Optimal Portfolio Selection using Regularization*

Marine Carrasco and Nérée Noumon
Université de Montréal
Preliminary and incomplete

October 2010

Abstract

The mean-variance principle of Markowitz (1952) for portfolio selection gives disappointing results once the mean and variance are replaced by their sample counterparts. The problem is amplified when the number of assets is large and the sample covariance is singular or nearly singular. In this paper, we investigate four regularization techniques to stabilize the inverse of the covariance matrix: the ridge, spectral cut-off, Landweber-Fridman and LARS Lasso. These four methods involve a tuning parameter that needs to be selected. The main contribution is to derive a data-driven method for selecting the tuning parameter in an optimal way, i.e. in order to minimize a quadratic loss function measuring the distance between the estimated allocation and the optimal one. The cross-validation type criterion takes a similar form for the four regularization methods. Preliminary simulations show that regularizing yields a higher out-of-sample performance than the sample based Markowitz portfolio and often outperforms the 1 over N equal weights portfolio.

*We thank Raymond Kan, Bruce Hansen, and Marc Henry for their helpful comments.

1 Introduction

In his seminal paper of 1952, Markowitz stated that the optimal portfolio selection strategy should be an optimal trade-off between return and risk instead of an expected return maximization only. In his theoretical framework, Markowitz made the important assumption that the belief about the future performance of asset returns are known. However in practice these beliefs have to be estimated. The damaged caused by the so-called parameter uncertainty has been pointed out by many authors, see for instance Kan and Zhou (2007). Solving the mean variance problem leads to estimate the covariance matrix of returns and take its inverse. This results in estimation error, amplified by two facts. First, the number of securities is typically very high and second, these security returns may be highly correlated. This results in a ill-posed problem in the sense that a slight change in portfolio return target implies a huge change in the optimal portfolio weights. To tackle these issues, various solutions have been proposed. Some authors have taken a bayesian approach, see Frost and Savarino (1986). Some have used shrinkage, more precisely Ledoit and Wolf (2003, 2004a,b) propose to replace the covariance matrix by a weighted average of the sample covariance and some structured matrix. Tu and Zhou (2009) take a combination of the naive 1/N portfolio with the Markowitz portfolio. Alternatively, Brodie, Daubechies, De Mol, Giannone and Loris (2008) and Fan, Zhang, and Yu (2009) use a method called Lasso which consists in imposing a constraint on the sum of the absolute values (l_1 norm) of the portfolio weights. This constraint has for consequence to generate a sparse portfolio which degree of sparsity depends on a tuning parameter.

In this paper, we investigate various regularization (or stabilization) techniques borrowed from the literature on inverse problems. Indeed, inverting a covariance matrix can be regarded as solving an inverse problem. Inverse problems are encountered in many fields and have been extensively studied. Here, we will apply the three regularization techniques that are the most used: the ridge which consists in adding a diagonal matrix to the covariance matrix, the spectral cut-off which consists in discarding the eigenvectors associated with the smallest eigenvalues, and Landweber Fridman iterative method. For completeness, we also consider a form of Lasso where we penalize the l_1 norm of the optimal portfolio weights. These various regularization techniques have been used and compared in the context of forecasting macroeconomic time series using a large number of predictors by i.e. Stock and Watson (2002), Bai and Ng (2008), and De Mol, Giannone, and Reichlin (2008). The four methods under consideration involve a regularization (or tuning) parameter that needs to be selected. Little has been said so far on how to choose the tuning parameter to perform optimal portfolio selection. For example using the Lasso, Brodie et al. (2008), Fan et al. (2009) show that by tuning the penalty term one could construct portfolio with desirable sparsity but do not give a systematic rule on how to select it in practice. Ledoit and Wolf (2004) choose the tuning parameter in order to minimize the mean-square error of the shrinkage covariance

matrix, however this approach may not be optimal for portfolio selection.

The main objective of this paper is to derive a data-driven method for selecting the regularization parameter in an optimal way. As the portfolio performance is usually assessed by the Sharpe ratio, we believe that most investors would like to select the tuning parameter in order to minimize the mean square error of the estimated optimal allocation. This would insure that the optimal allocation is evaluated as accurately as possible. The mean square error can not be derived analytically. Our contribution is to provide an estimate of the MSE that uses only the observations. This estimate is a type of generalized cross-validation criterion. The advantage of our criterion is that it applies to all the methods mentioned above and gives a basis to compare the different methods.

The rest of the paper is organized as follows. Section 2 reviews the mean-variance principle. Section 3 describes three regularization techniques of the inverse of the covariance matrix. Section 4 discusses stabilization techniques that take the form of penalized least-squares. Section 5 derives the optimal selection of the tuning parameter. Section 6 presents simulations results and Section 7 empirical results. Section 8 concludes.

2 Markowitz paradigm

Markowitz (1952) proposes the mean-variance rule, which can be viewed as a trade-off between expected return and the variance of the returns. For a survey, see Brandt (2009). Consider N risky assets with random return vector R_{t+1} and a riskfree asset with known return R_t^f. Define the excess returns $r_{t+1} = R_{t+1} - R_t^f$ and denote their conditional means and covariance matrix by μ and Σ, respectively. The investor allocates a fraction x of wealth to risky assets and the remainder to $(1 - 1_N' x)$ to the risk-free asset, where 1_N denotes a $N-$vector of ones. The portfolio return is therefore $x' r_{t+1} + R_t^f$. The mean-variance problem consists in choosing the vector x to minimize the variance of the resulting portfolio $r_{p,t+1} = x' r_{t+1}$ for a pre-determined target expected return of the portfolio μ_p:

$$\begin{array}{l} \min_x \frac{1}{2} Var\left[r_{p,t+1}\right] = x'\Sigma x \\ \text{st } E\left[r_{p,t+1}\right] = x'\mu = \mu_p \end{array} \tag{1}$$

The optimal portfolio is given by

$$x^* = \frac{\mu_p}{\mu'\Sigma^{-1}\mu}\Sigma^{-1}\mu. \tag{2}$$

The combination that maximizes the Sharpe ratio of the overall portfolio defined as $E\left[r_{p,t+1}\right]/std\left[r_{p,t+1}\right]$ is obtained for the so-called tangency portfolio and corresponds to $\mu_p = \mu'\Sigma^{-1}\mu / 1_N'\Sigma^{-1}\mu$. Note that for this portfolio, x^* satisfies

$$x^* = \frac{\Sigma^{-1}\mu}{1_N'\Sigma^{-1}\mu}$$

and the Sharpe ratio takes the simple form $\sqrt{\mu'\Sigma^{-1}\mu}$. We mainly focus on the tangency portfolio in the sequel.

In order to solve the mean variance problem (1), the expected return and the covariance matrix of the vector of security return, which are unknown, need to be estimated from available data set. In particular, an estimate of the inverse of the covariance matrix is needed. The sample covariance often used in practice may be the worst choice because it is typically nearly singular, and sometimes not even invertible. The issue of ill-conditioned covariance matrix must be addressed because inverting such matrix increases dramatically the estimation error and then makes the mean variance solution unreliable. Many regularization techniques can stabilize the inverse. They can be divided into two classes: regularization directly applied to the covariance matrix and regularization expressed as a penalized least-squares.

3 Regularization as approximation to an inverse problem

3.1 Inverse problem

Let r_t, $t = 1, \cdots, T$ be the observations of asset returns and R be the $T \times N$ matrix with tth row given by r_t'. We can replace the unknown expectation μ by the sample average $\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} r_t$ and the covariance Σ by the sample covariance $\hat{\Sigma} = (R - 1_T\hat{\mu}')'(R - 1_T\hat{\mu}')/T \equiv \tilde{R}'\tilde{R}$. Jobson and Korkie (1983) and later on Britten-Jones (1999) showed that the optimal allocation $x^* = \hat{\Sigma}^{-1}\hat{\mu}/1_N'\hat{\Sigma}^{-1}\hat{\mu}$ for the tangency portfolio can be rewritten as

$$\widehat{x^*} = \hat{\beta}/1_N'\hat{\beta}$$

where $\hat{\beta}$ is the OLS estimate of β in the regression

$$1 = \beta' r_{t+1} + u_{t+1}$$

or equivalently

$$1_T = R\beta + u \tag{3}$$

where R is the $T \times N$ matrix with rows composed of r_t'. In other words, it is not necessary to center r_t in the calculation of x^*. Finding β can be thought as finding the minimum least-squares solution to the equation:

$$R\beta = 1_T. \tag{4}$$

It is a typical inverse problem.

The stability of the previous problem depends on the characteristics of the matrix $\hat{\Omega} = R'R$. Two difficulties may occur: the assets could be highly correlated (i.e. the

population covariance matrix Σ is nearly singular) or the number of assets could be too large relative to the sample size (i.e. the sample covariance is (nearly) singular even though the population covariance is not). In such cases, $\hat{\Omega}$ typically has some singular values close to zero resulting in an ill posed problem, such that the optimization of the portfolio becomes a challenge. These difficulties are summarized by the condition number which is the ratio of the maximal and minimal eigenvalue of $\hat{\Omega}$. A large condition number leads to unreliable estimate of the vector of portfolio weights x.

The inverse problem literature, that usually deals with infinite dimensional problems, has proposed various regularization techniques to stabilize the solution to (4). For an overview on inverse problems, we refer the readers to Kress (1999) and Carrasco, Florens, and Renault (2007). We will consider here the three most popular regularization techniques: ridge, spectral cut-off, and Landweber Fridman. Each method will give a different estimate of β, denoted $\hat{\beta}_\tau$ and estimate of x^*, denoted $\widehat{x}_\tau^* = \hat{\beta}_\tau / 1_N' \hat{\beta}_\tau$.

Let $\left(\lambda_j, \phi_j, v_j\right)$, $j = 1, 2, ..., N$ be the singular system of R, i.e. $\left(\lambda_j^2, \phi_j\right)$ denote the eigenvalues and eigenvectors of $R'R$ and $\left(\lambda_j^2, v_j\right)$ are the nonzero eigenvalues and eigenvectors of RR'. Let $\tau > 0$ be a regularization parameter.

3.2 Ridge regularization

It consists in adding a diagonal matrix to $\hat{\Omega}$.

$$\begin{aligned} \hat{\beta}_\tau &= \left(R'R + \tau I\right)^{-1} R' 1_T, \\ \hat{\beta}_\tau &= \sum_{j=1}^{N} \frac{\lambda_j}{\lambda_j^2 + \tau} \left(1_T' v_j\right) \phi_j. \end{aligned} \tag{5}$$

This regularization has a baysesian interpretation, see i.e. De Mol et al (2008).

3.3 Spectral cut-off regularization

This method discards the eigenvectors associated with the smallest eigenvalues.

$$\hat{\beta}_\tau = \sum_{\lambda_j > \tau} \frac{1}{\lambda_j} \left(1_T' v_j\right) \phi_j.$$

Interestingly, v_j are the principal components of $\hat{\Omega}$, so that if r_t follows a factor model, v_1, v_2,... estimate the factors.

3.4 Landweber-Fridman regularization

The solution to (4) can be computed iteratively as

$$\psi_k = \left(I - cR'R\right) \psi_{k-1} + cR' 1_T$$

with $0 < c < 1/\|R\|^2$. Alternatively, we can write

$$\hat{\beta}_\tau = \sum \frac{1}{\lambda_j} \left\{ 1 - \left(1 - c\lambda_j^2\right)^{1/\tau} \right\} \left(1_T' v_j\right) \phi_j.$$

Here, the regularization parameter τ is such that $1/\tau$ represents the number of iterations.

The three methods involve a regularization parameter τ which needs to converge to zero with T at a certain rate for the solution to converge.

3.5 Shrinkage

In this subsection, we compare our methods with a popular alternative called shrinkage. Shrinkage can also be regarded as a form of regularization. Ledoit and Wolf (2004a) propose to estimate the returns covariance matrix by a weighted average of the sample covariance matrix $\hat{\Sigma}$ and an estimator with a lot of structure F, based on a model. The first one is easy to compute and has the advantage to be unbiased. The second one contains relatively little estimation error but tends to be misspecified and can be severely biased. The shrinkage estimator takes the form of a convex linear combination : $\delta F + (1-\delta)\hat{\Sigma}$, where δ is a number between 0 and 1. This method is called shrinkage since the sample covariance matrix is shrunk toward the structured estimator. δ is referred to as the shrinkage constant. With the appropriate shrinkage constant, we can obtain an estimator that performs better than either extreme (invertible and well-conditioned).

Many potential covariance matrices F could be used. Ledoit and Wolf (2004a) suggested the single factor model of Sharpe (1963) which is based on the assumption that stock returns follow the model (Market model):

$$r_{it} = \alpha_i + \beta_i r_{0t} + \varepsilon_{it}$$

where residuals ϵ_{it} are uncorrelated to market returns r_{0t} and to one another, with a constant variance $Var(\epsilon_{it}) = \delta_{ii}$. The resulting covariance matrix is

$$\Phi = \sigma_0^2 \beta\beta' + \Delta$$

Where σ_0^2 is the variance of market returns and $\Delta = diag(\delta_{ii})$. σ_0^2 is consistently estimated by the sample variance of market returns, β by OLS, and δ_{ii} by the residual variance estimate. A consistent estimate of Φ is then

$$F = s_0^2 bb' + D$$

Instead of F derived for a factor model, the constant correlation model[1] (Ledoit and Wolf (2004a)), and identity matrix $F = I$ (Ledoit and Wolf (2004b)) can as well be

[1] All the pairwise covariances are identical.

used. They give comparable results but are easier to compute.
In the particular case where the shrinkage target is the identity matrix, the shrinkage method is equivalent to Ridge regularization since the convex linear combination $\delta I + (1-\delta)\hat{\Sigma}$ can be rewritten :

$$\Sigma_{Shrink} = c\left(\hat{\Sigma}+\alpha I\right),$$

and

$$\Sigma_{Shrink}^{-1} = c\left(\hat{\Sigma}+\alpha I\right)^{-1},$$

where c is a constant.

Once the shrinkage target is determined one has to choose the Optimal Shrinkage intensity δ^*. Ledoit and Wolf (2004b) propose to select δ^* so that it minimizes the expected L^2 distance between the resulting shrinkage estimator $\Sigma_{Shrink} = \hat{\delta}^* F + (1-\hat{\delta}^*)\hat{\Sigma}$ and the true covariance matrix Σ. The limitation of this criterion is that it only focuses on the statistical properties of Σ, and in general could fail to be optimal for the portfolio selection.

4 Regularization scheme as penalized least-square

The traditional optimal Markowitz portfolio x^* is the solution to (1) that can be reformulated by exploiting the relation $\Sigma = E(r_t r_t') - \mu\mu'$ as

$$\begin{aligned} x^* &= \arg\min_x E\left[\left|\mu_p - x' r_t\right|^2\right] \\ &\text{st } x'\mu = \mu_p \end{aligned}$$

If one replaces the expectation by sample average $\hat{\mu}$, the problem becomes:

$$\begin{aligned} x^* &= \arg\min_x \tfrac{1}{T}\left\|\mu_p 1_T - Rx\right\|_2^2 \\ &\text{st } x'\hat{\mu} = \mu_p \end{aligned} \tag{6}$$

As mentioned before, the solution of this problem may be very unreliable if $R'R$ is nearly singular. To avoid having explosive solutions, we can penalize the large values by introducing a penalty term applied to a norm of x. Depending on the norm we choose, we end up with different regularization techniques.

4.1 Bridge method

For $\gamma > 0$ the Bridge estimate is given by

$$\begin{aligned} \widehat{x}_\tau^* &= \arg\min_x \left\|\mu_p 1_T - Rx\right\|_2^2 + \tau\sum_{i=1}^N |x_i|^\gamma \\ &\text{st } x'\hat{\mu} = \mu_p \end{aligned}$$

where τ is the penalty term.

The Bridge method includes two special cases. For $\gamma = 1$ we have the Lasso regularization, while $\gamma = 2$ leads to the Ridge method. The term $\sum_{i=1}^{N} |x_i|^\gamma$ can be interpreted as a transaction cost. It is linear for Lasso, but quadratic for the ridge.

4.2 Least Absolute Shrinkage and Selection Operator (LASSO)

The Lasso regularization technique introduced by Tibshirani (1996) is the l_1-penalized version of the problem (6). The Lasso regularized solution is obtained by solving:

$$\begin{aligned} &\widehat{x}_\tau^* = \arg\min_x \left\|\mu_p 1_T - Rx\right\|_2^2 + \tau \left\|x\right\|_1 \\ &\text{st } x'\hat{\mu} = \mu_p \end{aligned}$$

The main feature of this regularization scheme is that it induces sparsity. It has been studied by Brodie, Daubechies, De Mol, Giannone and Loris (2008) to compute portfolio involving only a small number of securities. For two different penalty constants τ_1 and τ_2 the optimal regularized portfolio satisfies: $(\tau_1 - \tau_2)\left(\left\|x^{[\tau_2]}\right\|_1 - \left\|x^{[\tau_1]}\right\|_1\right) \geq 0$ then the higher the l_1-penalty constant (τ), the sparser the optimal weights. So that a portfolio with non negative entries corresponds to the largest values of τ and thus to the sparsest solution. In particular the same solution can be obtained for all τ greater than some value τ_0.

Brodie et al. consider models without a riskfree asset. Using the fact that all the wealth is invested $(x'1_N = 1)$, they use the equivalent formulation for the objective function as:

$$\left\|\mu_P 1_T - Rx\right\|_2^2 + 2\tau \sum_{i\ with\ x_i<0} |x_i| + \tau$$

which is equivalent to a penalty on the short positions. The Lasso regression then regulate the amount of shorting in the portfolio designed by the optimization process, so that the problem stabilizes.

The general form of the l_1-penalized regression with linear constraints is:

$$\widehat{x}_\tau^* = \arg\min_{x \in H} \left\|b - Ax\right\|_2^2 + \tau \left\|x\right\|_1$$

H is an affine subspace defined by linear constraints. The regularized optimal portfolio can be found using an adaptation of the homotopy / LARS algorithm as described in Brodie et al (2008). In appendix A, we provide a detailed description of this algorithm.

4.3 Ridge method

The Ridge regression has been introduced by Hoerl and Kennard (1970) as a more stable alternative to the standard least-squares estimator with potential lower risk. It repre-

sents a different way to penalize the problem using the l_2 norm. The Ridge regression is then given by :

$$\begin{array}{l} \widehat{x}_\tau^* = \arg\min_x \left\| \mu_p 1_T - Rx \right\|_2^2 + \tau \left\| x \right\|_2^2 \\ \text{st } x'\hat{\mu} = \mu_p. \end{array} \tag{7}$$

Contrary to the Lasso regularization, the Ridge does not deliver a sparse portfolio, but selects all the securities with possibly short-selling.

Proposition 1 *The solutions in presence of a risk free asset are:*

$$\begin{cases} x_\tau^* = \mu_p \frac{\hat{\Omega}_\tau^{-1}\hat{\mu}}{\hat{\mu}'\hat{\Omega}_\tau^{-1}\hat{\mu}} & \textit{for a given } \mu_p \\ x_\tau^* = \frac{\hat{\Omega}_\tau^{-1}\hat{\mu}}{1_N'\hat{\Omega}_\tau^{-1}\hat{\mu}} & \textit{for the tangency portfolio} \end{cases}$$

where $\hat{\Omega}_\tau = (R'R + \tau I)$.

We see that the solution for the tangency portfolio corresponds to $\hat{\beta}_\tau / 1_N'\hat{\beta}_\tau$ where $\hat{\beta}_\tau$ is computed using the ridge regularization (5). The equivalence between the two forms of ridge regularizations has been established a long time ago for the unconstrained case and is shown to hold here in a constrained setting.

Proof of Proposition 1. Solving the problem (7) is equivalent to minimizing the Lagrangian:

$$\min_{\{x,\lambda\}} L(x,\lambda) = \frac{1}{2}\left(\|y - Rx\|_2^2 + \tau \|x\|_2^2 \right) + \lambda(\mu_p - x'\hat{\mu})$$

where $y = \mu_p 1_T$. The solutions are given by the system of equations:

$$\begin{cases} \hat{\Omega}_\tau x - R'y - \lambda\hat{\mu} = 0 \\ x'\hat{\mu} = \mu_p \end{cases}.$$

We then have:

$$x = \hat{\Omega}_\tau^{-1} R'y + \lambda \hat{\Omega}_\tau^{-1}\hat{\mu}$$

and using the fact that $x'\mu = \mu_p$ we obtain that:

$$\lambda = \frac{\mu_p - \hat{\mu}'\hat{\Omega}_\tau^{-1}R'y}{\hat{\mu}'\hat{\Omega}_\tau^{-1}\hat{\mu}}.$$

Then for a given expected excess return on the risky portfolio, the optimal investment strategy on the risky asset is:

$$x = \hat{\Omega}_\tau^{-1}R'y + \frac{\mu_p - \hat{\mu}'\hat{\Omega}_\tau^{-1}R'y}{\hat{\mu}'\hat{\Omega}_\tau^{-1}\hat{\mu}}\Omega_\tau^{-1}\hat{\mu} = \frac{\mu_p \hat{\Omega}_\tau^{-1}\hat{\mu}}{\hat{\mu}'\hat{\Omega}_\tau^{-1}\hat{\mu}}.$$

If instead the investor is interested in the tangency portfolio, from the relation $x'1_N = 1$ we have that $\lambda = \frac{1-1'_N\hat{\Omega}_\tau^{-1}R'y}{1'_N\hat{\Omega}_\tau^{-1}\hat{\mu}}$ which leads to the investment rule:

$$x = \hat{\Omega}_\tau^{-1}R'y + \frac{1 - 1'_N\hat{\Omega}_\tau^{-1}R'y}{1'_N\hat{\Omega}_\tau^{-1}\hat{\mu}}\hat{\Omega}_\tau^{-1}\hat{\mu} = \frac{\hat{\Omega}_\tau^{-1}\hat{\mu}}{1'_N\hat{\Omega}_\tau^{-1}\hat{\mu}}.$$

The implied portfolio return is then given by $\mu_p = x'\mu$.

5 Optimal selection of the regularization parameter

5.1 Loss function of estimated allocation

An investor will want to choose τ so that the selected optimal portfolio $\widehat{x}_\tau^*$ is as close as possible to the optimal allocation x^* obtained if μ and Σ were known. To achieve this goal, we select τ in order to minimize a quadratic loss function defined by

$$E\left[(\widehat{x}_\tau^* - x^*)' R'R\,(\widehat{x}_\tau^* - x^*)\right] = E\left[\|R\,(\widehat{x}_\tau^* - x^*)\|^2\right]. \tag{8}$$

The difference between $\hat{x}_\tau^*$ and x^* is weighted by the square matrix $R'R$ as it gives a natural interpretation to the criterion. This criterion is close to the following

$$E\left[\left\|R\widehat{x}_\tau^* - (\mu' x^*)\,1_T\right\|^2\right] = \left\|E\,(R\widehat{x}_\tau^*) - (\mu' x^*)\,1_T\right\|^2 + Var\,(R\widehat{x}_\tau^*)$$

where the first term on the rhs is the squared bias of the return of the estimated optimal portfolio and the second term is its squared risk.

Our goal is to give a convenient expression for the criterion (8). Consider $\widehat{x}_\tau^* = \hat{\beta}_\tau / 1'_N\hat{\beta}_\tau$ where $\hat{\beta}_\tau$ is given by

$$\widehat{\beta}_\tau = \widehat{\Omega}_\tau^{-1}R'1_T \tag{9}$$

where $\widehat{\Omega}_\tau^{-1}$ is a regularized inverse of $\hat{\Omega} = R'R$.

Using the notation $\beta = E\,(R'R)^{-1}\,E\,(R'1_T)$, the optimal allocation x^* can be rewritten as $\beta / 1'_N\beta$. Indeed, we have seen earlier that $x^* = \Sigma^{-1}\mu / 1'_N\Sigma^{-1}\mu$. Let $\Omega = E\,(R'R)$ and $\Omega_T = \Omega / T$.

$$\begin{aligned}
\Sigma^{-1}\mu &= (\Omega_T - \mu\mu')^{-1}\,\mu \\
&= \left(\Omega_T^{-1} + \frac{\Omega_T^{-1}\mu\mu'\Omega_T^{-1}}{1 - \mu'\Omega_T^{-1}\mu}\right)\mu \\
&= \frac{\Omega_T^{-1}\mu}{1 - \mu'\Omega_T^{-1}\mu}
\end{aligned}$$

where the second equality follows from the updating formula for an inverse matrix (see Greene, 1993, p.25). Hence

$$x^* = \frac{\Omega^{-1}\mu}{1'_N\Omega^{-1}\mu} = \frac{\beta}{1'_N\beta}.$$

The criterion (8) can be rewitten as

$$E\left[\left\|R\left(\frac{\widehat{\beta}_\tau}{1'_N\widehat{\beta}_\tau}-\frac{\beta}{1'_N\beta}\right)\right\|^2\right] \tag{10}$$

Minimizing (8) is equivalent to minimizing $E\left[\left\|R\left(\frac{\widehat{\beta}_\tau-\beta}{1'_N\beta}\right)\right\|^2\right]$ which itself is equivalent to minimizing

$$E\left[\left\|R\left(\widehat{\beta}_\tau-\beta\right)\right\|^2\right]. \tag{11}$$

Term (11) depends on the unknown β and hence needs to be approximated. Interestingly, (11) is equal to the prediction error of model (3) plus a constant and has been extensively studied. To approximate (11), we use results on cross-validation from Craven and Wahba (1979), Li (1986, 1987), and Andrews (1991) among others.

We can write

$$R\widehat{\beta}_\tau = M_T(\tau)\,1_T$$

with $M_T(\tau) = R\widehat{\Sigma}_\tau^{-1}R'$. The rescaled MSE

$$\frac{1}{T}E\left[\left\|R\left(\widehat{\beta}_\tau-\beta\right)\right\|^2\right]$$

can be approximated by generalized cross validation criterion:

$$GCV(\tau) = \frac{1}{T}\frac{\left\|\left(I_T - M_T(\tau)\right)1_T\right\|^2}{\left(1-tr\left(M_T(\tau)\right)/T\right)^2}.$$

To obtain the optimal value of τ, it suffices to minimize $GCV(\tau)$ with respect to τ.

5.2 Explicit expression of the cross validation criterion

When $\widehat{\Sigma}_\tau^{-1}$ is a regularized inverse of Σ, $M_T(\tau)$ can be expressed as a function of the orthonormal eigenvectors v_j and eigenvalues λ_j^2, $j = 1, ..., T$ of the matrix RR'. If $N < T$, it is easier to compute v_j^* and λ_j^2, $j = 1, ..., N$ the orthonormal eigenvectors and eigenvalues of the matrix $R'R$ and deduce the spectrum of RR'. Indeed, the eigenvectors of RR' are $v_j = Rv_j^*/\lambda_j$ associated with the same nonzero eigenvalues λ_j^2. We have

$$M_T(\tau)\,w = \sum_{j=1}^{T} q(\tau,\lambda_j)\left(w'v_j\right)v_j$$

for any $T-$ vectors w. Moreover, $trM_T(\tau) = \sum_{j=1}^{T} q(\tau,\lambda_j)$. The function q takes a different form depending on the type of regularization. For Ridge, $q(\tau,\lambda_j) = \lambda_j^2/\left(\lambda_j^2+\tau\right)$.

For Spectral cut-off, $q(\tau, \lambda_j) = I\left(\lambda_j^2 \geq \tau\right)$. For Landweber Fridman, $q(\tau, \lambda_j) = 1 - \left(1 - c\lambda_j^2\right)^{1/\tau}$.

The Lasso estimator does not take the simple form (9). However, Tibshirani (1996) shows that it can be approximated by a ridge type estimator and suggests using this approximation for cross-validation. Let $\tilde{\beta}(\tau)$ be the Lasso estimator for a value τ. By writing the penalty $\sum |\beta_j|$ as $\sum \beta_j^2 / |\beta_j|$, we see that $\tilde{\beta}(\tau)$ can be approximated by

$$\beta^* = \left(R'R + \tau(c) W^-(\tau)\right)^{-1} R' 1_T$$

where c is the upper bound $\sum |\beta_j|$ in the constrained problem equivalent to the penalized Lasso and $W(\tau)$ is the diagonal matrix with diagonal elements $\left|\tilde{\beta}_j(\tau)\right|$, W^- is the generalized inverse of W and $\tau(c)$ is chosen so that $\sum_j \left|\beta_j^*\right| = c$. Since $\tau(c)$ representes the Lagrangian multiplier on the constraint $\sum_j \left|\beta_j^*\right| \leq c$, we always have this constraint binding when $\tau(c) \neq 0$ (ill-posed cases). Let

$$p(\tau) = tr\left\{R\left(R'R + \tau(c) W^-(\tau)\right)^{-1} R'\right\}.$$

The generalized cross-validation criterion for Lasso is

$$GCV(\tau) = \frac{1}{T} \frac{\left\| 1_T - R\widetilde{\beta}(\tau) \right\|^2}{\left(1 - p(\tau)/T\right)^2}.$$

Tibshirani (1996) shows in simulations that the above formula gives good results.

6 Simulations

We use simulated data to assess the performance of each of the investment strategies $x^*(\tau^*)$ we proposed. Precisely we find that most of the times, the strategies obtained through regularization outperform the very tough benchmark of the equally weighted portfolio. As it is done in the literature, one way to compare different strategies is to look at their out-of-sample performances. An alternative way is to investigate the in-sample performance as Fan and Yu (2009).

6.1 A three-factor model

In this section, we use a three factor ($K = 3$) model to assess the in-sample performance of our strategies through a Monte Carlo study. Precisely, we suppose that the N excess returns of assets are generated by the model:

$$r_{it} = b_{i1} f_{1t} + b_{i2} f_{2t} + \cdots + b_{iK} f_{Kt} + \varepsilon_{it} \quad for \quad i = 1, \cdots, N \tag{12}$$

or in a contracted form:

$$R = BF + \varepsilon$$

where b_{ij} are the factors loading of the i^{th} asset on the factor f_j, ε_i is the idiosyncratic noise independent of the three factors and independent of each other.

We assume further a trivariate normal distribution for the factor loading coefficients and for the factors: $b_i \sim N(\mu_b, \Sigma_b)$ and $f_t \sim N(\mu_f, \Sigma_f)$. The ε_i are supposed to be normally distributed with level σ_i drawn from a uniform distribution, so their covariance matrix is $\Sigma_\varepsilon = diag(\sigma_1^2, \cdots, \sigma_p^2)$. As a consequence the covariance matrix of returns is given by:

$$\Sigma = B\Sigma_f B' + \Sigma_\varepsilon$$

The parameters μ_f ,Σ_f, μ_b and Σ_b used in the model (12) are calibrated to market data from July 1980 to June 2008. The data sets used consist of 20 years monthly returns of Fama-French three factors and of 30 industry portfolio from French data library. As pointed out in Fan et al (2008) a natural idea for estimating Σ is to use the least-squares estimators of B, Σ_f and Σ_ϵ and obtain a substitution estimator:

$$\hat{\Sigma} = \hat{B}\hat{\Sigma}_f\hat{B}' + \hat{\Sigma}_\varepsilon$$

where $\hat{B} = RF'(FF')^{-1}$ is the matrix of estimated regression coefficients, Σ_f is the covariance matrix of the three Fama-French factors. Namely the excess return of the proxy of the market portfolio over the one-month treasury bill, the difference of return between large and small capitalization, that capture the size effect, and the difference of returns between high and low book-to-market ratios, that capture the valuation effect. We then select the level of the idiosyncratic noise so that the generated asset returns exhibit three principal components. This means in practice that the covariance matrix of the generated returns have three dominant eigenvalues. We choose idiosyncratic noise to be normally distributed with level σ_i uniformly distributed between 0.01 and 0.05. Once generated, the factors and the factor loadings are kept fixed throughout replications. Table 1 summarizes the calibrated mean and covariance matrix for the factors and the factors loadings.

Parameters for factor loadings				Parameters for factor returns			
μ_b		Σ_b		μ_f		Σ_f	
0.9919	0.0344	0.0309	0.0005	0.0060	0.0019	0.0003	-0.0005
0.0965	0.0309	0.0769	0.0042	0.0014	0.0003	0.0009	-0.0003
0.1749	0.0005	0.0042	0.0516	0.0021	-0.0005	-0.0003	0.0012

Table 1: Calibrated parameters used in simulations

6.2 Estimation methods and tuning parameters

We start by a serie of simulations to assess the performance of the different strategies proposed. This is done relative to the benchmark naive 1 over N strategy and the sample based Markowitz portfolio that is well known to perform poorly. The portfolios considered are the naive evenly weighted portfolio (1oN), the sample-based mean variance portfolio (M), the Lasso portfolio (L), the ridge-regularized portfolio (Rdg), the spectral cut-off regularized portfolio (SC) and the Landweber-Fridman portfolio (LF) as summarized in Table 2.

#	Model	Abbreviations
1	Naive evenly weighted portfolio	1oN
2	Sample-based mean variance portfolio	M
3	Lasso Portfolio	L
4	Optimal Ridge portfolio	Rdg
5	Optimal Spectral cut-off Portfolio	SC
6	Optimal Landweber-Fridman Portfolio	LF

Table 2: List of investment rules

The three regularization techniques introduced to improve the optimality of the sample-based Markowitz portfolio involve a regularization parameter which have each a particular value that corresponds to the sample-based Markowitz portfolio. So our approach can be considered as a generalization that aims to stabilize while improving the performance of the sample-based mean-variance portfolio. Here we give some insights about the effect from tuning different regularization parameters.

The ridge, the spectral cut-off and the Landweber-Fridman schemes have a common feature that they transform the eigenvalues from the singular decomposition of returns covariance matrix so that the resulting estimate has a more stable inverse. This transformation is done with a damping function $q(\tau, \lambda)$ specific to each approach as introduced previously.

The Ridge is the easiest regularization to implement and recovers the sample-based mean-variance minimizer for $\tau = 0$.

For SC, minimizing GCV with respect to τ is equivalent to minimizing with respect to p, the number of eigenvalues ranked in decreasing order. The higher the number of eigenvectors kept, the closer we are to the sample based Markowitz portfolio. For values of τ lower than the smallest eigenvalue, the SC portfolio is identical to the classical sample-based portfolio.

The Landweber-Fridman regularization technique can be implemented in two equivalent ways. Either we perform a certain number l of iterations or we transform the eigenvalues using the function $q(\frac{1}{l}, \lambda)$. Consequently, a larger number of iterations cor-

responds to smaller value the penalty term τ that belong to interval $]0,1[$. Besides, for a large number of iterations ($\tau \approx 0$) the regularized portfolio $\hat{x}_\tau$ obtained becomes very close to the sample-based optimal portfolio $\hat{x}$. In the Landweber-Fridman case we seek the optimal number of iterations so that $\hat{x}_\tau$ is the closest to the theoretically optimal rule x^*. In the ill-posed case we typically have a very few number of iterations which corresponds to a value of τ close to one. That is, $\hat{x}_\tau$ is far from the Markowitz allocation $\hat{x}$ known to perform very poorly.

The effect of tuning the penalty τ in the l_1-penalized regression have been extensively studied Brodie et al (2009). Our approach is different in the fact that the portfolio we are interested in is the tangency portfolio which can be derived up to a normalization using an unconstrained regression. We solve this problem using the unconstrained version of the Homotopy/Lars Algorithm (see Appendix for a detailed description). For a given value of the penalty term, the algorithm determine the number of assets (from 1 to N) to be included in the portfolio as well the weights associated up to a normalization. For illustration, Figure 1 plots the portfolio weights constructed using the 10 industry portfolios from Fama and French data library, against the penalty term. The data set used is the French 10 industry Portfolios from January 1990 to September 2009. We seek to determine the portfolio weights associated with each of the 10 industry portfolios. The case $\tau = 0$ corresponds to the sample-based tangency portfolio. The higher the penalty the smaller the number of industries included in the optimal portfolio. Furthermore, there exists a threshold (here $\tau \geq 3.3$) beyond which all the industry portfolios are ruled out. By order of entrance in the optimal portfolio, the industry considered are: HiTec, Enrgy, Hlth, NoDur, Other, Durbl, Shops, Telcm and Utils.

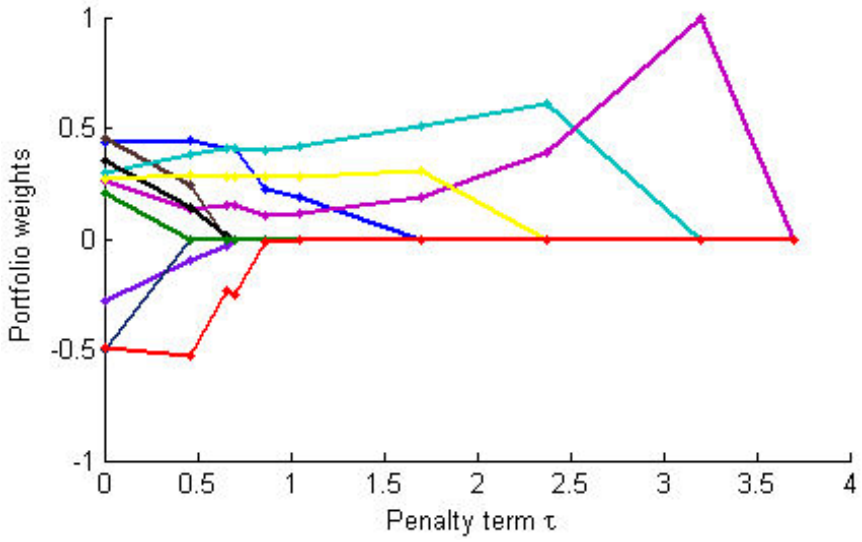

Figure 1: Evolution of the tangency portfolio weights with the penalty term τ values.

Figure 2 gives an idea of the shape of the GCV for a single series with $N = 100$ and $T = 120$. In our computations, the GCV for the Rdg the SC and the LF portfolios are minimized respectively with respect to τ, the number p of eigenvectors kept in the spectral decomposition of returns covariance matrix, and the number of iterations l.

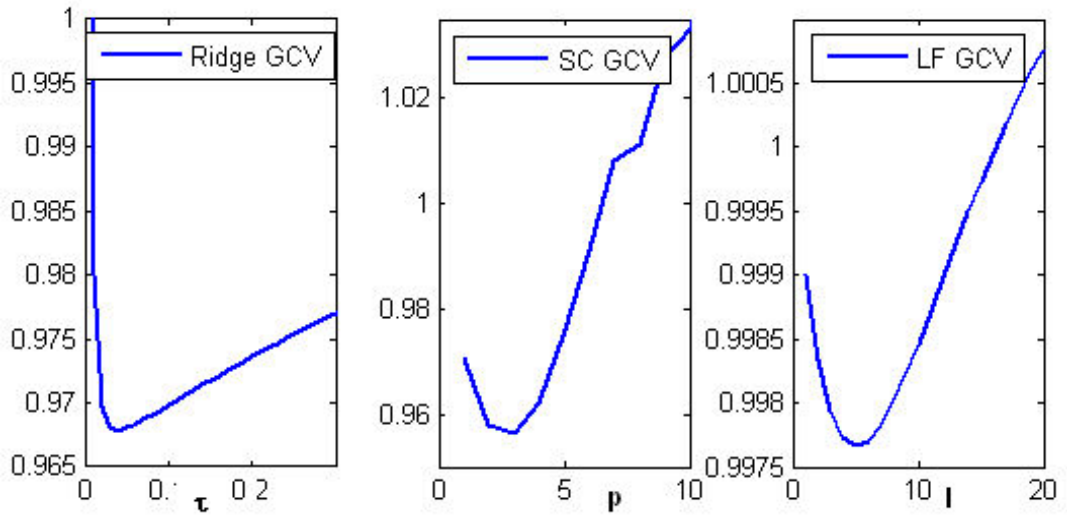

Figure 2: GCV as function of regularization parameters for the Ridge, the SC and the LF schemes. In this figure we consider a model with N=100 assets and T=120 observations of asset returns. The minima are obtained for $\tau = 0.04, p = 3, l = 5$.

6.3 In-sample performance

We perform 1000 replications. In each of the replications, model (12) is used along with the parameters in Table 1 to generate $T = 120$ monthly observations of $N = 100$ asset excess returns. This setting corresponds to an ill-posed case with a large number of assets and a number observations relatively small. We compare three different versions of the Sharpe ratio. The theoretically optimal Sharpe ratio $SR(x^*) = \sqrt{\mu' \Sigma^{-1} \mu} \equiv SR^*$ using the theoretically optimal weights x^*, the actual Sharpe ratio that results from using regularized strategies $SR(\hat{x}_\tau) = \frac{\hat{x}'_\tau \mu}{\sqrt{\hat{x}_\tau \Sigma \hat{x}_\tau}}$ and the empirical Sharpe ratio $SR_T(\hat{x}_\tau) = \frac{\hat{x}'_\tau \hat{\mu}}{\sqrt{\hat{x}'_\tau \hat{\Sigma} \hat{x}_\tau}}$ obtained using regularized portfolio and sample-based moments. We report descriptive statistics on the Sharpe ratios across replications in Tables 3 and 4.

Tables 3 and 4 display descriptive statistic on empirical and actual Sharpe ratios obtained through replications. It appears that the Markowitz sample-based strategy does not provide the best actual Sharpe ratio as stressed in the literature. Clearly when no regularization is used there is a huge discrepancy between the empirical Sharpe ratio

and the actual Sharpe ratio. Precisely using the mean-variance strategy, we obtain an empirical version of the Sharpe ratio which is overly optimistic. This is consistent with the observation made by Fan et al (2009) concerning empirical risk: the empirical risk is under evaluated when no constraint on portfolio weights is imposed implying an over evaluation of the Sharpe ratio. Consequently, using the empirical Sharpe ratio may lead to wrong conclusions. On the other hand, using the Rdg, SC or LF optimal strategies we propose, the discrepancy between the empirical and the actual Sharpe ratio is greatly reduced and the resulting investment rules become very close to the theoretically optimal portfolio. For instance, the true Sharpe ratio is $SR = 0.1847$ and the actual Sharpe ratio for the Rdg, SC and the LF are on average respectively 0.15775, 0.16891 and 0.15316. So on average we get Sharpe ratio higher than the Sharpe ratio provided by the 1 over N rule except for the LF which remains of a comparable order. Concerning the Lasso, the regularized portfolio obtained performs better than the sample-based Markowitz portfolio but is still far from what is theoretically optimal. Patently, the adaptation of GCV criterion proposed by Tibshirani (1996) does not provide a good approximation to the Lasso penalty term that minimizes the MSE of allocations in presence of a large number of assets relative to the sample size. An alternative way to proceed is to consider a two-stage procedure. That is apply the Rdg, the SC or the LF scheme to subsets of assets selected by the Lasso and then select the portfolio that maximizes the Sharpe ratio over subsets of assets. Simulations show that using the Lasso first and then applying Rdg, SC or LF regularization techniques reduce the number of assets and provides better results than the case where the regularization is directly applied to the whole set of assets.

6.4 Monte Carlo assessment of GCV

A question that we seek to answer through simulations is whether the generalized cross validation (GCV) criterion minimizer is a good approximation of the theoretically optimal τ that minimizes the MSE of allocations. To address this issue, we use the 1000 samples generated in the previous section ($N = 100$ and $T = 120$). For each of the samples, we compute the GCV as a function of τ and determine its minimizer $\hat{\tau}$. We provide some statistics for $\hat{\tau}$ in Table 5.

To compute the MSE of $\hat{\tau}$, we need to derive the true optimal regularization parameter τ_0. To do so, we use our 1000 samples to approximate τ_0 as the minimizer $\hat{\tau}_0$ of the sample counterpart of the MSE of allocations:

$$\hat{E}\left[\left\|R\left(\widehat{x}\left(\tau\right)-x^{*}\right)\right\|^{2}\right]$$

where $\hat{E}$ is an average over the 1000 replications and x^* is the true optimal allocation. This first step provides us with an estimation of the true parameter which is a function of the number of assets N and the sample size T under consideration.

Simulations reported in Table 5 show that the minimizer of the GCV function is indeed a good approximation to the minimizer of the theoretical mean squared error of allocations. For each regularization scheme the true optimal parameter is approximated by the value that minimizes the sample counterpart of the MSE of allocation. In general, regularization parameters have a relatively small variability across replications and they are accurate in the sense that they are close to the estimations of their theoretical value. Indeed, the Rdg optimal penalty term has a standard deviation of order 10^{-3} and a mean squared error of order 10^{-5}. In the SC case, the GCV criterion selects often $p = 3$ which is the minimizer of the theoretical MSE, namely the number of factors used. Concerning the Landweber-Fridman approach, the optimal allocations are obtained on average after $l = 6$ iterations which is consistent with the average number of iterations 6.20 that minimizes the GCV. However the facts that more 25% of samples have required more than 7 iterations and that the MSE of l is equal 10 indicate the presence of outliers. This is most likely due to the small sample size.

6.5 Out-of-sample performance

From now on, we adopt the rolling sample approach in Mackinlay and Pastor (2000) and DeMiguel et al (2007). Given a dataset of size T and a window size M, we obtain a set of $T - M$ out-of-sample returns, each generated recursively at time $t + 1$ using the M previous returns. The time series of out-of-sample returns obtained can then be used to compute out-of-sample performance for each strategy. As pointed out by Brodie et al (2009), this approach can be seen as an investment exercise to evaluate the effectiveness of an investor who bases his strategy on the M last periods returns and a given optimal strategy.

First we generate a single sample with 40 asset returns and $T = 1000$ monthly returns to which we apply the rolling window strategy described above. We compare the naive evenly weighted portfolio, the Markowitz sample-based portfolio, the optimal Ridge portfolio and we represent the mean variance frontier from the true moments, from the estimated moments (in sample using 1000 time periods) and estimated moments out-of-sample using 120 time periods in Figure 3. The Ridge uses the true value of τ obtained via 10000 replications of samples with $N = 40$ and $T = 120$. The objective of this first experiment is to have more insights about the effect of regularization by comparing with theoretically optimal portfolios obtained using the known true moments of asset returns.

Figure 3 shows how dramatic estimation errors can affect the optimality of the mean variance optimizer. Even by using all the 1000 months available there is still huge discrepancy between the theoretically optimal mean-variance frontier and the estimated ones. The effect are even worse when estimation is done using a rolling windows of size M=120. In this case, we can see that the sample based optimal Markowitz portfolio can be very far from the tangency portfolio. The main message from Figure 3 is that regularization reduces the distance between the sample based tangency portfolio and

the theoretical tangency portfolio.

In the final set of simulations we use the generalized cross validation criterion to determine the optimal tuning parameter for the Ridge, the spectral cut-off and the Landweber-Fridman. The resulting investment rules are then compared with respect to their out-of-sample Sharpe ratios. We use a sample size of $T = 1000$ and rolling windows length $M = \{60, 120\}$ months which correspond respectively to five and ten years.

Table 6 reports the out-of-sample Sharpe ratio of strategies listed in table 2 for sub-periods extending over 5 years each. Optimal value for regularization parameter are obtained by minimizing the GCV. These Sharpe ratios reflect the performance an investor would have if he were trading on assets generated by model (12). It appears that whith respect to the sharpe ratio they generate, the three rules Rdg, SC and LF tremendously improve the Markowitz portfolio and even outperform most of the time the naive investment rule. The lasso strategy using the approximated GCV do not provide considerable improvement to the Markowitz portfolio as noticed in the in-sample simulation excercise. Table 7 displays the outcome of simulated investment exercise that use a larger estimation window $M = 120$. The same conclusions as in the case $M = 60$ hold.

7 Empirical Application

In this section we apply the methodology we propose to historical monthly returns on 48 industry sector portfolio (abbreviated to FF48) compiled by Fama and French . This dataset ranges from July 1969 to June 2009. As previously, the optimal portfolios listed in Table 2 are constructed at the end of June every year from 1974 to 2009 for a rolling window of size $M = 60$ months and from 1979 to 2009 for a rolling window of size $M = 120$. The risk free rate is taken to be the one month T-bill rate. We use an appropriate range for τ to carry out our optimizations depending on the regularization scheme. For the Ridge we use a grid on $[0, 2]$, for the SC the parameter p ranges from 1 to $p_{max} = N-1$ while for LF we use a maximal number of iterations equal to $l_{\max} = 20$. For instance for $M = 60$ our first portfolios are constructed in June 1974. From the $T \times N$ matrix of excess returns R, empirical mean $\hat{\mu}$, and regularized inverse matrix of excess returns $\Sigma(\tau)$ are computed using historical returns from July 1969 to June 1974. We then deduce the tangency portfolio for each of the optimal rules. The portfolio obtained is kept from July 1974 to June 1975 and its returns recorded. We repeat the same process using data from July 1970 to June 1975 to predict portfolio return from July 1975 to June 1976. The simulated investment exercise is done recursively until the last set of portfolio constructed at the end of June 2009.

From Tables 8 to 10, we see that using a regularized portfolio is a more stable alternative to the Markowitz sample-based portfolio. In this empirical study the LF

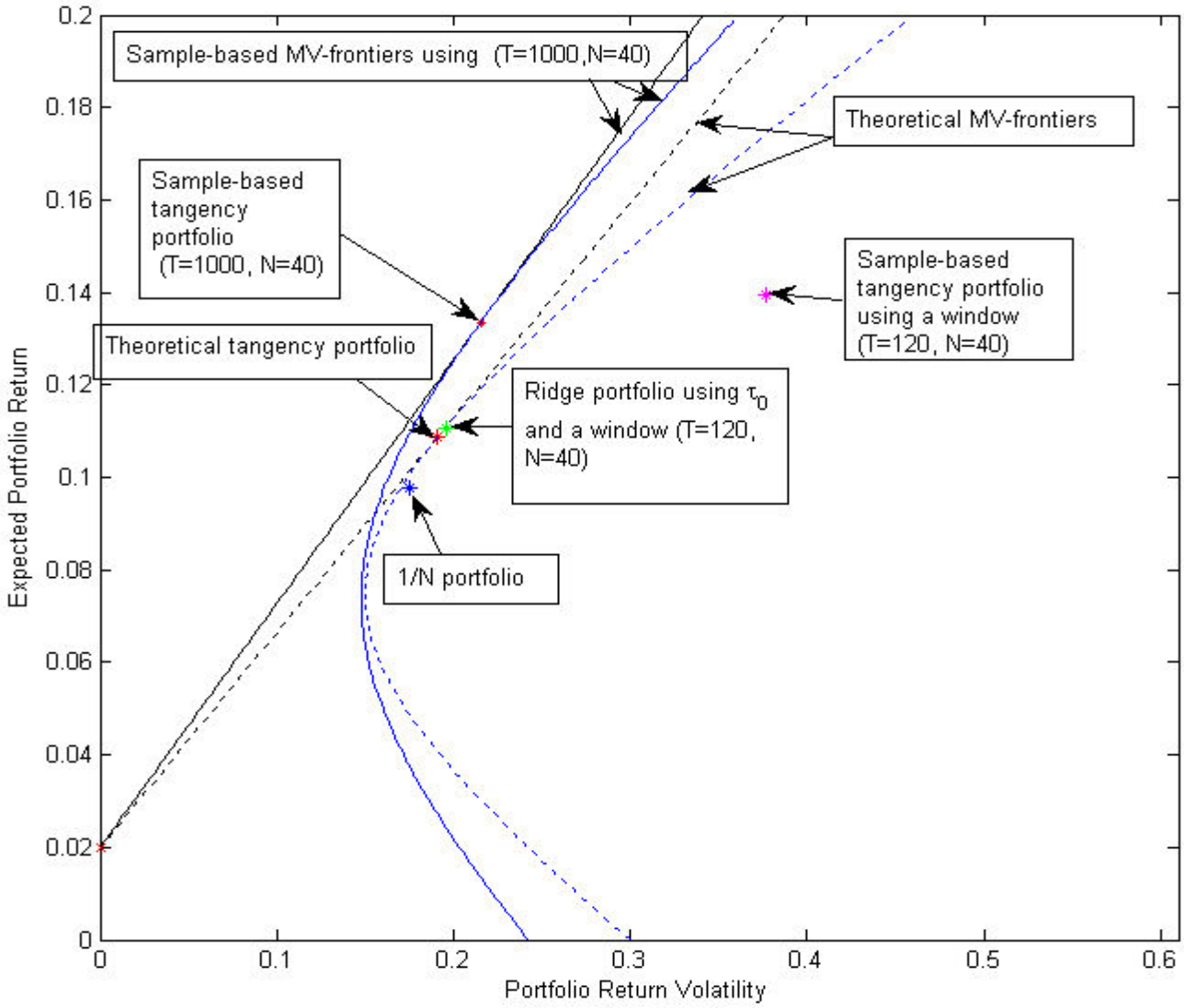

Figure 3: Effect of regularization using the the sample counterpart of the minimizer of the MSE of allocations.

turns out to perform the best. However the performance of these regularized portfolio are not as good as in case where we suppose that excess return follow the market model. This may be explained by the fact many characteristics of returns are not taken into account while solving the mean-variance optimization problem.

8 Conclusion

In this paper, we address the issue of error estimation in the framework of the mean-variance analysis. We propose to regularize the portfolio choice problem using regularization techniques from inverse problem literature. These regularization techniques namely the ridge, the spectral cut-off, and Landweber-Fridman involve a regularization parameter or penalty term whose optimal value is selected to minimize the mean squared error of allocations. We show that this is equivalent to select the penalty term as the minimizer of a generalized cross validation criterion.

Our simulations and empirical study show that in ill-posed cases a regularization to covariance matrix drastically improve the performance of mean-variance problem and in many cases outperforms the naive portfolio and Lasso portfolios. The methodology we propose can be used as well for any investment rule that requires an estimate of the covariance matrix and given a performance criterion. The appeal of the investment rules we propose is that they are easy to implement and provide comparable and very often better results than the existing asset allocation strategies.

Table 3: Descriptive statistics of the empirical Sharpe ratio derived from 1000 replications and using a three-factor model ($SR^* = 0.1847$)

	Empirical Sharpe ratio for Optimal strategies					
Statistics	1oN	M	Lasso	Rdg	SC	LF
Mean	0.13519	0.39758	0.48643	0.53786	0.28921	0.18064
Std	0.00831	2.36727	0.90460	0.18107	0.04421	0.01918
mse	0.00252	5.64370	0.90853	0.15748	0.01288	0.00038
q1	0.12972	-2.16625	0.28526	0.35823	0.26490	0.16927
median	0.13526	1.91352	0.86901	0.51813	0.29331	0.17705
q3	0.14047	2.43239	1.08892	0.71851	0.31593	0.18790

Table 4: Descriptive statistics of the actual Sharpe ratio derived from 1000 replications and using a three-factor model ($SR^* = 0.1847$)

	Actual Sharpe ratio for Optimal strategies					
Statistics	1oN	M	Lasso	Rdg	SC	LF
Mean	0.15503	0.00565	0.04094	0.15775	0.16891	0.15316
Std	0.00000	0.02215	0.05389	0.00605	0.01111	0.00149
mse	0.00088	0.03255	0.02357	0.00076	0.00037	0.00100
q1	0.15503	-0.01063	0.02514	0.15328	0.15896	0.15218
median	0.15503	0.00734	0.05274	0.15885	0.17385	0.15307
q3	0.15503	0.02304	0.07129	0.16258	0.17855	0.15414

Table 5: Descriptive statistics of optimal regularization parameters obtained from the 1000 samples usind a three factor model

Regularization	$\hat{\tau}_0$	Mean	Std	mse	mode	q1	median	q3
Rdg (τ)	0.03	0.0344	0.0329	0.0011	0.01	0.01	0.02	0.05
SC (p)	3	3.661	3.0851	9.945	3	2	3	4
LF (l)	6	6.207	3.2503	10.597	5	4.5	5	7

Period	**1oN**	**M**	**L**	**Rdg**	**Sc**	**LF**
t=61-120	-3.3	4.1	-2.0	5.3	0.7	4.8
t=121-180	22.8	-3.3	4.8	22.8	21.8	23.9
t=181-240	-1.0	7.3	-15.4	-2.3	-1.0	-6.1
t=241-300	7.7	4.4	-9.2	2.5	7.3	5.5
t=301-361	26.2	17.4	21.0	26.4	26.0	26.5
t=361-420	17.8	15.3	-22.0	22.2	15.8	16.9
t=421-480	10.6	10.5	-7.1	9.2	4.9	10.3
t= 481-540	4.3	-4.8	-23.7	-3.6	3.4	-1.5
t= 541-600	51.3	19.0	5.5	34.8	39.3	32.9
t= 601-660	9.1	5.4	5.2	8.2	8.7	8.6
t= 661-720	15.3	8.3	-17.9	14.0	15.2	10.1
t= 721-780	19.1	-8.9	-7.6	19.6	16.7	19.8
t= 781-840	0.8	-2.2	6.8	7.1	6.1	1.2
t= 841-900	11.7	7.2	-6.3	9.1	22.7	6.7
t= 901-960	5.8	-6.8	-1.3	6.6	5.6	0.6

Table 6: Out-of-sample Sharpe-ratio in percent from simulated data using a rolling window of size 60 months .

Period	1oN	M	L	Rdg	Sc	LF
t=121-240	11.0	-7.5	-5.9	12.1	9.9	11.3
t=240-360	16.1	-8.2	0.0	16.6	15.8	16.6
t=361-480	14.4	-2.9	2.6	12.3	10.0	10.6
t=481-600	23.9	-5.2	22.9	25.2	18.8	20.4
t=601-720	11.9	-0.4	0.4	11.9	13.1	12.5
t=721-840	11.1	11.9	0.1	9.5	12.8	15.6
t=841-960	8.4	2.9	-3.1	9.9	13.1	8.2

Table 7: Optimal portfolio out-of-sample Sharpe ratios using a rolling windows of size 10 years

Period	**1oN**	**M**	**Ns**	**Rdg**	**Sc**	**LF**
07/36 - 06/46	16.3	-3.8	15.8	15.3	-12.0	15.7
07/46 - 06/56	28.9	14.6	31.8	29.5	29.5	29.3
07/56 - 06/66	23.8	4.4	23.6	22.2	17.1	21.5
07/66 - 06/76	3.1	-9.1	3.6	4.8	1.8	4.8
07/76 - 06/86	16.8	-4.4	19.1	17.9	8.1	17.9
07/86 - 06/96	17.2	6.1	12.2	15.8	10.8	16.5
07/96 - 06/06	14.2	3.0	19.6	11.7	11.8	12.1

Table 8: Performances in term of Sharpe ratio(%) for FF10 using a rolling window length of 10 years

Period	**1oN**	**M**	**Ns**	**Rdg**	**Sc**	**LF**
07/69 - 06/74	13.3	8.0	3.9	13.2	-26.1	28.0
07/74 - 06/79	10.7	-28.8	13.0	8.8	5.0	12.3
07/79 - 06/84	18.2	13.0	2.8	29.5	27.4	20.7
07/84 - 06/89	10.8	3.3	-3.9	4.9	-20.4	11.0
07/89 - 06/94	29.1	4.3	18.6	34.2	29.7	32.8
07/99 - 06/04	7.7	3.8	-11.7	7.4	17.3	8.7
07/04 - 06/09	-1.2	-24.9	3.4	-1.9	-9.1	2.2

Table 9: Performances in term of Sharpe ratio(expressed in %) for FF48 using a rolling window length of 5 years

Period	1oN	M	Ns	Rdg	Sc	LF
07/69 - 06/79	14.7	-1.8	14.4	13.3	14.9	13.2
07/79 - 06/89	20.0	-5.7	26.3	24.3	16.9	21.9
07/89 - 06/09	2.7	-4.9	-0.4	-5.6	-9.8	-0.6

Table 10: Performances in term of Sharpe ratio(%) for FF48 using a rolling window length of 10 years

References

[1] Andrews, D. (1991) "Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors", *Journal of Econometrics*, 359-377.

[2] J. Bai and S. Ng "Forecasting economic time series using targeted predictors", *Journal of Econometrics*, 146, 304-317, 2008.

[3] W. Brandt, *Portfolio choice problems*, http://home.uchicago.edu/ lhansen/handbook.htm, 2004.

[4] M. Britten-Jones "The sampling error in estimates of mean-variance efficient portfolio weights", *Journal of Finance* 54, 655-671, 1999.

[5] J. Brodie, I. Daubechies, C. De Mol, D. Giannone, I. Loris, "Sparse and stable Markowitz portfolios", Proceedings of the National Academy of Sciences of the USA 2009 106:12267-12272.

[6] M. Carrasco, "A regularization approach to the many instrument problem", *mimeo,* 2009.

[7] M. Carrasco, J-P. Florens, E. Renault "Linear Inverse Problems and Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization", *Handbook of Econometrics*, Vol. 6B, 2007.

[8] P. Craven and G. Wahba "Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of the generalized cross-validation", *Numer. Math.* 31, 377-403, 1979.

[9] V. DeMiguel, L. Garlappi, R. Uppal, "Optimal Versus Naive Diversification : How Inefficient is the 1/N Portfolio Strategy ? ", *The review of Financial studies*, 2007.

[10] C. De Mol, D. Giannone, and L. Reichlin (2009) "Forecasting using a large number of predictors: Is Bayseian shrinkage a valid alternative to principal components?", *Journal of Econometrics*, 146, 318-328, 2008.

[11] A. Frost E. Savarino, "An Empirical Bayes Approach to Efficient Portfolio Selection", *Journal of Financial and Quantitative Analysis*, 1986.

[12] Jobson, J.D. and B. Korkie (1983) "Statistical Inference in Two-Parameter Portfolio Theory with Multiple Regression Software"", *Journal of Financial and Quantitative Analysis*, 18, 189-197.

[13] R. Kan and G. Zhou "Optimal portfolio choice with parameter uncertainty", *Journal of Financial and Quantitative Analysis*, 42, 621-656, 2007.

[14] O. Ledoit, M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection", *The Journal of Empirical Finance*, 10,603-621, 2003.

[15] O. Ledoit, M. Wolf, "Honey, I shrunk the sample covariance matrix", *The Journal of Portfolio Management*, 31, 2004a.

[16] O. Ledoit, M. Wolf, " A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices", *Journal of Multivariate Analysis*, 88, 365-411, 2004b.

[17] K-C. Li (1986) "Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing", *The Annals of Statistics*, 14, 1101-1112.

[18] K-C. Li (1987) "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete Index Set", *The Annals of Statistics*, 15, 958-975.

[19] A. Craig MacKinlay, Lubos Pastor, "Asset Pricing Models: Implications for Expected Returns and Portfolio Selection", *The Review of financial studies*, Vol. 13, No. 4, PP 883-916, 2000.

[20] H. M. Markowitz, "Portfolio selection", *The Journal of Finance*, Vol. 7, No. 1 , pp. 77-91, 1952.

[21] Jean Luc Prigent, *Portfolio Optimization and performance analysis*, Chapman and Hall/CRC, 2007.

[22] J. Stock and M. Watson "Forecasting Using Principal Components from a Large Number of Predictors", *Journal of the American Statistical Association*, 2002.

[23] Robert Tibshirani, "Regression Shrinkage and Selection via the Lasso "*J. R. Statist. Soc. B*, 1996.

[24] J. Tu and G. Zhou, " Markowitz Meets Talmud: A combination of Sophisticated and Naive Diversification Strategies", working paper, 2009.

[25] Hal Varian, "A Portfolio of Nobel Laureates: Markovitz, Miller and Sharpe ", *The Journal of economic Perspectives*, Vol 7, No. 1, pp 159-169, 1996.

Appendix A: Homotopy - LARS Algorithms for penalized least-squares

Homotopy (Continuation) is a general approach for solving a system of equation by tracking the solution of nearby system of parametrized equation. In the penalized Lasso case the Homotopy variable is the penalty term. We give below a detailed description of the Homotopy/LARS algorithm which provides the solution path to the l1-penalized least-squares objective function:

$$\tilde{x}(\tau) = \arg\min_x \|y - Rx\|_2^2 + \tau \|x\|_1 .$$

The solution to this minimization problem $\tilde{x}(\tau)$ is provided as a continuous piecewise function of the penalty τ satisfying the variational equations given by:

$$\begin{cases} (R'(y - Rx))_i = \frac{\tau}{2}\text{sgn}(x_i) & x_i \neq 0 \\ |(R'(y - Rx))_i| \leq \frac{\tau}{2} & x_i = 0 \end{cases}$$

Meaning that the residual correlations $b_i = (R'(y - Rx))_i$ corresponding to non zero weights are equal to $\tau/2$ in absolute value, while the absolute residual correlation corresponding the zero weights must by bounded by $\tau/2$. Throughout the algorithm, it is critical to identify the set of active elements, that is the components with non zeros weights. At a given iteration k of the algorithm this set is denoted by $J_k = \left\{ i \;\; for\; which\, |b_i| = \frac{\tau^k}{2} \right\}$, and also corresponds to the set of maximal residual correlations components.

The algorithm starts with an initial solution satisfying the variational equations, for a penalty term suitably chosen. The obvious initial solution is obtained by setting all the weights to zeros. The corresponding penalty term τ_0, must then satisfy $\tau_0 \equiv 2\max_i |(R'y)_i|$. Hence we have that $\tilde{x}(\tau) = 0$ for all $\tau \geq \tau_0$. This allow us to set $J_1 = \{i^*\}$, where $i^* = arg\max_i |(R'y)_i|$.

From one iteration k to the next the algorithm manages to update the active set J_k, which represents the support of $\tilde{x}(\tau_k)$, so that the first-order conditions remains satisfied. Hence in each iteration $k+1$, the vector b decreases at the same rate γ^{k+1} in the active set to preserve the same level of correlation for active elements.

$$(b^{k+1})_{J_{k+1}} = (b^k)_{J_{k+1}} - \gamma^{k+1}(sign(b^k))_{J_{k+1}}$$

This result is obtained by updating the optimal weights while moving along a walking direction u^{k+1}:

$$x(\tau^{k+1}) = x(\tau^k) + \gamma^{k+1} u^{k+1}$$

Denote R_J the submatrix consisting of the columns J of R, the walking direction u^{n+1} is a solution to a linear system:

$$R'_{J_{k+1}} R_{J_{k+1}} (u^{k+1})_{J_{k+1}} = (\mathrm{sgn}(b^k)_{J_{k+1}}) = \left(\mathrm{sgn}(b_j^k)_{j \in J_{k+1}}\right) = v^{k+1}$$

The remaining components of u^{k+1} are set to *zero* that is:

$$u_i^{k+1} = 0 \quad for \;\; i \notin J_{k+1}$$

The step γ_{k+1} to make in direction u^{k+1} to find $x(\tau^{k+1})$ is the minimum value such that an inactive element becomes active or the reverse.

If an inactive element i becomes active, it means that its correlation reached the maximal correlation in the descent procedure. And then is must be case that:

$$\left|b_i^k - \gamma^{k+1} r_i'^{k+1}\right| = \left|(b^k)_{J_{k+1}} - \gamma^{k+1} v_{J_{k+1}}\right| = \tau_{k+1} = \frac{\tau^k}{2} - \gamma^{k+1}$$

with r_i is the i^{th} column of R. This implies that:

$$\gamma^{k+1} = \frac{\frac{\tau^k}{2} - b_i^k}{1 - r_i'^{k+1}} \; or \; \gamma^{n+1} = \frac{\frac{\tau^k}{2} + b_i^k}{1 + r_i'^{k+1}}$$

The optimal step is then given by :

$$\gamma_+^{k+1} = \min_{i \in J^c}{}^{+} \left\{ \frac{\frac{\tau^k}{2} - b_i^k}{1 - r_i'^{k+1}} \; ; \frac{\frac{\tau^k}{2} + b_i^k}{1 + r_i'^{k+1}} \right\}$$

On the other hand, if γ^{n+1} is such that an active element i reaches zero then (**??**) implies that:

$$\gamma_-^{k+1} = -\frac{x_i^k}{u_i^{k+1}}$$

The smallest step to make so that an element leaves the active set:

$$\gamma_-^{k+1} = \min_{i \in J_{k+1}}{}^{+} \left(-\frac{x_i^k}{u_i^{k+1}} \right)$$

Finally the next step is given by:

$$\gamma^{k+1} = \min\left\{\gamma_+^{k+1}, \gamma_-^{k+1}\right\}$$

At the end of each stage the corresponding penalty term is $\tau^{k+1} = \tau^k - 2\gamma^{k+1}$ which is smaller than τ^k. We stop when τ^{k+1} becomes negative. After $q+1$ iterations the Algorithm provides $q+1$ breakpoints $\tau_0 > \tau_1 > \cdots > \tau_q$ and their corresponding minimizers $x(\tau_i)$. From there, the optimal solution for any τ can be deduced by linear interpolation.