the two additional constraints, a modification of this algorithm was devised by Brodie *et al*. (2009) to make it suitable for solving the portfolio optimization problem (2.13). For the technical details, we refer the interested reader to the supplementary appendix of that paper.

2.4 Empirical Validation

The sparse portfolio methodology described in the previous Section 2.3 has been validated by an empirical exercise, the results of which are succinctly described here. For a complete description, we refer the reader to the original paper by Brodie *et al*. (2009).

Sparse portfolios were constructed using two benchmark datasets compiled by Fama and French and available from the site http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. They are ensembles of 48 and 100 portfolios and will be referred to as FF48 and FF100, respectively. The out-of-sample performances of the portfolios constructed by solving (2.13) were assessed and compared to the tough benchmark of the Talmudic or equal-weight portfolios for the same period. Using annualized monthly returns from the FF48 and FF100 datasets, the following simulated investment exercise was performed over a period of 30 years between 1976 and 2006. In June of each year, sparse optimal portfolios were constructed for a wide range of values of the regularization parameter τ in order to get different levels of sparsity, namely portfolios containing different numbers K of active positions. To run the regression, historical data from the preceding 5 years (60 months) were used. At the time of each portfolio construction, the target return, ρ, was set to be the average return achieved by the naive, equal-weight portfolio over the same historical period. Once constructed, the portfolios

were held until June of the next year, and their monthly out-of-sample returns were observed. The same exercise was repeated each year until June 2005. All the observed monthly returns of the portfolios form a time series from which one can compute the average monthly return $\widehat{\rho}$ (over the whole period or a subperiod), the corresponding standard deviation $\widehat{\sigma}$, and the Sharpe ratio $S = \widehat{\rho}/\widehat{\sigma}$. We report some Sharpe ratios obtained when averaging over the whole period 1976–2006. For FF48, the best one was $S = 41$ and was obtained with the no-short portfolio, comprising a number of active assets varying over the years, but typically ranging between 4 and 10. Then, when looking at the performances of sparse portfolios with a given number K of active positions, their Sharpe ratios, lower than for the no-short portfolio, decreased with K, clearly outperforming the equal-weight benchmark (for which $S = 27$) as long as $K \lesssim 25$ but falling below for K larger. For FF100, a different behavior was observed. The Sharpe ratios were maximum and of the order of 40 for a number of active positions K around 30, thus including short positions, whereas $S = 30$ for the no-short portfolio. The sparse portfolios were outperforming the equal-weight benchmark with $S = 28$ as long as $K \lesssim 60$.

In parallel and independently of the paper by Brodie *et al*. (2009), DeMiguel *et al*. (2009b) performed an extensive comparison of the improvement in terms of the Sharpe ratio obtained through various portfolio construction methods, and in particular by imposing constraints on some specific norm of the weight vector, including ℓ_2 and ℓ_1 norms. Subsequent papers confirmed the good performances of the sparse portfolios, also on other and larger datasets and in somewhat different frameworks, such as those by Fan *et al*. (2012), by Gandy and Veraart (2013) and by Henriques and Ortega (2014).

2.5 Variations on the Theme

2.5.1 Portfolio Rebalancing

The empirical exercise described in Section 2.4 is not very realistic in representing the behaviour of a single investor since a sparse portfolio would be constructed from scratch each year. Its aim was rather to assess the validity of the investment strategy, as it would be carried out by different investors using the same methodology in different years.

More realistically, an investor already holding a portfolio with weights $\mathbf{w}$ would like to adjust it to increase its performance. This means that one should look for an adjustment $\Delta\mathbf{w}$, so that the new rebalanced portfolio weights are $\mathbf{w} + \Delta\mathbf{w}$. The incurred transaction costs concern only the adjustment and hence can be modelled by the ℓ_1 norm of the vector $\Delta\mathbf{w}$. This means that we must now solve the following optimization problem:

$$\begin{aligned} \Delta\mathbf{w}_{sparse} = \arg \min_{\Delta\mathbf{w}} [\|\rho\mathbf{1}_T - \boldsymbol{R}(\mathbf{w} + \Delta\mathbf{w})\|_2^2 + \tau\|\Delta\mathbf{w}\|_1] \\ \text{s. t. } \Delta\mathbf{w}^\top \widehat{\boldsymbol{\mu}} = 0 \\ \Delta\mathbf{w}^\top \mathbf{1}_N = 0 \end{aligned}$$

ensuring sparsity in the number of weights to be adjusted and conservation of the total unit capital invested as well as of the target return. The methodology proposed by Brodie *et al*. (2009) can be straightforwardly modified to solve this problem. An empirical exercise on sparse portfolio rebalancing is described by Henriques and Ortega (2014).

2.5.2 Portfolio Replication or Index Tracking

In some circumstances, an investor may want to construct a portfolio that replicates the performances of a given portfolio or of a financial index such as the S&P 500, but is easier to manage, for example because it contains less assets. In such a case, the investor will have at his disposal a time series of index values or global portfolio historical returns, which can be put in a $T \times 1$ column vector $\mathbf{y}$. The time series of historical returns of the assets that he can use to replicate $\mathbf{y}$ will be put in a $T \times N$ matrix $\boldsymbol{R}$, as before. The problem can then be formulated as the minimization of the mean square tracking error augmented by a penalty on the ℓ_1 norm of $\mathbf{w}$, representing the transaction costs and enforcing sparsity:

$$\mathbf{w}_{track} = \arg\min_{\mathbf{w}} [\|\mathbf{y} - \boldsymbol{R}\mathbf{w}\|_2^2 + \tau \|\mathbf{w}\|_1], \quad \text{s. t. } \mathbf{w}^\top \mathbf{1}_N = 1. \tag{2.15}$$

This is a constrained lasso regression that can again be solved by means of the methodology described in Section 2.3. A rebalancing version of this tracking problem could also be implemented.

2.5.3 Other Penalties and Portfolio Norms

A straightforward modification of the previous scheme consists of introducing weights in the ℓ_1 norm used as penalty (i.e. replacing it with):

$$\|\mathbf{w}\|_{1,s} = \sum_{i=1}^{N} s_i |w_i| \tag{2.16}$$

where the positive weights s_i can model either differences in transaction costs or some preferences of the investor. Another extension, considered for example by Daubechies *et al*. (2004) for unconstrained lasso regression, is to use ℓ_p-norm penalties with $1 \leq p \leq 2$, namely of the type

$$\|\mathbf{w}\|_p^p = \sum_{i=1}^{N} |w_i|^p \tag{2.17}$$

yielding as special cases lasso for $p = 1$ or ridge regression for $p = 2$. The use of values of p less than 1 in (2.17) would reinforce the sparsifying effect of the penalty but would render the optimization problem nonconvex and therefore a lot more cumbersome.

A well-known drawback of variable selection methods relying on an ℓ_1-norm penalty or constraint is the instability in selection in the presence of collinearity among the variables. This means that, in the empirical exercise described here, when recomposing each year a new portfolio, the selection will not be stable over time within a group of potentially correlated assets. The same effect has been noted by De Mol *et al*. (2008) when forecasting macroeconomic variables based on a large ensemble of time series. When the goal is forecasting and not variable selection, such effect is not harmful and would not, for example, affect the out-of-sample returns of a portfolio. When stability in the selection matters, however, a possible remedy to this problem is the so-called elastic net strategy proposed by Zou and Hastie (2005) which consists of adding to the ℓ_1-norm penalty a ℓ_2-norm penalty, the role of which

is to enforce democracy in the selection within a group of correlated assets. Since all assets in the group thus tend to be selected, it is clear that, though still sparse, the solution of the scheme using both penalties will in general be less sparse than when using the ℓ_1-norm penalty alone. An application of this strategy to portfolio theory is considered by Li (2014).

Notice that for applying the elastic net strategy as a safeguard against selection instabilities, there is no need to know in advance which are the groups of correlated variables. When the groups are known, one may want to select the complete group composed of variables or assets belonging to some predefined category. A way to achieve this is to use the so-called mixed $\ell_1 - \ell_2$ norm, namely

$$\|\mathbf{w}\|_{1,2} = \sum_j (\sum_l |w_{j,l}|^2)^{1/2} \tag{2.18}$$

where the index j runs over the predefined groups and the index l runs inside each group. Such strategy, called "group lasso" by Yuan and Lin (2006), will sparsify the groups but select all variables within a selected group. For more details about these norms ensuring "structured sparsity" and the related algorithmic aspects, see, for example, the review paper by Bach *et al*. (2012).

2.6 Optimal Forecast Combination

The problem of sparse portfolio construction or replication bears strong similarity with the problem of linearly combining individual forecasts in order to improve reliability and accuracy, as noticed by Conflitti *et al*. (2015). These forecasts can be judgemental (i.e., provided by experts asked in a survey to provide forecasts of some economic variables such as inflation) or else be the output of different quantitative prediction models.

The idea is quite old, dating back to Bates and Granger (1969) and Granger and Ramanathan (1984), and has been extensively discussed in the literature (see, e.g., the review by Clemen 1989 and Timmermann 2006).

The problem can be formulated as follows. We denote by y_{t+h} the variable to be forecast at time t, assuming that the desired forecast horizon is h. We have at hand N forecasters, each delivering at time t a forecast $\hat{y}_{i,t+h}$, using the information about y_t they have at time t. We form with these individual forecasts $\widehat{y}_{i,t+h}, i = 1, \cdots, N$, the $N \times 1$-dimensional vector $\mathbf{y}_{t+h}$. These forecasts are then linearly combined using time-independent weights $w_i, i = 1, \cdots, N$, which are assumed to satisfy the contraints $w_i \geq 0$ and $\sum_{i=1}^N w_i = 1$, and which are put into the $N \times 1$ vector $\mathbf{w}$. The aim is to minimize the mean square forecast error $\mathbf{E}[(y_{t+h} - \mathbf{w}^\top \widehat{\mathbf{y}}_{t+h})^2]$ achieved by the combination. In empirical applications, the expectation is replaced by the sample mean over some historical period for which both the forecasts and the realization of the real variable are available. Hence the optimal forecast combination problem can be formulated as

$$\begin{aligned} \mathbf{w}_{opt} = \arg \min_{\mathbf{w}} & \left[\sum_{t=1}^{T-h} (y_{t+h} - \mathbf{w}^\top \widehat{\mathbf{y}}_{t+h})^2 \right] \\ \text{s. t. } & \mathbf{w} \geq \mathbf{0} \\ & \mathbf{w}^\top \mathbf{1}_N = 1 \end{aligned} \tag{2.19}$$

assuming that the variable y_t is observed for $t = 1, \ldots, T$. The resulting combined forecast for the variable y_t at time $t = T + h$ is then given by $\mathbf{w}_{opt}^\top \widehat{\mathbf{y}}_{T+h}$.

With the vector of forecasts replacing the vector of returns, the problem is analogous to the problem of portfolio tracking described in Section 2.5, but with an additional no-shorting constraint. Besides, since by combining the two constraints we see that the ℓ_1 norm of the weight vector is fixed to one, problem (2.19) is equivalent to

$$\begin{aligned} \mathbf{w}_{opt} = \arg\min_{\mathbf{w}} & \left[\sum_{t=1}^{T-h} (y_{t+h} - \mathbf{w}^\top \hat{\mathbf{y}}_{t+h})^2 + \tau \|\mathbf{w}\|_1 \right] \\ \text{s. t. } & \mathbf{w} \geq \mathbf{0} \\ & \mathbf{w}^\top \mathbf{1}_N = 1 \end{aligned} \tag{2.20}$$

for any value of the regularization parameter τ, which means that the weight vector will be sparse. Hence we have to solve a constrained lasso regression, and the modified LARS algorithm proposed by Brodie *et al*. (2009) can again be used to this purpose. Notice, however, that the sparsity level cannot be tuned by adjusting the value of τ. Possible remedies to this drawback would be to give up the nonnegativity constraints on the weights or else to use exact sparse simplex projections as in the paper by Kyrillidis *et al*. (2013).

An empirical exercise using survey data from the Survey of Professional Forecasters (SPF) for the Euro area and concerning the forecast of inflation and of GDP (Gross Domestic Product) growth is described in detail in the paper by Conflitti *et al*. (2015). The findings are that the optimal combinations of more than 50 individual forecasts perform well compared to the equal-weight combinations currently used by the European Central Bank. Nevertheless, the corresponding gains are relatively modest, which shows that the $1/N$ puzzle applies to this situation as well. The paper by Conflitti *et al* (2015) also addresses the problem of optimally combining density forecasts, in which case the least-squares objective function is replaced by a Kullback–Leibler Information Criterion between densities or by a derived "log-score" criterion.

Acknowlegments

I would like to thank my coauthors of the sparse portfolio paper on which most of the material of this chapter is based, namely Joshua Brodie, Ingrid Daubechies, Domenico Giannone, and Ignace Loris. Useful comments by an anonymous referee are also gratefully acknowledged.

This work was supported by the research contracts ARC-AUWB/2010-15/ULB-11 and IAP P7/06 StUDys.

References

Bach, F., Jenatton, R., Mairal, F. and Obozinski, G. (2012). Structured sparsity through convex optimization. *Statistical Science*, 27, 450–468.

Bates, J.M. and Granger, C.W.J. (1969). The combination of forecasts. *Operations Research Quarterly*, 20, 451–468.

Bertero, M. and Boccacci, P. (1998). *Introduction to inverse problems in imaging*. London: Institute of Physics Publishing.

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

Campbell, J.Y., Lo, A.W. and MacKinlay, C.A. (1997). *The econometrics of financial markets*. Princeton, NJ: Princeton University Press.

Carrasco, M. and Noumon, N. (2012). Optimal portfolio selection using regularization. https://www.webdepot.umontreal.ca/Usagers/carrascm/MonDepotPublic/carrascm/index.htm

Chen, S., Donoho, D. and Saunders, M. (2001). Atomic decomposition by basis pursuit. *SIAM Review*, 43, 129–159.

Clemen, R.T. (1989). Combining economic forecasts: a review and annotated bibliography. *International Journal of Forecasting*, 5, 559–583.

Conflitti, C., De Mol, C. and Giannone, D. (2015). Optimal combination of survey forecasts. *International Journal of Forecasting*, 31, 1096–1103.

Daubechies, I., Defrise, M. and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57, 1416–1457.

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

DeMiguel, V., Garlappi, L., Nogales, F.J. and Uppal, R. (2009b). A generalized approach to portfolio optimization: improving performance by constraining portfolio norms. *Management Science*, 55, 798–812.

De Mol, C., Giannone, D. and Reichlin, L. (2008). Forecasting using a large number of predictors: is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146, 318–328.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32, 407–499.

Fan, J., Zhang, J. and Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of American Statistical Association*, 107, 592–606.

Foucart, S. and Rauhut, H. (2013). *A mathematical introduction to compressive sensing*. Basel: Birkhauser.

Gandy, A. and Veraart, L.A.M. (2013). The effect of estimation in high-dimensional portfolios. *Mathematical Finance*, 23, 531–559.

Granger, C.W.J. and Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3, 197–204.

Henriques, J. and Ortega, J-P. (2014). Construction, management, and performances of Markowitz sparse portfolios. *Studies in Nonlinear Dynamics and Econometrics*, 18, 383–402.

Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: why imposing the wrong constraints helps. *Journal of Finance*, 58, 1651–1684.

Kyrillidis, A., Becker, S., Cevher, V. and Koch, C. (2013). Sparse projections onto the simplex. Proceedings of the 30th International Conference on Machine Learning (ICML 2013). *JMLR W&CP*, 28, 235–243.

Li, J. (2014). Sparse and stable portfolio selection with parameter uncertainty. *Journal of Business and Economic Statistics*, 33, 381–392.

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

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000a). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20, 389–403.

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000b). On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9, 319–337.

Park, S. and O'Leary, D.P. (2010). Portfolio selection using Tikhonov filtering to estimate the covariance matrix. *SIAM Journal on Financial Mathematics*, 1, 932–961.

Rosenbaum, M. and Tsybakov, A.B. (2010). Sparse recovery under matrix uncertainty. *Annals of Statistics*, 38, 2620–2651.

Ruppert, D. (2004). *Statistics and finance: an introduction*. Berlin: Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Timmermann, A. (2006). Forecast combination: *Handbook of economic forecasting*, Vol. 1 (ed. G. Elliott, C. Granger and A. Timmermann). Amsterdam: North Holland.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68, 49–67.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67, 301–320.