Sparse and stable Markowitz portfolios

Joshua Brodie[a], Ingrid Daubechies[a,b,1], Christine De Mol[c], Domenico Giannone[d], and Ignace Loris[c,e]

[a]Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544-1000; [b]Department of Mathematics, Princeton University, Princeton, NJ 08544-1000; [c]Department of Mathematics, European Center for Advanced Research in Economics and Statistics, Université Libre de Bruxelles, 1050 Brussels, Belgium; [d]European Central Bank, European Center for Advanced Research in Economics and Statistics, and Centre for Economic Policy Research, London EC1V ODG, United Kingdom; and [e]Department of Mathematics, Vrije Universiteit Brussel, 1050 Elsene, Belgium

Contributed by Ingrid Daubechies, April 23, 2009 (sent for review December 3, 2008)

We consider the problem of portfolio selection within the classical Markowitz mean-variance framework, reformulated as a constrained least-squares regression problem. We propose to add to the objective function a penalty proportional to the sum of the absolute values of the portfolio weights. This penalty regularizes (stabilizes) the optimization problem, encourages sparse portfolios (i.e., portfolios with only few active positions), and allows accounting for transaction costs. Our approach recovers as special cases the no-short-positions portfolios, but does allow for short positions in limited number. We implement this methodology on two benchmark data sets constructed by Fama and French. Using only a modest amount of training data, we construct portfolios whose out-of-sample performance, as measured by Sharpe ratio, is consistently and significantly better than that of the naïve evenly weighted portfolio.

penalized regression | portfolio choice | sparsity

In 1951, Harry Markowitz ushered in the modern era of portfolio theory by applying simple mathematical ideas to the problem of formulating optimal investment portfolios (1). He argued that single-minded pursuit of high returns constitutes a poor strategy, and suggested that rational investors must, instead, balance their desires for high returns and for low risk, as measured by variability of returns.

It is not trivial, however, to translate Markowitz's conceptual framework into a portfolio selection algorithm in a real-world context. The recent survey (2) examined several portfolio construction algorithms inspired by the Markowitz framework. Given a reasonable amount of training data, the authors found none of the surveyed algorithms able to significantly or consistently outperform the naïve strategy where each available asset is given an equal weight in the portfolio. This disappointing performance is partly due to the structure of Markowitz's optimization framework. Specifically, the optimization at the core of the Markowitz scheme is empirically unstable: small changes in assumed asset returns, volatilities, or correlations can have large effects on the output of the optimization procedure. In this sense, the classic Markowitz portfolio optimization is an ill-posed (or ill-conditioned) inverse problem. Such problems are frequently encountered in other fields; a variety of regularization procedures have been proposed to tame the troublesome instabilities (3).

In this article, we discuss a regularization of Markowitz's portfolio construction. We will restrict ourselves to the traditional Markowitz mean-variance approach. (Similar ideas could also be applied to different portfolio construction frameworks considered in the literature.) Moreover, we focus on one particular regularization method, and highlight some very special properties of the regularized portfolios obtained through its use.

Our proposal consists of augmenting the original Markowitz objective function by adding a penalty term proportional to the sum of the absolute values of the portfolio weights. This term is known in the mathematical literature as an ℓ_1 norm. We allow ourselves to adjust the importance of this penalty with a "tunable" coefficient. For large values of this coefficient, optimization of the penalized objective function turns out to be equivalent to solving the original (unpenalized) problem under an additional positivity condition on the weights. As the tunable coefficient is decreased, the optimal solutions are given more latitude to include short positions. The optimal solutions for our penalized objective function can thus be seen as natural generalizations of the "no-short-positions" portfolios considered in ref. 4. We show that these regularized portfolios are sparse, i.e., they have few active positions (few nonzero weights).

In addition to stabilizing the optimization problem (5) and generalizing no-short-positions–constrained optimization, the ℓ_1 penalty facilitates treatment of transaction costs. For large investors, whose principal cost is a fixed bid–ask spread, transaction costs are effectively proportional to the gross market value of the selected portfolio, i.e., to the ℓ_1 penalty term. For small investors, volume-independent "overhead" costs cannot be ignored, and thus transaction costs are best modeled via a combination of an ℓ_1 penalty term and the number of assets transacted; minimizing such a combination is tantamount to searching for sparse solutions (sparse portfolios or sparse changes to portfolios), a goal that, we will argue, is also achieved by our use of an ℓ_1 penalty term.

We use the methodology to compute efficient investment portfolios with two sets of portfolios constructed by Fama and French as our assets: the 48 industry portfolios and the 100 portfolios formed on size and book-to-market. Using data from 1971 to 2006, we construct an ensemble of portfolios for various values of our tunable coefficient and track their out-of-sample performances. We find a consistent and significant increase in Sharpe ratio compared with the naïve equal-weighting strategy. With the 48 industry portfolios as our assets, the best portfolios we construct have no short positions. With the 100 portfolios as our assets, the best portfolios constructed by our methodology *do* include short positions.

We are not alone in proposing the use of regularization in the context of Markowitz-inspired portfolio construction; ref. 6 discusses several different regularization techniques for the portfolio construction problem, including the imposition of constraints on appropriate norms of the portfolio weight vector. Our work* differs from ref. 6 in that our goal is not only regularization: we are interested in particular in the stable construction of sparse portfolios, which is achieved by ℓ_1 penalization, as demonstrated by our analysis and examples.

Sparse Portfolio Construction

We consider N securities, denoting their returns at time t by the $N \times 1$ vector $\mathbf{r}_t = (r_{1,t}, \ldots, r_{N,t})^\top$. We write $\mathbf{E}[\mathbf{r}_t] = \boldsymbol{\mu}$ for the vector of expected returns of the different assets, and $\mathbf{E}[(\mathbf{r}_t - \boldsymbol{\mu})(\mathbf{r}_t - \boldsymbol{\mu})^\top] = \boldsymbol{C}$ for the covariance matrix of returns.

Author contributions: I.D., C.D.M., D.G., and I.L. designed research; J.B., I.D., C.D.M., D.G., and I.L. performed research; J.B., D.G., and I.L. analyzed data; and J.B., I.D., C.D.M., D.G., and I.L. wrote the paper.

The authors declare no conflict of interest.

[1] To whom correspondence should be addressed. E-mail: ingrid@math.princeton.edu.

*The first presentation of our work is given in ref. 7, independent of and simultaneous with ref. 6. An earlier reference pointing to regularization in portfolio estimation is ref. 8.

This article contains supporting information online at www.pnas.org/cgi/content/full/0904287106/DCSupplemental

(For the financial background and terminology used throughout the article we refer the reader to ref. 9.)

A portfolio is defined as a list of weights w_i, for assets $i = 1, \ldots, N$, that represent the amount of capital to be invested in each asset. We assume that one unit of capital is available and require it be fully invested, i.e., $\sum_{i=1}^{N} w_i = 1$. We collect the weights in an $N \times 1$ vector $\mathbf{w} = (w_1, \ldots, w_N)^\top$. The normalization constraint on the weights can thus be rewritten as $\mathbf{w}^\top \mathbf{1}_N = 1$, with $\mathbf{1}_N$ the $N \times 1$ vector in which every entry equals 1. The expected return and variance, for portfolio $\mathbf{w}$, are equal to $\mathbf{w}^\top \boldsymbol{\mu}$ and $\mathbf{w}^\top \boldsymbol{C}\mathbf{w}$, respectively.

In the traditional Markowitz portfolio optimization, the objective is to find a portfolio that has minimal variance for a given expected return $\rho = \mathbf{w}^\top \boldsymbol{\mu}$. More precisely, one seeks $\widetilde{\mathbf{w}}$ satisfying:

$$\widetilde{\mathbf{w}} = \arg\min_{\mathbf{w}} (\mathbf{w}^\top \boldsymbol{C}\mathbf{w}) \text{ such that } \mathbf{w}^\top \boldsymbol{\mu} = \rho, \mathbf{w}^\top \mathbf{1}_N = 1.$$

Since $\boldsymbol{C} = \mathbf{E}\big[\mathbf{r}_t \mathbf{r}_t^\top\big] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$, this minimization is equivalent to

$$\widetilde{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{E}\big[|\rho - \mathbf{w}^\top \mathbf{r}_t|^2\big] \text{ such that } \mathbf{w}^\top \boldsymbol{\mu} = \rho, \mathbf{w}^\top \mathbf{1}_N = 1.$$

For the empirical implementation, we replace expectations by sample averages. Set $\widehat{\boldsymbol{\mu}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{r}_t$; define $\boldsymbol{R}$ as the $T \times N$ matrix of which row t equals $\mathbf{r}_t^\top$, that is, $\boldsymbol{R}_{t,i} = (\mathbf{r}_t)_i = r_{i,t}$. Given this notation, we thus have the following optimization problem

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{T} \|\rho \mathbf{1}_T - \boldsymbol{R}\mathbf{w}\|_2^2 \text{ s. t. } \mathbf{w}^\top \widehat{\boldsymbol{\mu}} = \rho, \mathbf{w}^\top \mathbf{1}_N = 1, \qquad [1]$$

where, for a vector $\mathbf{a}$ in $\mathbb{R}^T$, we denote by $\|\mathbf{a}\|_2^2$ the sum $\sum_{t=1}^{T} \mathbf{a}_t^2$.

This problem requires the solution of a constrained multivariate regression involving many potentially collinear variables. Although this problem is analytically simple, it can be quite challenging in practice, depending on the nature of the matrix $\boldsymbol{R}$. Specifically, the condition number—defined to be the ratio of the largest to smallest singular values of a matrix—of $\boldsymbol{R}$ can effectively summarize the difficulty we will face when trying to perform this optimization in a stable way. When the condition number of $\boldsymbol{R}$ is small, the problem is numerically stable and easy to solve. However, when the condition number is large, a nonregularized numerical procedure will amplify the effects of noise, leading to an unstable and unreliable estimate of the vector $\mathbf{w}$. As asset returns tend to be highly correlated, the smallest singular value of $\boldsymbol{R}$ can be quite small, leading to a very large condition number and thus very unstable optimizations in a financial context. It is this sort of instability that likely plagues many of the algorithms reviewed in ref. 2.

To obtain meaningful, stable results for such ill-conditioned problems, one typically adopts a *regularization* procedure. One fairly standard approach is to augment the objective function of interest with a penalty term, which can take many forms and ideally should have a meaningful interpretation in terms of the specific problem at hand. We propose here to add an ℓ_1 penalty to the original Markowitz objective function in Eq. **1**. We thus seek to find a vector of portfolio weights $\mathbf{w}$ that solves

$$\mathbf{w}^{[\tau]} = \arg\min_{\mathbf{w}} \big[\|\rho \mathbf{1}_T - \boldsymbol{R}\mathbf{w}\|_2^2 + \tau \|\mathbf{w}\|_1\big] \qquad [2]$$

$$\text{such that } \mathbf{w}^\top \widehat{\boldsymbol{\mu}} = \rho \qquad [3]$$

$$\mathbf{w}^\top \mathbf{1}_N = 1. \qquad [4]$$

Here, the ℓ_1 norm of a vector $\mathbf{w}$ in $\mathbb{R}^N$ is defined by $\|\mathbf{w}\|_1 := \sum_{i=1}^{N} |w_i|$, and τ is a parameter that allows us to adjust the relative importance of the ℓ_1 penalization in our optimization. Note that we absorbed the factor $1/T$ from Eq. **1** in the parameter τ. The particular problem of minimizing an (unconstrained) objective function of the type given by Eq. **2** was named *lasso regression* in ref. 10.

Adding an ℓ_1 penalty to the objective function in Eq. **1** has several useful consequences:

- It promotes *sparsity*. The sparsifying effect arising from penalizing or minimizing ℓ_1 norms has long been observed in statistics (see e.g., ref. 11 and references therein). Minimization of ℓ_1-penalized objective functions is now a widely used technique when sparse solutions are desirable. Sparsity should also play a key role in the task of formulating investment portfolios: investors frequently want to be able to limit the number of positions they must create, monitor, and liquidate. By considering suitably large values of τ in Eq. **2**, one can achieve just such an effect within our framework.
- It regulates the amount of *shorting* in the portfolio designed by the optimization process. Because of the constraint **4**, an equivalent form of the objective function in Eq. **2** is

$$\|\rho \mathbf{1}_T - \boldsymbol{R}\mathbf{w}\|_2^2 + 2\tau \sum_{i \text{ with } w_i < 0} |w_i| + \tau, \qquad [5]$$

in which the last term is, of course, irrelevant for the optimization process. Under the constraint **4**, the ℓ_1 penalty is thus equivalent to a penalty on short positions. The no-short-positions optimal portfolio, obtained by solving **1** under the *three* constraints given not only by Eq. **3** and Eq. **4**, but also the additional restriction $w_i \geqslant 0$ for $i = 1, \ldots, N$, is in fact the optimal portfolio for Eq. **5** in the limit of extremely large values of τ. As the high τ limit of a sparsity-promoting framework, it is completely natural that the optimal no-short-positions portfolio should be quite sparse, as indeed also observed in practice (see below). We note that the literature has focused on the stability of positive solutions, but seems to have overlooked the sparsity of such solutions. This may possibly be due to the use of iterative numerical optimization algorithms and stopping criteria that halt the optimization before most of the components have converged to their zero limit. By decreasing τ in the ℓ_1-penalized objective function, one relaxes the constraint without removing it completely; it then no longer imposes positivity absolutely, but still penalizes overly large negative weights.
- It *stabilizes* the problem. By imposing a penalty on the size of the coefficients of $\mathbf{w}$ in an appropriate way, we reduce the sensitivity of the optimization to the possible collinearities between the assets. In ref. 5, it is proved (for the unconstrained case) that any ℓ_p penalty on $\mathbf{w}$, with $1 \leqslant p \leqslant 2$, suffices to stabilize the minimization of Eq. **1** by regularizing the inverse problem. The stability induced by the ℓ_1 penalization is extremely important; indeed, it is such stability property that makes practical, empirical work possible with only limited training data. For example, ref. 12 shows that this regularization method can be used to produce accurate macroeconomic forecasts by using many predictors.
- It incorporates a proxy for the *transaction costs* into the minimization procedure. In addition to the choice of the securities they trade, real-world investors must also concern themselves with the transaction costs they will incur when acquiring and liquidating the positions they select. Transaction costs in a liquid market can be modeled by a two-component structure: one that is a fixed "overhead," independent of the size of the transaction, and a second one, given by multiplying the transacted amount with the marketmaker's bid–ask spread applicable to the size of the transaction.

For large investors, the overhead portion can be neglected; in that context, the total transaction cost paid is just $\sum_{i=1}^{N} s_i |w_i|$, the sum of the products of the absolute trading volumes $|w_i|$ and bid–ask spreads s_i for the securities $i = 1, \ldots, N$. We assume that the bid–ask spread is the same for all assets and constant for a wide range of transaction sizes. In that case, the transaction cost is effectively captured by the ℓ_1 norm of $\mathbf{w}$. (Our method easily generalizes to asset-dependent bid–ask spreads—see the section on possible generalizations.)

For small investors, the overhead portion of the transaction costs is nonnegligible; for a very small investor, this portion may even be the only one worth considering. If the transaction costs are asset-independent, then the total cost is simply proportional to the number K of assets selected (i.e., corresponding to nonzero weights), a number sometimes referred to as $\|\mathbf{w}\|_0$, the ℓ_0 norm of the weight vector. Like an ℓ_1 sum, this ℓ_0 sum can be incorporated into the objective function to be minimized; however, ℓ_0-penalized optimization is computationally intractable when more than a handful of variables are involved, because its complexity is essentially combinatorial in nature, and grows superexponentially with the number of variables. For this reason, one often replaces the ℓ_0 penalty, when it occurs, by its much more tractable (convex) ℓ_1 penalty cousin, which has similar sparsity-promoting properties. In this sense, our ℓ_1 penalization is thus "natural" even for small investors.

Optimization Strategy

We first quickly review the unconstrained case, i.e., the minimization of the objective function in Eq. **2**, and then discuss how to deal with constraints **3** and **4**.

Various algorithms can be used to solve problem **2**. For the values of the parameters encountered in the portfolio construction problem, a particularly convenient algorithm is the homotopy method (13, 14), also known as **L**east **A**ngle **R**egression (LARS) (15). This algorithm solves problem **2** for a range of τ, starting from a very large value, and gradually decreasing τ until the desired value is attained. As τ evolves, the optimal solution $\mathbf{w}^{[\tau]}$ moves through $\mathbb{R}^N$, on a piecewise affine path. To find the whole locus of solutions for $\mathbf{w}^{[\tau]}$ we need only find the critical points where the slope changes. These slopes are thus the only quantities that need to be computed explicitly, besides the breakpoints of the piecewise linear (vector-valued) function. For every value of τ, the entries j for which $w_j \neq 0$, are said to constitute the *active set* $\mathcal{A}_\tau$. Typically, the number of elements of $\mathcal{A}_\tau$ increases as τ decreases. However, this is not always the case: at some breakpoints, entries may need to be removed from $\mathcal{A}_\tau$ (see, e.g., ref. 15).

When the desired minimizer contains only a small number K of nonzero entries, this method is very fast: the procedure involves solving linear systems of k equations with k unknowns, k being the number of active variables, that increases until K is reached.

The homotopy/LARS algorithm applies to *unconstrained* ℓ_1-penalized regression. The problem of interest to us, however, is the minimization problem **2** *under the constraints* **3** and **4**, in which case the original algorithm does not apply. The supporting information (SI) *Appendix* shows how to modify the homotopy/LARS algorithm to deal with a general ℓ_1-penalized minimization problem with linear constraints, allowing us to find:

$$\mathbf{w}^{[\tau]} = \arg\min_{\mathbf{w}\in H}\left[\|\mathbf{y} - \boldsymbol{R}\mathbf{w}\|_2^2 + \tau\|\mathbf{w}\|_1\right] \qquad \textbf{[6]}$$

where H is a prescribed affine subspace, defined by the linear constraints. The adapted algorithm consists again of starting with large values of τ, and shrinking τ gradually until the desired value is reached, monitoring the solution, which is still piecewise linear, and solving a linear system at every breakpoint in τ. Because of the constraints, the initial solution (for large values of τ) is now more complex (in the unconstrained case, it is simply equal to zero); in addition, extra variables (Lagrange multipliers) have to be introduced that are likewise piecewise linear in τ.

In the particular case of the minimization problem **2** under the constraints **3** and **4**, an interesting interplay takes place between Eq. **4** and the ℓ_1-penalty term. When the weights w_i are all nonnegative, the constraint **4** is equivalent to setting $\|\mathbf{w}\|_1 = 1$. Given that the ℓ_1-penalty term takes on a fixed value in this case, minimizing the quadratic term only (as in Eq. **1**) is thus equivalent to minimizing the penalized objective function in Eq. **2**, *for nonnegative weights* w_i. This is consistent with the observation made in ref. 4 that a restriction to nonnegative weights only can have a regularizing effect on Markowitz's portfolio construction.

The following mathematical observations have interesting consequences. Suppose that the two weight vectors $\mathbf{w}^{[\tau_1]}$ and $\mathbf{w}^{[\tau_2]}$ are minimizers for **2**, corresponding to the values τ_1 and τ_2, respectively, and both satisfy the two constraints **3** and **4**. By using the respective minimization properties of $\mathbf{w}^{[\tau_1]}$ and $\mathbf{w}^{[\tau_2]}$, we obtain

$$\begin{aligned}
&\left\|\rho\mathbf{1}_T - \boldsymbol{R}\mathbf{w}^{[\tau_1]}\right\|_2^2 + \tau_1\left\|\mathbf{w}^{[\tau_1]}\right\|_1 \\
&\quad\leqslant \left\|\rho\mathbf{1}_T - \boldsymbol{R}\mathbf{w}^{[\tau_2]}\right\|_2^2 + \tau_1\|\mathbf{w}^{[\tau_2]}\|_1 \\
&\quad= \left\|\rho\mathbf{1}_T - \boldsymbol{R}\mathbf{w}^{[\tau_2]}\right\|_2^2 + \tau_2\left\|\mathbf{w}^{[\tau_2]}\right\|_1 + (\tau_1 - \tau_2)\left\|\mathbf{w}^{[\tau_2]}\right\|_1 \\
&\quad\leqslant \left\|\rho\mathbf{1}_T - \boldsymbol{R}\mathbf{w}^{[\tau_1]}\right\|_2^2 + \tau_2\left\|\mathbf{w}^{[\tau_1]}\right\|_1 + (\tau_1 - \tau_2)\|\mathbf{w}^{[\tau_2]}\|_1 \\
&\quad= \left\|\rho\mathbf{1}_T - \boldsymbol{R}\mathbf{w}^{[\tau_1]}\right\|_2^2 + \tau_1\left\|\mathbf{w}^{[\tau_1]}\right\|_1 + (\tau_1 - \tau_2)\left(\left\|\mathbf{w}^{[\tau_2]}\right\|_1 - \left\|\mathbf{w}^{[\tau_1]}\right\|_1\right),
\end{aligned}$$

which implies that

$$(\tau_1 - \tau_2)\left(\left\|\mathbf{w}^{[\tau_2]}\right\|_1 - \left\|\mathbf{w}^{[\tau_1]}\right\|_1\right) \geqslant 0. \qquad \textbf{[7]}$$

If all the $w_i^{[\tau_1]}$ are nonnegative, but some of the $w_i^{[\tau_2]}$ are negative, then we have $\|\mathbf{w}^{[\tau_2]}\|_1 > |\sum_{i=1}^N w_i^{[\tau_2]}| = 1$ and $\|\mathbf{w}^{[\tau_1]}\|_1 = 1$, implying $\|\mathbf{w}^{[\tau_2]}\|_1 > \|\mathbf{w}^{[\tau_1]}\|_1$. In view of Eq. **7**, this means that $\tau_1 \geqslant \tau_2$. It follows that the optimal portfolio with nonnegative entries obtained by our minimization procedure corresponds to the *largest* values of τ, and thus typically to the *sparsest* solution (since the penalty term, promoting sparsity, is weighted more heavily). This particular portfolio is a minimizer for **2**, under the constraints **3** and **4**, for all τ larger than some critical value τ_0. For smaller τ the optimal portfolio will contain at least one negative weight and will typically become less sparse. However, as in the unconstrained case, this need not happen in a monotone fashion.

Although other optimization methods could be used to compute the sparse portfolios we define, the motivation behind our choice of a constrained homotopy/LARS algorithm is the fact that we are only interested in computing portfolios involving a small number of securities and that we use the parameter τ to tune this number. Whereas other algorithms would require separate computations to find solutions for each value of τ, a particularly nice feature of our LARS-based algorithm is that, by exploiting the piecewise linear dependence of the solution on τ, it obtains, in one run, the weight vectors for all values of τ (i.e., for all numbers of selected assets) in a prescribed range.

Another strategy often used in constrained least-squares optimization consists in reparametrizing the variables so as to automatically satisfy the constraints; we chose not to do this, because this would mess up the ℓ_1 penalty.

Empirical Application

In this section we apply the methodology described above to construct optimal portfolios and evaluate their out-of-sample performance. We present two examples, each of which uses a universe of investments compiled by Fama and French. In the first example, we use 48 industry sector portfolios (abbreviated to FF48 in the remainder of this article). In the second example, we use 100 portfolios formed on size and book-to-market (FF100). (These portfolios are the intersections of 10 portfolios formed on size and 10 portfolios formed on the ratio of book equity to market equity.) In both FF48 and FF100, the portfolios are constructed at the end of June in their construction year. (See below for details.)

Example 1: FF48. By use of the above notation, $r_{i,t}$ is the annualized return in month t of industry i, where $i = 1, \ldots, 48$. We evaluate our methodology by looking at the out-of-sample performances of

Evaluation	Eq. weight			$w_i \geqslant 0$			bin=8–16			bin=17–24			bin=25–32			bin=33–40			bin=41–48		
period	m		S	m		S	m		S	m		S	m		S	m		S	m		S
07/76–06/06	17	61	27	17	41	41	16	40	40	14	40	34	12	43	28	12	47	26	12	54	22
07/76–06/81	29	66	44	23	48	49	20	41	50	18	39	46	19	40	49	22	43	50	23	50	46
07/81–06/86	18	58	31	23	41	57	25	42	58	23	44	52	24	46	52	23	50	46	22	56	39
07/86–06/91	5	72	7	9	45	20	8	43	18	7	44	15	4	47	9	4	51	7	5	57	8
07/91–06/96	18	41	44	16	26	62	15	26	57	13	27	47	12	33	36	12	41	30	11	52	21
07/96–06/01	11	67	17	16	40	40	16	41	38	9	42	22	3	50	6	2	54	4	0	61	0
07/01–06/06	18	60	30	13	43	30	13	42	30	12	41	29	10	38	28	10	37	27	12	43	27

Evaluation	Eq. weight			$w_i \geqslant 0$			bin=11–20			bin=21–30			bin=31–40			bin=41–50			bin=51–60		
period	m		S	m		S	m		S	m		S	m		S	m		S	m		S
06/76–06/06	17	59	28	16	53	30	16	50	33	19	48	39	19	49	40	20	52	39	21	60	34
07/76–06/81	23	61	38	12	59	21	11	52	22	12	51	24	12	55	22	11	56	20	7	66	10
07/81–06/86	20	53	38	24	49	49	26	41	64	31	38	81	31	40	77	31	43	72	33	49	67
07/86–06/91	9	71	13	10	65	15	9	63	14	9	61	16	11	62	18	12	64	19	12	71	17
07/91–06/96	18	34	53	19	31	61	20	29	70	22	25	86	20	28	73	22	31	70	25	36	67
07/96–06/01	16	63	26	18	52	35	18	53	35	23	52	44	29	47	61	31	50	62	34	63	54
07/01–06/06	12	64	19	11	55	21	11	53	22	15	51	29	13	51	26	13	56	23	14	64	22

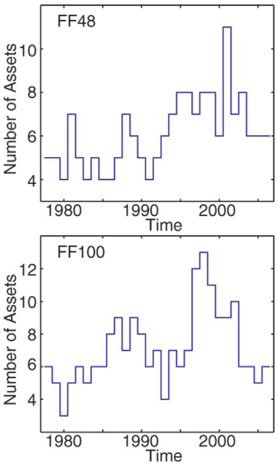

Fig. 1. Empirical results for FF48 (*Upper*) and FF100 (*Lower*). For each of the two examples, the table on the left lists the monthly mean return m, standard deviation of monthly return σ, and corresponding monthly Sharpe ratio S (expressed in %), for the optimal portfolios with equal weights for the N assets, for binned portfolios, and for the optimal portfolio without short positions. The figures on the right show, for both examples, the number of assets $K_{\text{pos.}}$ in $\mathbf{w}^{\text{pos.}}$, the optimal portfolio without short positions, from year to year.

our portfolios during the past 30 years in a simulated investment exercise.

For each year from 1976 to 2006, we construct a collection of optimal portfolios by solving an ensemble of minimizations of the objective function in Eq. **2** with constraints **3** and **4**. For each time period, we carry out our optimization for a sufficiently wide range of τ to produce an ensemble of portfolios containing different numbers of active positions; ideally, we would like to construct portfolios with K securities, for all values of K between 2 and 48. As explained below, we do not always obtain all the low values of K; typically, we find optimal portfolios only for K exceeding a minimal value $K_{\min}$, that varies from year to year (Fig. 1). To estimate the necessary return and covariance parameters, we use data from the preceding 5 years (60 months). At the time of each portfolio construction, we set the target return, ρ, to be the average return achieved by the naïve, evenly weighted portfolio over the previous 5 years.

For example, our first portfolio construction takes place at the end of June 1976. To determine $\boldsymbol{R}$ and $\widehat{\boldsymbol{\mu}}$, we use the historical returns from July 1971 until June 1976. We then solve the optimization problem by using this matrix and vector, targeting an annualized return of 6.60% ($\rho = 0.066$), equal to the average historical return, from July 1971 until June 1976, obtained by a portfolio in which all industry sectors are given the equal weight 1/48. We compute the weights of optimal solutions $\mathbf{w}^{[\tau]}$ for τ ranging from large to small values. We select these portfolios according to some criterion we would like to meet. We could, e.g., target a fixed total number of active positions, or limit the number of short positions; see below for examples. Once a portfolio is thus fixed, it is kept from July 1976 until June 1977, and its returns are recorded. At the end of June 1977, we repeat the same process, using training data from July 1972 to June 1977 to compute the composition of a new collection of portfolios. These portfolios are observed from July 1977 until June 1978 and their returns are recorded. The same exercise is repeated every year with the last ensemble of portfolios constructed at the end of June 2005.

Once constructed, the portfolios are thus held through June of the next year and their monthly out-of-sample returns are observed. These monthly returns, for all the observation years together, constitute a time series; for a given period (whether it is the full 1976–2006 period, or subperiods), all the monthly returns corresponding to this period are used to compute the average monthly return m, its standard deviation σ, and their ratio m/σ, which is then the Sharpe ratio measuring the trade-off, corresponding to the period, between returns and volatility of the constructed portfolios.

We emphasize that the *sole purpose* of carrying out the portfolio construction multiple times, in successive years, is to collect data from which to evaluate the effectiveness of the portfolio construction strategy. These constructions from scratch in consecutive years are *not* meant to model the behavior of a single investor; they model, rather, the results obtained by *different* investors who would follow the same strategy to build their portfolio, starting in different years. A single investor might construct a starting portfolio according to the strategy described here, but might then, in subsequent years, adopt a sparse portfolio adjustment strategy such as described in the next section.

We compare the performance of our strategy with that of a benchmark strategy constituting an equal investment in each available security. This $1/N$ strategy is a tough benchmark because it has been shown to outperform a host of optimal portfolio strategies constructed with existing optimization procedures (2). To evaluate the $1/N$ strategy portfolios for the FF48 assets, we likewise observe the monthly returns for a certain period (a 5-year break-out period or the full 30-year period), and use them to compute the average mean return m, the standard deviation σ, and the Sharpe ratio m/σ.

We carried out the full procedure using several possible guidelines. The first such guideline is to pick the optimal portfolio $\mathbf{w}^{\text{pos.}}$ that has only nonnegative weights w_i, i.e., the optimal portfolio without short positions. As shown in the previous section, this portfolio corresponds to the largest values of the penalization constant τ; it typically is also the optimal portfolio with the fewest assets. Fig. 1 reports the number of active assets of this optimal no-short-positions portfolio from year to year. This number varies from a minimum of 4 to a maximum of 11; note that this is quite sparse in a 48-asset universe. The top table in Fig. 1 reports statistics to evaluate the performances of the optimal no-short-positions portfolio. We give the statistics for the whole sample period and for consecutive subperiods extending over 5 years each, comparing these with the portfolio that gives equal weight to the 48 assets. The table shows that the optimal no-short-positions portfolio significantly outperforms the benchmark both in terms of returns and in terms of volatility; this result holds for the full sample period as well as for the subperiods. Note that most of the gain comes from the smaller variance of the sparse portfolio around its target return, ρ.

A second possible guideline for selecting the portfolio construction strategy is to target a particular number of assets, or a particular narrow range for this number. For instance, users could decide to pick, every year, the optimal portfolio that has always more than 8 but at most 16 assets. Or the investor may decide to select an

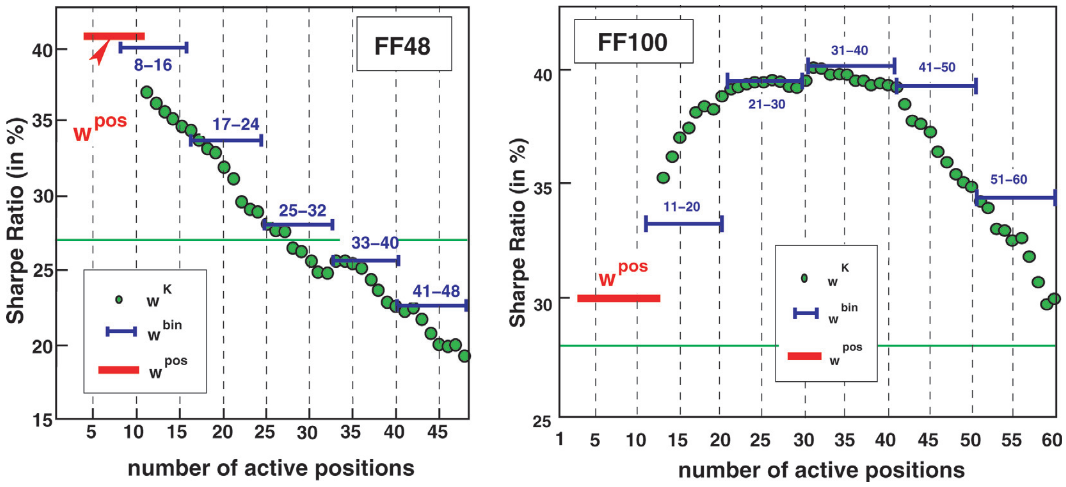

Fig. 2. Shown are the FF48 example (*Left*) and the FF100 example (*Right*), the Sharpe ratio, for the full period 1976–2006, for several portfolios: $\boldsymbol{w}^{\text{pos.}}$, the optimal portfolio without short positions (red); the "binned" portfolios (blue, the extent of the line covers the bin width); and the portfolios $\boldsymbol{w}^K$ with a fixed number K of active positions (green dots circled in black). In both cases $\boldsymbol{w}^{\text{pos.}}$ is indicated by a fat solid horizontal red bar, stretching from its minimum to maximum number of assets (see also Fig. 1). In both examples, optimal sparse portfolios that allow short positions significantly outperform the evenly weighted portfolio; in the FF100 case, they also significantly outperform the optimal no-short-positions portfolio.

optimal portfolio with, say, exactly 13 assets. In this case, we would carry out the minimization, decreasing τ until we reach the breakpoint value where the number of assets in the portfolio reaches 13. We shall denote the corresponding weight vector by $\mathbf{w}^{13}$.

For a "binned" portfolio, such as the 8-to-16 asset portfolio, targeting a narrow range rather than an exact value for the total number of assets, we define the portfolio $\mathbf{w}^{8-16}$ by considering each year the portfolios $\mathbf{w}^K$ with K between 8 and 16 (both extremes included), and selecting the one that minimizes the objective function in Eq. **1**; if there are several possibilities, the minimizer with smallest ℓ_1 norm is selected. The results are summarized in Fig. 2, which shows the average monthly Sharpe ratio of different portfolios of this type for the entire 30-year exercise. For several portfolio sizes, we are able to significantly outperform the evenly weighted portfolio (the Sharpe ratio of which is indicated by the horizontal line at 27%). Detailed statistics are reported in the upper table in Fig. 1.

Notice that, according to this table, the no-short-positions portfolio outperforms all binned portfolios for the full 30-year period; this is not systematically true for the breakout periods, but even in those breakout periods where it fails to outperform all binned portfolios, its performance is still close to that of the best performing (and sparsest) of the binned portfolios. This observation no longer holds for the portfolio constructions with FF100, our second exercise—see Fig. 2.

Example 2: FF100. Except for using a different collection of assets, this exercise is identical in its methodology to that of FF48, so that we do not repeat the full details here. The lower table and figure in Fig. 1, and the right-hand figure in Fig. 2 summarize the results.

From the results of our two exercises we see that:

- Our sparse portfolios (with a relatively small number of assets and moderate τ) outperform the naïve $1/N$ strategy significantly and consistently over the entire evaluation period. This gain is achieved for a wide range of portfolio sizes, as indicated in Fig. 2. Note that the best performing sparse portfolio we constructed is *not* always the no-short-positions portfolio.
- When we target a large number of assets in our portfolio, the performance deteriorates. We interpret this as a result of so-called "overfitting." Larger target numbers of assets correspond to smaller values of τ. The ℓ_1 penalty is then having only a negligible effect and the minimization focuses essentially on the variance term. Hence, the solution becomes unstable and is overly sensitive to the estimation errors that plague the original (unpenalized) Markowitz optimization problem **1**.

Numerical experiments showed that this overall behavior is quite robust with respect to the choice of the target return.

Possible Generalizations

In this section, we describe, in brief, some extensions of our approach. It should be pointed out that the relevance and usefulness of the ℓ_1 penalty is not limited to a stable implementation of the usual Markowitz portfolio selection scheme described above. Indeed, there are several other portfolio construction problems that can be cast in similar terms or otherwise solved through the minimization of a similar objective function. We now list a few examples:

Partial Index Tracking. In many situations, investors want to create a portfolio that efficiently tracks an index. In some cases, this will be an existing financial index whose level is tied to a large number of tradable securities but which is not yet tradable en masse as an index future or other single instrument. In such a situation, investors need to find a collection of securities whose profit-and-loss profile accurately tracks the index level. Such a collection need not be a full replication of the index in question; indeed, it is frequently inconvenient or impractical to maintain a full replication.

In other situations, investors will want to monetize some more abstract financial time series: an economic time series, an investor sentiment time series, etc. In that case, investors will need to find a collection of securities that is likely to remain correlated to the target time series.

Either way, the investor will have at his disposal a time series of index returns, which we will write as a $T \times 1$ column vector, $\mathbf{y}$. Also, the investor will have at his disposal the time series of returns for every available security, which we will write as a $T \times N$ matrix $\boldsymbol{R}$, as before.

In that case, an investor seeking to minimize expected tracking error would want to find

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\| \mathbf{y} - \boldsymbol{R}\mathbf{w} \right\|_2^2.$$

However, this problem is simply a linear regression of the target returns on the returns of the available assets. As the available assets may be collinear, the problem is subject to the same instabilities that we discussed above. As such, we can augment our objective function with an ℓ_1 penalty and seek instead

$$\mathbf{w}^{[\tau]} = \arg\min_{\mathbf{w}} \left[\left\| \mathbf{y} - \boldsymbol{R}\mathbf{w} \right\|_2^2 + \tau \|\mathbf{w}\|_1 \right],$$

subject to the appropriate constraints. This simple modification stabilizes the problem and enforces sparsity, so that the index can be stably replicated with few assets.

Moreover, one can enhance this objective function in light of the interpretation of the ℓ_1 term as a model of transaction costs. Let s_i is the transaction cost (bid–ask spread) for the ith security. In that case, we can seek

$$\mathbf{w}^{[\tau]} = \arg\min_{\mathbf{w}} \left[\left\| \mathbf{y} - \boldsymbol{R}\mathbf{w} \right\|_2^2 + \tau \sum_i s_i |w_i| \right].$$

By making this modification, the optimization process will "prefer" to invest in more liquid securities (low s_i) whereas it will "avoid" investments in less liquid securities (high s_i). A slightly modified version of the algorithm described above can cope with such weighted ℓ_1 penalty and generate a list of portfolios for a wide range of values for τ. For each portfolio, the investor could then compare the expected tracking error per period ($\frac{1}{T}\|\mathbf{y} - \boldsymbol{R}\mathbf{w}\|_2^2$) with the expected cost of creating and liquidating the tracking portfolio ($\sum_i s_i |w_i|$). The investor could then select a portfolio that suits both his risk tolerance and cost constraints.

Portfolio Hedging. Consider the task of hedging a given portfolio using some subset of a universe of available assets. As a concrete example, imagine trying to efficiently hedge out the market risk in a portfolio of options on a single underlying asset, potentially including many strikes and maturities. An investor would be able to trade the underlying asset and any options desired. In this context, it would be possible to completely eliminate market risk by negating the initial position. However, this may not be feasible given liquidity (transaction cost) constraints.

Instead, an investor may simply want to reduce his risk in a cost-efficient way. One could proceed as follows: Generate a list of scenarios. For each scenario, determine the change in the value of the existing portfolio. Also, determine the change in value for a unit of each available security. Store the former in a $M \times 1$ column vector $\mathbf{y}$ and store the latter in a $M \times N$ matrix, $\boldsymbol{X}$. Also, determine a probability, p_i for $i = 1, \ldots, M$ of each scenario, and store *the square root of these values* in a diagonal $M \times M$ matrix, $\boldsymbol{P}$. These probabilities can be derived from the market or assumed subjectively according to an investor's preference. As before, denoting by s_i the transaction costs for each tradable security, we can seek

$$\mathbf{w}^{[\tau]} = \arg\min_{\mathbf{w}} \left[\|\boldsymbol{P}(\mathbf{y} + \boldsymbol{X}\mathbf{w})\|_2^2 + \tau \sum_i s_i |w_i| \right].$$

As before, the investor could then apply one of the algorithms above to generate a list of optimal portfolios for a wide range of values of τ. Then, just as in the index tracking case, the investor could observe the attainable combinations of expected mark to market variance ($\|\boldsymbol{P}(\mathbf{y} + \boldsymbol{X}\mathbf{w})\|_2^2$) and transaction cost ($\sum_i s_i |w_i|$). One appealing feature of this method is that it does not explicitly determine the number of assets to be included in the hedge portfolio. The optimization naturally trades off portfolio volatility for transaction cost, rather than imposing an artificial cap on either.

Portfolio Adjustment. Thus far, we have assumed that investors start with no assets, and must construct a portfolio to perform a particular task. However, this is rarely the case in the real world. Instead, investors frequently hold a large number of securities and must modify their existing holdings to achieve a particular goal. In this context, the investor already holds a portfolio $\mathbf{w}$ and must make an adjustment $\Delta_{\mathbf{w}}$. In that case, the final portfolio will be $\mathbf{w} + \Delta_{\mathbf{w}}$, but the transaction costs will be relevant only for the adjustment $\Delta_{\mathbf{w}}$. The corresponding optimization problems is given by

$$\Delta_{\mathbf{w}}^{[\tau]} = \arg\min_{\Delta_{\mathbf{w}}} \left[\|\rho \mathbf{1}_T - \boldsymbol{R}(\mathbf{w} + \Delta_{\mathbf{w}})\|_2^2 + \tau \|\Delta_{\mathbf{w}}\|_1 \right]$$
$$\text{s. t.} \quad \Delta_{\mathbf{w}}^\top \widehat{\boldsymbol{\mu}} = 0 \quad \text{and} \quad \Delta_{\mathbf{w}}^\top \mathbf{1}_N = 0.$$

It is easy to modify our methodology to handle this situation.

Conclusion

We have devised a method that constructs stable and sparse portfolios by introducing an ℓ_1 penalty in the Markowitz portfolio optimization. We obtain as special cases the no-short-positions portfolios that also comprise few active assets. To our knowledge, such a sparsity property of the nonnegative portfolios has not been previously noticed in the literature. The portfolios we propose can be seen as natural extensions of the no-short-positions portfolios and maintain or improve their performances while preserving their sparse nature as much as possible.

We have also described an efficient algorithm for computing the optimal, sparse portfolios, and we have implemented it using as assets two sets of portfolios constructed by Fama and French: 48 industry portfolios and 100 portfolios formed on size and book-to-market. We found empirical evidence that the optimal sparse portfolios outperform the evenly weighted portfolios by achieving a smaller variance; moreover, they do so with only a small number of active positions, and the effect is observed over a range of values for this number. This shows that adding an ℓ_1 penalty to objective functions is a powerful tool for various portfolio construction tasks. This penalty forces our optimization scheme to select, on the basis of the training data, few assets forming a stable and robust portfolio, rather than being "distracted" by the instabilities because of collinearities and responsible for meaningless artifacts in the presence of estimation errors.

Many variants and improvements are possible on the simple procedure described and illustrated above. This goes beyond the scope of the present article which was to propose a methodology and to demonstrate its validity.

ACKNOWLEDGMENTS. We thank Tony Berrada, Laura Coroneo, Simone Manganelli, Sergio Pastorello, Lucrezia Reichlin, and Olivier Scaillet for helpful suggestions and discussions. J. B. thanks Bobray Bordelon, Jian Bai, Josko Plazonic, and John Vincent for their technical assistance. This research was supported, in part, by "Action de Recherche Concertée" 02/07-281 (C.D.M., D.G.), the Francqui Foundation (I.L.), the "Geconcerteerde Onderzoeksactie" 09/07-62 (I.D., C.D.M., I.L.), the National Bank of Belgium (D.G., C.D.M.), and by the National Science Foundation Grant DMS-0354464 (to I.D.).

1. Markowitz H (1952) Portfolio selection. *J Finance* 7:77–91.
2. DeMiguel V, Garlappi L, Uppal R (2007) Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Rev Financial Stud*, preprint.
3. Bertero M, Boccacci P (1998) *Introduction to Inverse Problems in Imaging* (Institute of Physics Publishing, London).
4. Jagannathan R, Ma T (2003) Risk reduction in large portfolios: Why imposing the wrong constraints helps. *J Finance* 58:1651–1684.
5. Daubechies I, Defrise M, De Mol C (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun Pure Appl Math* 57:1416–1457.
6. DeMiguel V, Garlappi L, Nogales FJ, Uppal R (2009) A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Manage Sci* 55:798–812.
7. Brodie J, Daubechies I, De Mol C, Giannone D (2007) Sparse and stable Markowitz portfolios, preprint arXiv:0708.0046v1; http://arxiv.org/abs/0708.0046.
8. Lauprete GJ (2001) Portfolio risk minimization under departures from normality. PhD thesis (Massachusetts Institute of Technology, Cambridge, MA).
9. Campbell JY, Lo AW, MacKinlay CA (1997) *The Econometrics of Financial Markets* (Princeton Univ Press, Princeton).
10. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B* 58:267–288.
11. Chen SS, Donoho D, Saunders M (2001) Atomic decomposition by basis pursuit. *SIAM Rev* 43:129–159.
12. De Mol C, Giannone D, Reichlin L (2008) Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *J Economet* 146:318–328.
13. Osborne MR, Presnell B, Turlach BA (2000) A new approach to variable selection in least squares problems. *IMA J Numer Anal* 20:389–403.
14. Osborne MR, Presnell B, Turlach BA (2000) On the Lasso and its dual. *J Comput Graphical Stat* 9:319–337.
15. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. *Ann Stat* 32:407–499.

A Constrained minimization algorithm

Before discussing our solution method for the linearly constrained ℓ_1-penalized least-squares problem, we briefly recall the homotopy/LARS method which manages to recover the unconstrained minimizer of the ℓ_1-penalized least-squares objective function

$$\bar{\mathbf{w}}(\tau) = \arg\min_{\mathbf{w}} \left[\|\boldsymbol{R}\mathbf{w} - \mathbf{y}\|_2^2 + \tau\|\mathbf{w}\|_1\right]$$

for a whole range of values of the (positive) penalty parameter τ.

The variational equations describing the minimizer $\bar{\mathbf{w}}(\tau)$ are:

$$\begin{aligned} (\boldsymbol{R}^\top(\mathbf{y} - \boldsymbol{R}\mathbf{w}))_i &= \frac{\tau}{2}\operatorname{sgn} w_i & w_i &\neq 0 & (1)\\ |(\boldsymbol{R}^\top(\mathbf{y} - \boldsymbol{R}\mathbf{w}))_i| &\leq \frac{\tau}{2} & w_i &= 0. & (2) \end{aligned}$$

The minimizer $\bar{\mathbf{w}}(\tau)$ is a continuous piecewise linear function of τ. We shall denote the breakpoints by $\tau_0 > \tau_1 > \dots$ and the corresponding minimizers by $\bar{\mathbf{w}}(\tau_0), \bar{\mathbf{w}}(\tau_1), \dots$ The breakpoints occur where a new component enters or leaves the support of $\bar{\mathbf{w}}(\tau)$. We will use $\mathbf{b}$ to denote the residual $\mathbf{b} = \boldsymbol{R}^\top(\mathbf{y} - \boldsymbol{R}\mathbf{w})$.

The homotopy/LARS method for solving these equations starts by considering the point $\mathbf{w} = 0$, which satisfies the equations (1,2) for all $\tau \geq \tau_0 \equiv 2\max_i |(\boldsymbol{R}^\top\mathbf{y})_i|$. Hence $\bar{\mathbf{w}}(\tau \geq \tau_0) = 0$.

Given a breakpoint $\bar{\mathbf{w}}(\tau_n)$, it is possible to construct the next breakpoint $\bar{\mathbf{w}}(\tau_{n+1})$ by solving a small linear system. Let $J = \{i \text{ for which } |\mathbf{b}_i| = \tau_n/2\}$ (i.e. the set of maximal residual), $\boldsymbol{R}_J$ the submatrix consisting of the columns J of $\boldsymbol{R}$. We define the walking direction $\mathbf{u}$ by

$$\boldsymbol{R}_J^\top \boldsymbol{R}_J\, \mathbf{u}_J = \operatorname{sgn}(\mathbf{b}_J)$$

and $u_i = 0$ for $i \notin J$ ($\operatorname{sgn}(\mathbf{b}_J)$ denotes the vector $(\operatorname{sgn}(b_j)_{j\in J})$). In this way, a step $\mathbf{w} \to \mathbf{w} + \gamma\mathbf{u}$ results in a change in the residual $\mathbf{b} \to \mathbf{b} - \gamma\mathbf{v}$, where $v_j = \operatorname{sgn}(b_j)$ for $j \in J$. In other words, the maximal components of the residual decrease at the same rate. The step size $\gamma > 0$ is now determined to be the smallest number for which the absolute value of a component $|b_i - \gamma v_i|$ (with $i \notin J$) of the new residual becomes equal to $|b_j - \gamma v_j|$ for $j \in J$ (i.e. a new component joins the maximal residual set), or for which a nonzero component of $\mathbf{w}$ is turned into zero.

The new penalty parameter is then $\tau_{n+1} = \tau_n - 2\gamma$ (which is smaller than τ_n), and the corresponding minimizer is $\bar{\mathbf{w}}(\tau_{n+1}) = \bar{\mathbf{w}}(\tau_n) + \gamma\mathbf{u}$. By construction it is guaranteed to satisfy the variational equations (1,2).

The two main advantages of this method are thus that it is exact (in particular zero components are really zero) and that it yields the breakpoints (and hence the minimizers) for a whole range of values of the penalization parameters $\tau \geq \tau_{\text{stop}} \geq 0$. At each step, only a relatively small linear system has to be solved. If this procedure is carried through until the end, one finds $\lim_{\tau\to 0} \arg\min \|\boldsymbol{R}\mathbf{w} - \mathbf{y}\|_2^2 + \tau\|\mathbf{w}\|_1 = \arg \min\limits_{\mathbf{w} \text{ s.t. } \boldsymbol{R}^\top\boldsymbol{R}\mathbf{w} = \boldsymbol{R}^\top\mathbf{y}} \|\mathbf{w}\|_1$.

For the constrained case, i.e. the minimization problem

$$\widetilde{\mathbf{w}}(\tau) = \mathop{\arg\min}_{\mathbf{w} \text{ s.t. } \boldsymbol{A}\mathbf{w} = \mathbf{a}} \left[\|\boldsymbol{R}\mathbf{w} - \mathbf{y}\|_2^2 + \tau\|\mathbf{w}\|_1\right] \tag{3}$$

subject to the linear constraint $\boldsymbol{A}\mathbf{w} = \mathbf{a}$, we can devise a similar procedure. We assume, of course, that the constraint $\boldsymbol{A}\mathbf{w} = \mathbf{a}$ has a solution.

An approximation of the minimizer $\widetilde{\mathbf{w}}(\tau)$ can be obtained by applying the unconstrained procedure described above to the objective function

$$\widetilde{\widetilde{\mathbf{w}}}(\tau_\epsilon) = \arg\min_{\mathbf{w}} \left[\|\boldsymbol{A}\mathbf{w} - \mathbf{a}\|_2^2 + \epsilon\|\boldsymbol{R}\mathbf{w} - \mathbf{y}\|_2^2 + \tau_\epsilon\|\mathbf{w}\|_1\right]. \tag{4}$$

For sufficiently small ϵ, this will give a good approximation of the constrained minimizer $\widetilde{\mathbf{w}}(\tau)$ corresponding to the penalty $\tau = \tau_\epsilon/\epsilon$ (after first going through a number of breakpoints for which $\boldsymbol{A}\mathbf{w} \neq \mathbf{a}$, not even approximately). However, this is clearly an approximate method (often very good) whereas the unconstrained procedure did not involve any approximation.

We solve this issue, and provide an exact method, by solving the minimization problem (4) up to the first order in ϵ. In this approach ϵ is a small *formal* positive parameter. Now the minimizer $\widetilde{\widetilde{\mathbf{w}}}(\tau_\epsilon)$ and τ_ϵ both depend on ϵ. We can write $\mathbf{w} = \mathbf{w}^{(0)} + \epsilon\mathbf{w}^{(1)} + \mathcal{O}(\epsilon^2)$ and $\tau_\epsilon = \tau^{(0)} + \tau^{(1)}\epsilon + \mathcal{O}(\epsilon^2)$. We again follow the procedure for the unconstrained method, but take care to use arithmetic (addition, multiplication, comparison, ...) up to first order in ϵ.

As before, one starts from $\mathbf{w} = 0$, corresponding to a large initial value of τ_ϵ, and follows the path of descending τ_ϵ. The strategy consists of satisfying the variational equations

$$\begin{aligned} \left(\boldsymbol{A}^\top(\mathbf{a} - \boldsymbol{A}\mathbf{w}) + \epsilon \boldsymbol{R}^\top(\mathbf{y} - \boldsymbol{R}\mathbf{w})\right)_i &= \frac{\tau_\epsilon}{2} \operatorname{sgn} w_i \qquad\qquad w_i \neq 0 \qquad (5)\\ \left|\left(\boldsymbol{A}^\top(\mathbf{a} - \boldsymbol{A}\mathbf{w}) + \epsilon \boldsymbol{R}^\top(\mathbf{y} - \boldsymbol{R}\mathbf{w})\right)_i\right| &\leq \frac{\tau_\epsilon}{2} \qquad\qquad\qquad\quad w_i = 0 \qquad (6) \end{aligned}$$

at each breakpoint by carefully determining a walking direction $\mathbf{u} = \mathbf{u}^{(0)} + \mathbf{u}^{(1)}\epsilon + \mathcal{O}(\epsilon^2)$ and a step length $\gamma = \gamma^{(0)} + \gamma^{(1)}\epsilon + \mathcal{O}(\epsilon^2)$. Using $\mathbf{w} = \mathbf{w}^{(0)} + \epsilon\mathbf{w}^{(1)} + \mathcal{O}(\epsilon^2)$, we can rewrite equations (5) as

$$\begin{aligned} \left(\boldsymbol{A}^\top(\mathbf{a} - \boldsymbol{A}\mathbf{w}^{(0)})\right)_i &= \frac{\tau^{(0)}}{2} \operatorname{sgn} w_i \qquad (7)\\ \left(-\boldsymbol{A}^\top\boldsymbol{A}\mathbf{w}^{(1)} + \boldsymbol{R}^\top(\mathbf{y} - \boldsymbol{R}\mathbf{w}^{(0)})\right)_i &= \frac{\tau^{(1)}}{2} \operatorname{sgn} w_i\,. \qquad (8) \end{aligned}$$

From a known breakpoint $\mathbf{w}$ we can proceed to the following breakpoint by a step direction $\mathbf{u}$ and step size γ (both depending on ϵ). We again set

$$J = \arg\max_i \left|\left(\boldsymbol{A}^\top(\mathbf{a} - \boldsymbol{A}\mathbf{w}^{(0)}) + \epsilon(-\boldsymbol{A}^\top\boldsymbol{A}\mathbf{w}^{(1)} + \boldsymbol{R}^\top(\mathbf{y} - \boldsymbol{R}\mathbf{w}^{(0)}))\right)_i\right|.$$

As long as $\tau^{(0)} \neq 0$, the components J of $\mathbf{u}$ are determined by

$$\begin{pmatrix} \boldsymbol{R}_J^\top\boldsymbol{R}_J & \boldsymbol{A}_J^\top\boldsymbol{A}_J \\ \boldsymbol{A}_J^\top\boldsymbol{A}_J & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_J^{(0)} \\ \mathbf{u}_J^{(1)} \end{pmatrix} = \begin{pmatrix} 0 \\ \operatorname{sgn}(\mathbf{b}_J) \end{pmatrix} \qquad (9)$$

and the other components of $\mathbf{u}$ remain zero. The step size γ is again determined as before, i.e. when a new component enters the maximal residual set, or when a component leaves the active set. The penalty parameter τ_ϵ decreases as before: $\tau_\epsilon \to \tau_\epsilon - 2\gamma$.

At some point in this procedure, τ_ϵ will become zero in zeroth order: $\tau_\epsilon = 0 + \tau^{(1)}\epsilon + \mathcal{O}(\epsilon^2)$. The corresponding minimizer (more precisely the zeroth-order part of this breakpoint) will satisfy the constraint $\boldsymbol{A}\mathbf{w} = \mathbf{a}$ and we will have found the first constrained minimizer $\widetilde{\mathbf{w}}$ of (3), corresponding to $\tau_0 = \tau^{(1)}$ (i.e. the first-order part of the parameter τ_ϵ of the ϵ-dependent problem at this breakpoint). In the unconstrained case, no such calculations were necessary as the starting point was always equal to 0. Similarly to the unconstrained case, we have that $\widetilde{\mathbf{w}}(\tau > \tau_0) = \widetilde{\mathbf{w}}(\tau_0)$.

In principle, one could continue the ϵ-dependent algorithm, but now that the first breakpoint of $\widetilde{\mathbf{w}}(\tau)$ is determined, it is more advantageous to continue the descent of τ by introducing Lagrange multipliers $\boldsymbol{\lambda}$ for the problem (3):

$$\widetilde{\mathbf{w}}(\tau) = \mathop{\arg\min}_{\boldsymbol{\lambda},\ \mathbf{w}\text{ s.t. }\boldsymbol{A}\mathbf{w}=\mathbf{a}} \left[\|\boldsymbol{R}\mathbf{w} - \mathbf{y}\|_2^2 + \tau\|\mathbf{w}\|_1 + 2\boldsymbol{\lambda}^\top(\boldsymbol{A}\mathbf{w} - \mathbf{a})\right].$$

This minimization problem (analogous to (3)) amounts to solving the equations:

$$\begin{aligned} (\boldsymbol{R}^\top(\mathbf{y} - \boldsymbol{R}\mathbf{w}) + \boldsymbol{A}^\top\boldsymbol{\lambda})_i &= \frac{\tau}{2} \operatorname{sgn} w_i \qquad\qquad w_i \neq 0 \qquad (10)\\ |(\boldsymbol{R}^\top(\mathbf{y} - \boldsymbol{R}\mathbf{w}) + \boldsymbol{A}^\top\boldsymbol{\lambda})_i| &\leq \frac{\tau}{2} \qquad\qquad\qquad\quad w_i = 0 \qquad (11)\\ \boldsymbol{A}\mathbf{w} &= \mathbf{a}. \qquad (12) \end{aligned}$$

Equation (10) is the equivalent of equation (8) whereas equation (12) replaces equation (7). We now already have τ_0, $\widetilde{\mathbf{w}}(\tau \geq \tau^{(0)})$ and the initial Lagrange multipliers $\boldsymbol{\lambda} = -\boldsymbol{A}\widetilde{\widetilde{\mathbf{w}}}^{(1)}$ (from the first-order part of the last step of the ϵ-dependent problem).

To proceed from one breakpoint to the next ($\mathbf{w} \to \mathbf{w} + \gamma\mathbf{u}$, and $\boldsymbol{\lambda} \to \boldsymbol{\lambda} + \gamma\mathbf{s}$ as the multipliers also change), we again need to solve a linear system:

$$\begin{pmatrix} \boldsymbol{R}_J^\top\boldsymbol{R}_J & \boldsymbol{A}_J^\top \\ \boldsymbol{A}_J & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_J \\ \mathbf{s} \end{pmatrix} = \begin{pmatrix} \operatorname{sgn}(\tilde{\mathbf{b}}_J) \\ 0 \end{pmatrix} \qquad (13)$$

with $\tilde{\mathbf{b}} = \boldsymbol{R}^\top(\mathbf{y} - \boldsymbol{R}\mathbf{w}) + \boldsymbol{A}^\top\boldsymbol{\lambda}$. This will guarantee that $\mathbf{w} \to \mathbf{w} + \gamma\mathbf{u}$ and $\boldsymbol{\lambda} \to \boldsymbol{\lambda} + \gamma\mathbf{s}$ still satisfy the constraint (12) and the variational equations (10,11). The step size γ is determined by the same rule as before: stop when a new component enters the set $J = \arg\max_i |(\boldsymbol{R}^\top(\mathbf{y} - \boldsymbol{R}\mathbf{w}) + \boldsymbol{A}^\top\boldsymbol{\lambda})_i|$ or when a nonzero component of $\mathbf{w}$ is set to zero. Notice the differences and similarities between the linear systems (13) and (9).

At each breakpoint, this algorithm provides the penalty τ_n, the corresponding minimizer $\widetilde{\mathbf{w}}(\tau_n)$ and the Lagrange multipliers $\boldsymbol{\lambda}_n$. Unlike for the unconstrained case, it is now possible that $\widetilde{\mathbf{w}}(\tau)$ remains constant between two breakpoints (i.e. only the Lagrange multipliers $\boldsymbol{\lambda}$ change).

One simplifying assumption (not solved in the homotopy/LARS algorithm) was made in the above description of the algorithm: if the set of maximal residual and the support set differ by more than one component, one should carefully select the correct new components to enter the support. This can be done by using the variational equations, and our implementation handles this case.

One could argue that the starting point (i.e. the first breakpoint) for the constrained minimization problem is simply given by $\widetilde{\mathbf{w}}(\tau_0) = \underset{\mathbf{w} \text{ s.t. } \boldsymbol{A}\mathbf{w}=\mathbf{a}}{\arg\min} \|\mathbf{w}\|_1$, which could be calculated by letting the unconstrained solution procedure run its course: $\widetilde{\mathbf{w}}(\tau_0) = \lim_{\sigma\to 0} \arg\min_{\mathbf{w}} \left[\|\boldsymbol{A}\mathbf{w} - \mathbf{a}\|_2^2 + \sigma\|\mathbf{w}\|_1\right]$. Generically (i.e. excluding special cases), this is correct. However, the problem is that sometimes the minimizer $\text{argmin}_{\mathbf{w} \text{ s.t. } \boldsymbol{A}\mathbf{w}=\mathbf{a}} \|\mathbf{w}\|_1$ is not unique. In that case, the starting point for the constrained minimizer is not solely determined by $\boldsymbol{A}$ and $\mathbf{a}$ but also by $\boldsymbol{R}$ and $\mathbf{y}$. In this case, the ϵ-dependent algorithm still chooses the correct starting point from the set $\underset{\mathbf{w} \text{ s.t. } \boldsymbol{A}\mathbf{w}=\mathbf{a}}{\arg\min} \|\mathbf{w}\|_1$. This is important to mention because the special constraint $\sum w_i = 1$ used in this paper, gives rise to such cases.

Our algorithm is well-suited for the portfolio problems discussed in this paper. The size of the matrix, the number of constraints (just two) and, more importantly, the number of nonzero weights in the portfolios are such that a minimization run (i.e. finding the minimizer for a whole range of penalty parameters) can be done in a fraction of a second on a standard desktop.

We calculated the portfolio examples in this paper using both the formal ϵ approach (in Mathematica) and the approximate small ϵ approach (in Matlab). The outcomes were always consistent.