Simultaneous Pursuit of Out-of-Sample Performance and Sparsity in Index Tracking Portfolios

Akiko Takeda[1]*, Mahesan Niranjan[2], Jun-ya Gotoh[3], Yoshinobu Kawahara[4]

[1] Department of Administration Engineering, Keio University, Japan.
[2] School of Electronics and Computer Science, University of Southampton, United Kingdom.
[3] Department of Industrial and Systems Engineering, Chuo University, Japan.
[4] The Institute of Scientific and Industrial Research, Osaka University, Japan.

Abstract

Index tracking is a passive investment strategy in which an investor purchases a set of assets to mimic a market index. The tracking error, the difference between the performances of the index and the portfolio, may be minimized by buying all the assets contained in the index. However, this strategy results in a considerable amount of transaction cost and, accordingly, decreases the return of the constructed portfolio. On the other hand, a portfolio with small cardinality may result in a poor out-of-sample performance. Of interest is, thus, the minimization of tracking error, while keeping the number of assets invested in small (i.e., sparse).

In this paper, we develop a tracking portfolio model that addresses the conflicting requirements above by a combination of L0- and L2-norms. While L2-norm regularizes the overdetermined system to impose smoothness (and hence better out-of-sample) performance and shrinks the solution to the equally-weighted dense portfolio, L0-norm imposes a cardinality constraint that achieves sparsity (and hence lower transaction cost). We propose a heuristic method for parameter estimation in this model, which combines greedy search with an analytical formula embedded in it. We demonstrate that a sparse portfolio can achieve good tracking and generalization performance on historic data of weekly and daily returns on the Nikkei 225 index and its constituent companies.

Keywords: portfolio optimization, index tracking, norm constraint, regularization, sparse portfolio, greedy algorithm

1 Introduction

Efficiency of financial markets has been the topic of long running controversy in the literature. When efficiency holds, rapid flow of information should cause any arbitrage opportunity to be immediately neutralized. Empirically, however, inefficiencies are often observed over significant time-scales. Equilibrium models, such as the Capital Asset Pricing Model (CAPM), and its more sophisticated versions are formulated on the basis of an efficient market hypothesis, while a large volume of empirical finance literature including the use of time series analysis suggests otherwise. Strategies an investor chooses are driven by her belief in one of the above. If one were to believe in the efficient hypothesis, one would make a conservative investment strategy whereby the best return

*Corresponding author, 3-14-1 Hiyoshi, Kouhoku, Yokohama, Kanagawa 223-8522, Japan, TEL/FAX: +81 45 566 1633, takeda@ae.keio.ac.jp

on investment would simply be the average long term growth of the market. A *passive* investor who adopts this strategy would invest in the stock market index, or a portfolio approximating it and hope to benefit from the overall growth in the system. An *active* investor, on the other hand, would attempt to exploit inefficiencies by moving her portfolio dynamically, spotting opportunities for returns in excess of the overall market growth.

The obvious way of taking a passive investment strategy is to buy all the assets contained in a market index (e.g. S&P500 or TOPIX), so that the gains of the portfolio will be the same as that of the index. However, the transaction costs incurred in this strategy will make it prohibitive. To mitigate transaction costs in a passive strategy, an investor might seek to invest in a subset of stocks, in such a way that a combination of them might closely mimic the performance of the index. The smaller the number of assets needed to mimic the index, the smaller will be the incurred transaction costs. However, the tracking error, *i.e.* the difference between the performances of the index and the portfolio, is likely to be higher when a small number of assets is used. This trade-off is further compounded by non-stationarities in the markets, in the sense that a tracking portfolio constructed to mimic an index by estimating over a small window of time by regression analysis need not be the best portfolio at a different point in time, leading to what we refer to as generalization error. Balancing these conflicting demands is precisely the subject of this paper.

Several authors have considered the effects of transaction costs in setting up and tracking a market index. The simplest approach is to assume a linear cost function which can be easily embedded into any optimization scheme. A variety of more realistic nonlinear cost functions have also been considered. Konno and Wijayanayake have adopted a concave cost function in [13] and an S-shaped alternative in [14]. They have shown that the associated non-convex optimization problem can be tackled by a branch-and-bound method. Recently, Lobo, Fazel and Boyd [15] have applied a convex relaxation approach for treating a nonconvex cost function consisting of a fixed cost part and two linear floating cost parts. Specific cost functions are chosen to reflect observations of market behavior; the convex part in the S-shaped cost, for example, signifies market impact, while the concave part reflects economy of scale in commission. When formulating an optimization problem, setting of fixed costs, such as the one in [15] can also be considered to reflect economy of scale in a more specific form, and can be rewritten as a cost on the number of invested assets. Such a cost can be captured by a regularization imposed according to the L0-norm (or, equivalently, the cardinality) of a portfolio vector.

Apart from transaction costs, some investors do prefer to minimize the number of assets they wish to invest in. Thus several papers have addressed cardinality constraints on portfolios. Since the cardinality constraint adds a large combinatorial complexity to the optimization problem, most of those papers have formulated a quadratic (or linear) 0-1 mixed integer programming problem, and solved it by existing general purpose numerical solvers, or used sophisticated heuristic approaches, including genetic algorithms, tabu search and simulated annealing (e.g.,[3, 26]).

Such constraints to set up a portfolio with a subset of assets aside, the idea of tracking errors has been discussed in the literature. Here, what we are interested in is the *out-of-sample* tracking error of the discrepancy between the returns on the index and the portfolio, rather than how accurately the portfolio can mimic the return on the index in *in-sample* data. This, from a statistical modelling perspective, amounts to carrying out a linear regression analysis that does not suffer from over-fitting. As noted before, over-fitting is a potential consequence, even with a simple model such as linear regression, when we use a small temporal window of data to select an optimal tracking portfolio. The popular techniques for avoiding overfitting is *regularization* of the parameters to be estimated. Regularization is often implemented by adding a penalty term, as a function of the magnitude of the parameter vector to the original objective function. Ridge regression [12], employing the square of the L2-norm (i.e., the Euclidean norm) of the coefficient vector, is a popular regularizer of linear regression. These notions are closely related to shrinkage estimation in statistics (see, e.g., [25]).

In empirical finance work, too, the use of norm constraints regularizes models, and hence improves out-of-sample performance. For example, DeMiguel et al. [7] have reported that norm-constrained minimum variance portfolio optimizations improved the out-of-sample performance. Brodie et al. [2] have formulated a norm-constrained tracking portfolio model. Gotoh and Takeda [11] have recently studied the role of norm constraints from several viewpoints including computational learning theory. They provide a theoretical underpinning for the norm constraint in VaR- or CVaR-type loss minimization portfolio model by directly using the generalization theory (more concretely, nonparametric bounds for out-of-sample error) for support vector machines (e.g., [23]). Their numerical experiments show that the norm-constrained tracking portfolio optimization outperformed unconstrained portfolios. The generic idea in the above approaches is to use norms specified on the portfolio of weights to induce desirable properties such as good out-of-sample performance and sparsity, by formulations as convex optimization problems.

The summary above suggests that regularization (L2-norm, usually) achieves smoothness and hence good out-of-sample performance, while sparsity has been induced in portfolio models indirectly via the L1-norm, both leading to formulation as convex optimization problems. However, L1-norm is an indirect form of achieving this objective, often with no guarantee that the resulting portfolio will indeed be sparse. The constraint of interest, however, is cardinality (*i.e.* the number of non-zero components of the portfolio weight vector). Hence, in this work, we seek to develop sparse portfolios, with direct control of sparsity by cardinality (L0-norm) constraints, and additionally achieve regularization by L2-norm. In our formulation, L0- and L2-norm terms appear as regularizing constraints of a regression. We note that apart from avoiding overfitting, the L2-norm constraint also serves to mitigate against the so-called multicollinearity issue of some variables in the set being highly correlated. In a portfolio selection setting, this is very likely the case, with assets in a particular sector (e.g. technology) showing highly correlated returns over short intervals of time.

Naturally, it is computationally hard to optimize the portfolio under cardinality constraint in an exact manner. We thus propose a heuristic method on the basis of the combination of a greedy search embedding an analytic formula. We conducted computational experiments, demonstrating viability of the approach and achieving good performance with a sparse portfolio.

The remainder of this paper is organized as follows. In Section 2,we provide a brief overview of existing approaches to setting up tracking portfolio models as regularized regression problems. Then in Section 3,we introduce the new formulation of combining cardinal and quadratic regularizers, and present the search algorithm for optimization. Section 4 presents some modifications to the basic scheme allowing for rebalancing of portfolios and weighting of regression. Section 5 discusses empirical work using the *Nikkei 225* data and the numerical results obtained. We conclude in Section 6,with some pointers to our current work.

In the sequel, we will use the following notations. For any $q > 0$ and $\mathbf{x} \in \mathbb{R}^n$, let

$$\|\mathbf{x}\|_q := \left[\sum_{i=1}^{n} |x_i|^q\right]^{(1/q)}$$

and $\|\mathbf{x}\|_0 := \lim_{q \to 0+} \|\mathbf{x}\|_q$ =(the number of nonzero elements in $\mathbf{x}$). If $q \geq 1$, it is usual to refer to $\|\mathbf{x}\|_q$ as Lq-norm of $\mathbf{x} \in R^n$. By convention, even for $0 \leq q < 1$, we also call it Lq-norm. It should be noted that when $0 \leq q < 1$, $\|\mathbf{x}\|_q$ does not satisfy the axioms of norm, and consequently, it is nonconvex in $\mathbf{x}$.

2 Preliminaries

2.1 Portfolio Selection by Index Tracking

Let $N := \{1, ..., n\}$ be the index set of n investable assets. Let $\boldsymbol{R}_t := (R_{1,t}, \ldots, R_{n,t})^\top$ be the observed historical return vector of the n assets at time t and I_t be the observed market index return at time t ($t = 1, \ldots, T$). Let $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_n)^\top$ be the decision vector, each component indicating the proportion of the total amount of money devoted to asset $i \in N$.

The problem of portfolio selection by index tracking has been formulated as a linear regression problem, where a linear model $I = \boldsymbol{R}^\top \boldsymbol{\pi}$ (with parameter $\boldsymbol{\pi}$) is estimated based on a given set of observed data $(I_t, \boldsymbol{R}_t)$, $t = 1, \ldots, T$. The sum of squared deviations between portfolio returns and market index returns, is traditionally called *tracking error variance* [21] and it is defined as follows:

$$g_2(\boldsymbol{\pi}) := \frac{1}{T} \|\boldsymbol{I} - \boldsymbol{R}\boldsymbol{\pi}\|_2,$$

where $\boldsymbol{I} = (I_1, \ldots, I_T)^\top$ and $\boldsymbol{R}$ is the $(T \times n)$ matrix whose row vectors are $\boldsymbol{R}_1^\top, \ldots, \boldsymbol{R}_T^\top$. The squared error $g_2(\boldsymbol{\pi})$ is minimized with respect to $\boldsymbol{\pi}$ subject to $\boldsymbol{\pi} \in \Pi$, where $\Pi \subseteq \{\boldsymbol{\pi} : \sum_{i=1}^n \pi_i = 1\}$. Here Π denotes constraints on portfolio vector $\boldsymbol{\pi}$ and $\sum_{i=1}^n \pi_i = 1$ represents the budget constraint. If no constraints are imposed except the budget constraint, its solution can be obtained in an analytical form. Even when another constraints are imposed, the minimization of $g_2(\boldsymbol{\pi})$ results in a quadratic program as long as Π is given by a system of linear and/or convex quadratic inequalities.

Instead of the squared deviations $g_2(\boldsymbol{\pi})$, the absolute error $g_1(\boldsymbol{\pi})$ is also a popular choice in the literature (e.g., [10, 20, 22]). This is advantageous in that the optimization problem is reduced to a linear program, which can be efficiently solved. In practice, however, $g_2(\boldsymbol{\pi})$ is preferred to $g_1(\boldsymbol{\pi})$ as a well-accepted statistical criterion, especially when large errors are particularly undesirable. Therefore, we focus on the squared error $g_2(\boldsymbol{\pi})$ in this paper.

We can regard the standard Markowitz portfolio model as a special case of index tracking portfolio models, $\min_{\boldsymbol{\pi} \in \Pi} g_2(\boldsymbol{\pi})$. The Markowitz portfolio model [16] assumes that only the expectation and variability of return (i.e., mean and variance) matter to investors, with the variance acting as proxy for a measure of risk. The optimization problem here is given by

$$\min_{\boldsymbol{\pi}} \ \boldsymbol{\pi}^\top \Sigma \boldsymbol{\pi} \quad \text{s.t.} \quad \boldsymbol{r}^\top \boldsymbol{\pi} = \gamma, \quad \sum_{i=1}^n \pi_i = 1, \tag{1}$$

where $\gamma (> 0)$ is the investor's required return and $\boldsymbol{r} \in \mathbb{R}^n$ is the vector of expected returns. $\Sigma \in \mathbb{R}^{n \times n}$ is the covariance matrix of asset returns, and is estimated from historic data over a finite time window of length T. Noting that Σ can be computed by

$$\frac{1}{T} \boldsymbol{R}^\top \boldsymbol{R} - \frac{1}{T^2} \boldsymbol{R}^\top \boldsymbol{e}_T \boldsymbol{e}_T^\top \boldsymbol{R},$$

where $\boldsymbol{e}_n$ is the n-dimensional all-one vector. The objective function, $\boldsymbol{\pi}^\top \Sigma \boldsymbol{\pi}$, is transformed into the form $\frac{1}{T} \|\gamma \boldsymbol{e}_T - \boldsymbol{R}\boldsymbol{\pi}\|_2^2$, which follows from $\frac{1}{T} \boldsymbol{e}_T^\top \boldsymbol{R}\boldsymbol{\pi} = \boldsymbol{r}^\top \boldsymbol{\pi} = \gamma$. Therefore, we can regard (1) as a special case of the tracking portfolio models with $\boldsymbol{I} := \gamma \boldsymbol{e}_T$.

2.2 Existing Norm Constrained Portfolio Models

The standard approach to improve the prediction performance of the least-squares regression, formulated by $\min_{\boldsymbol{\pi}} g_2(\boldsymbol{\pi})$, is to impose a norm constraint on $\boldsymbol{\pi}$ using L_q-norm. It is known that this tends to shrink the values of a solution. These properties could give better prediction errors and increase the interpretability of the solution.

This line of approaches has been applied to portfolio problems by several researchers. Brodie et al. [2] studied L1-norm constrained portfolio optimization and furthermore introduced the L1-norm-constrained tracking portfolio model on the basis of the squared tracking errors $g_2(\boldsymbol{\pi})$:

$$\min_{\boldsymbol{\pi}} \frac{1}{T}\|\boldsymbol{I} - \boldsymbol{R}\boldsymbol{\pi}\|_2^2 \ \text{ s.t. } \ \sum_{i=1}^{n} \pi_i = 1, \ \|\boldsymbol{\pi}\|_1 \leq C_1, \tag{2}$$

where C_1 is a parameter satisfying $C_1 \geq 1$ (otherwise, (2) does not have a feasible solution). When $C_1 = 1$, the combination of the constraints $\sum_{i=1}^{n} \pi_i = 1$ and $\|\boldsymbol{\pi}\|_1 \leq C_1$ leads to $\boldsymbol{\pi} \geq \boldsymbol{0}$, which indicates that investors are not allowed to sell short certain assets, i.e., they cannot sell the assets that they do not currently own. We can regard (2) as a constrained lasso regression model. The L1-norm constraint encourages sparse portfolios, namely portfolios with only few non-zero weights. For small investors, volume independent overhead costs cannot be ignored, and controlling $\|\boldsymbol{\pi}\|_1$ and the number of assets is necessary to make transaction costs small.

In place of the L1-norm constraint $\|\boldsymbol{\pi}\|_1 \leq C_1$ in (2), we can use the L2-norm constraint $\|\boldsymbol{\pi}\|_2 \leq C_2$:

$$\min_{\boldsymbol{\pi}} \frac{1}{T}\|\boldsymbol{I} - \boldsymbol{R}\boldsymbol{\pi}\|_2^2 \ \text{ s.t. } \ \sum_{i=1}^{n} \pi_i = 1, \ \|\boldsymbol{\pi}\|_2 \leq C_2. \tag{3}$$

We can regard (3) as a constrained version of the ridge regression [12]. DeMiguel et al. [7] imposed the L2-norm constraint, $\|\boldsymbol{\pi}\|_2 \leq C_2$ on the portfolio for the variance minimization problem. They revealed that in a special case with $C_2 = 1/\sqrt{n}$, the L2-norm constraint and $\sum_{i=1}^{n} \pi_i = 1$ lead to the equally-weighted portfolio, a *naive* portfolio of the form $\pi_i = 1/n$ for all $i = 1, ..., n$. DeMiguel et al. [8] experimentally showed that among the various portfolio optimization models in the literature, there is no single model consistently better than the naive portfolio in terms of several out-of-sample performance criteria. Problem (3) shrinks the solution to the naive portfolio because (3) is equivalent to

$$\min_{\boldsymbol{\pi}} \frac{1}{T}\|\boldsymbol{I} - \boldsymbol{R}\boldsymbol{\pi}\|_2^2 \ \text{ s.t. } \ \sum_{i=1}^{n} \pi_i = 1, \ \|\boldsymbol{\pi} - \frac{1}{n}\boldsymbol{e}_n\|_2 \leq C_2. \tag{4}$$

Yen and Yen [27] have quite recently presented a portfolio optimization model with the elastic net-type regularization, i.e., a linear combination of the L1-norm and the squared L2-norm on $\boldsymbol{\pi}$ and proposed coordinate-wise descent algorithms for solving the model.

3 Sparse Tracking Portfolio Model

3.1 New Formulation

The L2-norm constraint (cf. the model (3)) could make the prediction performance significantly increase as shown in [24] and also in the numerical experiments of this paper. However, the obtained solution becomes completely dense, i.e., almost all elements of $\boldsymbol{\pi}$ are non zeros. That is, investors following this strategy have to keep almost assets while keeping small amount of each, which is not a preferable situation. As pointed out in Lobo et al. [15] and known in practice, considering the fixed cost, holding a position with a large number of assets each having small portion should be avoided.

Thus, for the portfolio design, sparsity is an important target constraint, which corresponds to the L0 regularization. As well as (2) and (3), this can be described as

$$\min_{\boldsymbol{\pi}} \frac{1}{T}\|\boldsymbol{I} - \boldsymbol{R}\boldsymbol{\pi}\|_2^2 \ \text{ s.t. } \ \sum_{i=1}^{n} \pi_i = 1, \ \|\boldsymbol{\pi}\|_0 = C_0, \tag{5}$$

where C_0 is an integer parameter satisfying $C_0 \geq 1$. The L0-constraint is also known as the cardinality constraint in the literature of financial optimization. It can be shown that the solution of the L0 formulation (5) achieves good prediction accuracy if the target function can be approximated by a sparse $\boldsymbol{\pi}$. However, a fundamental difficulty with this method is the computational cost, because the number of subsets of $\{1, \cdots, n\}$ of cardinality C_0 (corresponding to the nonzero components of $\boldsymbol{\pi}$) is exponential in C_0. There are no known efficient exact solution methods to solve the formulation (5). Due to the computational difficulty, in practice, we need to use some approximate approach to this problem. Heuristic algorithms based on generic algorithms, tabu search and simulated annealing for solving the cardinality constrained problem have been proposed in [3, 26].

One of the standard approaches for obtaining sparse portfolio is the L1-regularization (lasso), which corresponds to (2) for the tracking portfolio. It is known that L1 regularization often leads to sparse solutions. However, there are several problems with L1 regularization: first, the sparsity is not explicitly controlled; second, in order to obtain very sparse solution, one has to use a large regularization parameter that leads to (possibly) bad prediction accuracy because the L1 penalty not only shrinks irrelevant features to zero, but also shrinks relevant features to zero [28]. Fengmin et al. [9] recently proposed a sparse index tracking model based on L1/2-regularization, i.e., $\|\boldsymbol{\pi}\|_{1/2}$. The tracking portfolio problem with the L1/2-regularizer can be solved through transforming it into the solution of a series of the L1-regularizers. They proposed a hybrid half thresholding algorithm for solving the problem.

In this paper, we propose a new portfolio model aiming two objectives: i) to enhance out-of-sample performance; and ii) to achieve a sparse portfolio, by combining two different methods of regularization as

$$\begin{array}{ll} \min_{\boldsymbol{\pi}} & \|\boldsymbol{I} - \boldsymbol{R}\boldsymbol{\pi}\|_2^2 + \tau\|\boldsymbol{\pi}\|_2^2 \equiv f^{\tau}(\boldsymbol{\pi}) \\ \text{s.t.} & \displaystyle\sum_{i=1}^{n} \pi_i = 1, \ \|\boldsymbol{\pi}\|_0 \leq C_0, \end{array} \tag{6}$$

where C_0 is a positive integer parameter and τ is a positive parameter. We call the above problem as "L0+L2-norm model."

The portfolio model includes two positive parameters τ and C_0. When $\tau \to \infty$, the optimal solution becomes close to $\frac{1}{C_0}$-portfolio, i.e., each component of a portfolio is equal to $1/C_0$ (see, e.g.,[8] for the efficiency of a naive portfolio). When $\tau = 0$, this problem is equivalent to the L0 model (5) studied by [3, 26]. When $C_0 \to \infty$ it is equivalent to the L2 model, that is, a constrained ridge regression (3).

The L2-regularization term plays a very important role in this model. We can guess that multicollinearity occurs in the data matrix $\boldsymbol{R}$. Indeed, $\boldsymbol{R}$ consists of historical return vectors of n assets, and there are many pairs of assets with highly correlated returns. When multicollinearity occurs in $\boldsymbol{R}$, the least squares estimator, obtained with $\tau = 0$, is not good estimator in the sense of having a large variance. Adding the L2-regularization term is one of the ways of avoiding such problems.

It is possible to transform the L0+L2-norm model (6) into an equivalent mixed-integer optimization problem:

$$
\begin{array}{ll}
\min\limits_{\boldsymbol{\pi},\boldsymbol{y}} & \|\boldsymbol{I}-\boldsymbol{R}\boldsymbol{\pi}\|_2^2+\tau\|\boldsymbol{\pi}\|_2^2 \\
\text{s.t.} & \sum\limits_{i=1}^{n}\pi_i=1, \\
& -My_i\le\pi_i\le My_i,\quad \sum\limits_{i=1}^{n}y_i\le C_0,\quad y_i\in\{0,1\},\quad i=1,\dots,n
\end{array}
$$

using a sufficiently large positive value M and solve it with standard optimization tools such as branch-and-bound methods. It is, however, difficult to find its optimal solution within practical time limits especially when the asset size n becomes large. Therefore, in this paper, we propose applying a greedy algorithm that analytically solves subproblems induced from (6).

3.2 Greedy Algorithm

Here we interpret the formulation (6) of the L0+L2-norm model from a different viewpoint as a subset selection problem. Let S be a subset of $N:=\{1,\dots,n\}$ and $\widetilde{\boldsymbol{R}}_S$ be a $(T\times|S|)$-dimensional submatrix of $\boldsymbol{R}$ corresponding to the indices of S. Then problem (6) with $\pi_i=0$ for $i\in N\setminus S$ is reduced to

$$
g^{\tau}(S)\equiv\left|\begin{array}{ll}
\min\limits_{\widetilde{\boldsymbol{\pi}}} & \|\boldsymbol{I}-\widetilde{\boldsymbol{R}}_S\widetilde{\boldsymbol{\pi}}\|_2^2+\tau\|\widetilde{\boldsymbol{\pi}}\|_2^2 \\
\text{s.t.} & \boldsymbol{e}_{|S|}^{\top}\widetilde{\boldsymbol{\pi}}=1,\widetilde{\boldsymbol{\pi}}\in\mathbb{R}^{|S|}.
\end{array}\right. \tag{7}
$$

where $\boldsymbol{e}_k$ is the k-dimensional all-one vector. Then problem (6) is reduced to finding a subset S that minimizes $g^{\tau}(S)$ under the constraints $S\subseteq N$ and $|S|\le C_0$.

The function $g^{\tau}(S)$ is obviously a monotone set function. Indeed, let $S_{k-1}\subset S_k\subset N$. Let optimal solutions to (7) with S_{k-1} and S_k be $\widetilde{\boldsymbol{\pi}}^{\tau}_{S_{k-1}}$ and $\widetilde{\boldsymbol{\pi}}^{\tau}_{S_k}$, respectively. Note that $|S_k|$-dimensional vector $\widetilde{\boldsymbol{\pi}}^{\tau\top}=(\widetilde{\boldsymbol{\pi}}^{\tau\top}_{S_{k-1}},\boldsymbol{0}^{\top})$, where $\widetilde{\boldsymbol{\pi}}^{\tau}_{S_{k-1}}$ is in the position S_{k-1} and $\boldsymbol{0}$ is in $S_k\setminus S_{k-1}$, is a feasible solution to the problem (7) with S_k. Then, letting the objective function of (7) be $f^{\tau}_S(\widetilde{\boldsymbol{\pi}})$ and noting that $f^{\tau}_{S_k}(\widetilde{\boldsymbol{\pi}}^{\tau}_{S_k})$ is equal to $g^{\tau}(S_k)$, we get

$$
g^{\tau}(S_{k-1})=f^{\tau}_{S_{k-1}}(\widetilde{\boldsymbol{\pi}}^{\tau}_{S_{k-1}})=f^{\tau}_{S_k}(\widetilde{\boldsymbol{\pi}}^{\tau})\ge f^{\tau}_{S_k}(\widetilde{\boldsymbol{\pi}}^{\tau}_{S_k})=g^{\tau}(S_k),
$$

that shows the monotonicity of the problem (7). The monotonicity property implies that the optimal portfolio of (6) satisfies the constraint $\|\boldsymbol{\pi}\|_0\le C_0$ with equality.

If the indices corresponding to non-zero elements of $\boldsymbol{\pi}$ are determined in advance, the optimal solution to (6) is obtained explicitly with the use of the Lagrange-multiplier method. Indeed, let the set of indices of k non-zero elements be S_k. The optimal solution to (7) can be obtained as

$$
\widetilde{\boldsymbol{\pi}}^{\tau}_{S_k}=(\widetilde{\boldsymbol{R}}_{S_k}^{\top}\widetilde{\boldsymbol{R}}_{S_k}+\tau\boldsymbol{E}_k)^{-1}(\widetilde{\boldsymbol{R}}_{S_k}^{\top}\boldsymbol{I}-\lambda^*\boldsymbol{e}_k), \tag{8}
$$

where $\boldsymbol{E}_k$ is the $(k\times k)$-identity matrix and

$$
\lambda^*=\frac{\boldsymbol{e}_k^{\top}(\widetilde{\boldsymbol{R}}_{S_k}^{\top}\widetilde{\boldsymbol{R}}_{S_k}+\tau\boldsymbol{E}_k)^{-1}\widetilde{\boldsymbol{R}}_{S_k}^{\top}\boldsymbol{I}-1}{\boldsymbol{e}_k^{\top}(\widetilde{\boldsymbol{R}}_{S_k}^{\top}\widetilde{\boldsymbol{R}}_{S_k}+\tau\boldsymbol{E}_k)^{-1}\boldsymbol{e}_k}.
$$

Note that for any matrix $\widetilde{\boldsymbol{R}}_{S_k}$, $(\widetilde{\boldsymbol{R}}_{S_k}^{\top}\widetilde{\boldsymbol{R}}_{S_k}+\tau\boldsymbol{E}_k)$ is invertible as far as $\tau>0$. For notational convenience, we frequently use $f^{\tau}(\widetilde{\boldsymbol{\pi}}^{\tau}_{S_k})$ instead of $f^{\tau}_{S_k}(\widetilde{\boldsymbol{\pi}}^{\tau}_{S_k})$.

The greedy algorithm for solving (6) is described below as in Algorithm 1.

Algorithm 1 Greedy Algorithm

Initialize: $S_0 = \emptyset$ and $k = 1$. Set $f^\tau(\widetilde{\boldsymbol{\pi}}^\tau_{S_0})$ to a value large enough.
while $k \le C_0$ **do**
  For all $s \in N \setminus S_{k-1}$, compute $\widetilde{\boldsymbol{\pi}}^\tau_{S_{k-1}\cup\{s\}}$ using (8).
  Select s^* such that $\min_{s\in N\setminus S_{k-1}} f^\tau(\widetilde{\boldsymbol{\pi}}^\tau_{S_{k-1}\cup\{s\}})$ and set $\widetilde{\boldsymbol{\pi}}^\tau_{S_k} \leftarrow \widetilde{\boldsymbol{\pi}}^\tau_{S_{k-1}\cup\{s^*\}}$.
  Set $S_k \leftarrow S_{k-1} \cup \{s^*\}$ and $k \leftarrow k+1$.
end while
$\boldsymbol{\pi}^\tau_G \leftarrow \widetilde{\boldsymbol{\pi}}^\tau_{S_{C_0}}$ and $S^\tau_G \leftarrow S_{C_0}$.
return $\boldsymbol{\pi}^\tau_G$ and S^τ_G.

Note that in the above algorithm, $f(\widetilde{\boldsymbol{\pi}}^\tau_{S_{k-1}}) \ge f(\widetilde{\boldsymbol{\pi}}^\tau_{S_{k-1}\cup\{s\}})$ always holds for any $s \in N \setminus S_{k-1}$ because of the monotonicity of (7). The most time-consuming part in Algorithm 1 is the iterative inverse computations for obtaining $\widetilde{\boldsymbol{\pi}}^\tau_{S_{k-1}\cup\{s\}}$ via (8) for all $s \in N\setminus S_{k-1}$, $k = 1, \ldots, C_0$. However, we can reduce the computation of (8) by avoiding each inverse computation from scratch. Noticing that $\widetilde{\boldsymbol{R}}_{S_k}$ is constructed by appending a column vector $\boldsymbol{r}_k \in \mathbb{R}^T$ to $\widetilde{\boldsymbol{R}}_{S_{k-1}}$, the matrix $(\widetilde{\boldsymbol{R}}^\top_{S_k}\widetilde{\boldsymbol{R}}_{S_k} + \tau\boldsymbol{E}_k)$ is described as

$$\widetilde{\boldsymbol{R}}^\top_{S_k}\widetilde{\boldsymbol{R}}_{S_k} + \tau\boldsymbol{E}_k = \begin{pmatrix} \widetilde{\boldsymbol{R}}^\top_{S_{k-1}}\widetilde{\boldsymbol{R}}_{S_{k-1}} + \tau\boldsymbol{E}_{k-1} & \widetilde{\boldsymbol{R}}^\top_{S_{k-1}}\boldsymbol{r}_k \\ (\widetilde{\boldsymbol{R}}^\top_{S_{k-1}}\boldsymbol{r}_k)^\top & \boldsymbol{r}_k^\top\boldsymbol{r}_k + \tau \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{Q}_{k-1} & \boldsymbol{q}_k \\ \boldsymbol{q}_k^\top & q_{kk} \end{pmatrix}.$$

Then, by using Schur complement formula, the inverse is computed as

$$(\widetilde{\boldsymbol{R}}^\top_{S_k}\widetilde{\boldsymbol{R}}_{S_k} + \tau\boldsymbol{E}_k)^{-1} = \boldsymbol{Q}_k^{-1} = \begin{pmatrix} \boldsymbol{Q}_{k-1}^{-1} + \gamma_k\boldsymbol{s}_k\boldsymbol{s}_k^\top & -\gamma_k\boldsymbol{s}_k \\ -\gamma_k\boldsymbol{s}_k^\top & \gamma_k \end{pmatrix},$$

where $\gamma_k = \frac{1}{q_{kk} - \boldsymbol{q}_k^\top\boldsymbol{Q}_{k-1}^{-1}\boldsymbol{q}_k}$ and $\boldsymbol{s}_k = \boldsymbol{Q}_{k-1}^{-1}\boldsymbol{q}_k$. We easily get $\boldsymbol{Q}_k^{-1} = (\widetilde{\boldsymbol{R}}^\top_{S_k}\widetilde{\boldsymbol{R}}_{S_k} + \tau\boldsymbol{E}_k)^{-1}$ by using $\boldsymbol{Q}_{k-1}^{-1}$.

The greedy algorithm does not necessarily yield an optimal solution to the L0+L2-norm model (6). However we can expect that the greedy algorithm gives us a relatively good solution on the basis of the works [4, 5].

Das and Kempe [4] proved that the L0 model (5) without the equality constraint $\sum_{i=1}^n \pi_i = 1$ (that is, an L0-regularized regression problem) is equivalent to the problem of maximizing a monotone and submodular function. A real-valued function q defined on a power set of a finite set V is called *submodular* if and only if it satisfies the diminishing returns property:

$$q(S \cup \{x\}) - q(S) \ge q(T \cup \{x\}) - q(T),$$

for any $S \subseteq T (\subseteq V)$. We see from [4] that, when $\sum_{i=1}^n \pi_i = 1$ is removed and τ is set to 0 in (7), $q(S) := g^0(\emptyset) - g^0(S)$ is a normalized, monotone and submodular function (i.e., $g^0(S)$ is a *supermodular* set function)[1]. Thus, by maximizing $q(S)$, we could obtain a minimizer of $g^0(S)$. When applying the greedy algorithm to the L0-regularized regression problem, we need to compute optimal solutions to subproblems of the L0 problem by (8) with $\lambda^* = 0$ and $\tau = 0$. Minout [17] proposed an accelerated greedy algorithm for a submodular maximization problem. The algorithm reduces the computational time on average by dealing with a particular subset of $S \setminus S_{k-1}$ and, moreover, the obtained solution is the same as the greedy solution if the problem is a submodular maximization problem.

[1] *Normalized* means that f satisfies $q(\emptyset) = 0$.

Nemhauser et al. [18] showed that the greedy solution to the problem of maximizing a normalized, monotone and submodular function achieves $(1 - 1/e)$-approximation. Therefore, when $\sum_{i=1}^{n} \pi_i = 1$ is ignored (that is, set λ^* to 0), the greedy solution of the L0 model (5), $\boldsymbol{\pi}_G^0$, satisfies

$$f_{OPT}^0 \le f^0(\boldsymbol{\pi}_G^0) \le (1 - 1/e) f_{OPT}^0 + 1/e \times f^0(\mathbf{0}),$$

where f_{OPT}^0 is the optimal value of the L0-regularized regression problem. Thus, the submodularity is useful to get an approximation guarantee of the greedy algorithm to the maximization.

In cases where $\sum_{i=1}^{n} \pi_i = 1$ is not satisfied, we have no longer the above guarantee to the problem (5). However, recently Das and Kempe [5] introduced the concept of "submodularity ratio" that captures how close to a submodular function a set function is. With the analogy of this, we can define a similar criterion of g that measures the closeness of g to a supermodular function (hereafter, we call it *supermodularity ratio*). That is, the supermodularity ratio of g with respect to a set S_G^τ and a parameter C_0 can be defined as

$$\gamma_{S_G^\tau, C_0} = \min_{L \subseteq S_G^\tau, S: |S| \le C_0, S \cap L = \emptyset} \frac{\sum_{s \in S} \{g(L) - g(L \cup \{s\})\}}{g(L) - g(L \cup S)}.$$

It is shown that g is supermodular if and only if $\gamma_{S_G^\tau, C_0} \ge 1$. This concept is known to be useful to see the goodness of the greedy solution as well [5] because this gives a similar approximation guarantee of the greedy solution for an arbitrary monotone set function. That is, we can apply their theorem to the greedy solution $\boldsymbol{\pi}_G^\tau$ computed by using the monotone set function $g^\tau(S)$.

Theorem 3.1. *Let the optimal value of the L0+L2-norm model (4) be f_{OPT}^τ. Then, the following inequality holds:*

$$f_{OPT}^\tau \le f^\tau(\boldsymbol{\pi}_G^\tau) \le (1 - e^{-\gamma_{S_G^\tau, C_0}}) f_{OPT}^\tau + e^{-\gamma_{S_G^\tau, C_0}} \times f^\tau(\mathbf{0}).$$

The proof of this theorem is straightforward from [5] because we can just need to define a normalized, monotone and submodular function as the above.

4 Modifications for Sparse Tracking Portfolio

4.1 Sparse Rebalancing Portfolio

We have so far assumed that investors start with no assets. However, this is rarely the case in the real world. Investors frequently hold a large number of assets and must modify their existing portfolio to achieve a particular goal. In this context, the investor already holds a portfolio $\boldsymbol{\pi}_{t-1}^*$ and must make an adjustment $\Delta\boldsymbol{\pi}$. In that case, the portfolio of the stage will be $\boldsymbol{\pi}_{t-1}^* + \Delta\boldsymbol{\pi}$, but the transaction costs will be relevant only for the adjustment $\Delta\boldsymbol{\pi}$.

We propose a sparse rebalancing portfolio model to rebalance a portfolio from $\boldsymbol{\pi}_{t-1}^*$:

$$\begin{aligned} \min_{\Delta\boldsymbol{\pi}} \quad & \|\boldsymbol{I} - \boldsymbol{R}(\boldsymbol{\pi}_{t-1}^* + \Delta\boldsymbol{\pi})\|_2^2 + \tau\|\Delta\boldsymbol{\pi}\|_2^2 \\ \text{s.t.} \quad & \sum_{i=1}^{n} \Delta\pi_i = 0, \ \|\Delta\boldsymbol{\pi}\|_0 \le C_0. \end{aligned} \tag{9}$$

When we regard $\boldsymbol{I} - \boldsymbol{R}\boldsymbol{\pi}_{t-1}^*$ as our target to be tracked, this problem has the same structure of (6), though the variable vector is denoted by $\Delta\boldsymbol{\pi}$. Therefore, we can apply Algorithm 1 to the sparse rebalancing portfolio model. The computations of $\widetilde{\boldsymbol{\pi}}_{S_k}^\tau$ by (8) and its λ^* are modified as follows: $\boldsymbol{I}$ is replaced by $\boldsymbol{I} - \boldsymbol{R}\boldsymbol{\pi}_{t-1}^*$ and the second term in the numerator of λ^*, that is "-1", is deleted.

In summary, the proposed strategy is that if investors start with no assets so far, they use the tracking portfolio model (6) and then, use the sparse rebalancing portfolio model (9) in the following periods. To make portfolio turnover rate small, small C_0 is preferable in (9).

4.2 Weighted Regression Model

The objective function in (6), $f^{\tau}(\boldsymbol{\pi})$, can be rewritten as

$$f^{\tau}(\boldsymbol{\pi}) = \sum_{t=1}^{T}(I_t - \boldsymbol{R}_t^{\top}\boldsymbol{\pi})^2 + \tau\|\boldsymbol{\pi}\|_2^2.$$

In the objective function, the tracking error in each period is taken into account equally. In practice, however, the recent return data of each asset should be regarded important. For the purpose, we give larger weights to recent return data as

$$\sum_{t=1}^{T} w_t(I_t - \boldsymbol{R}_t^{\top}\boldsymbol{\pi})^2 + \tau\|\boldsymbol{\pi}\|_2^2,$$

where $0 \leq w_1 \leq w_2 \leq \ldots \leq w_T$. Using the diagonal matrix $\boldsymbol{W}$ with elements $(w_1, w_2, \ldots, w_T)$, we can rewrite the above equation as

$$f_W^{\tau}(\boldsymbol{\pi}) = (\boldsymbol{I} - \boldsymbol{R}\boldsymbol{\pi})^{\top}\boldsymbol{W}(\boldsymbol{I} - \boldsymbol{R}\boldsymbol{\pi}) + \tau\boldsymbol{\pi}^{\top}\boldsymbol{\pi}.$$

The L0+L2-norm model (6) whose objective function is replaced by $f_W^{\tau}(\boldsymbol{\pi})$ has an optimal solution:

$$\widetilde{\boldsymbol{\pi}} = (\boldsymbol{R}^{\top}\boldsymbol{W}\boldsymbol{R} + \tau\boldsymbol{E}_n)^{-1}(\boldsymbol{R}^{\top}\boldsymbol{W}\boldsymbol{I} - \lambda^*\boldsymbol{e}_n),$$

where

$$\lambda^* = \frac{\boldsymbol{e}_n^{\top}(\boldsymbol{R}^{\top}\boldsymbol{W}\boldsymbol{R} + \tau\boldsymbol{E}_n)^{-1}\boldsymbol{R}^{\top}\boldsymbol{W}\boldsymbol{I} - 1}{\boldsymbol{e}_n^{\top}(\boldsymbol{R}^{\top}\boldsymbol{W}\boldsymbol{R} + \tau\boldsymbol{E}_n)^{-1}\boldsymbol{e}_n}.$$

When applying the greedy algorithm to the problem, we need to solve induced subproblems. Then we can easily compute the inverse matrix $(\boldsymbol{R}^{\top}\boldsymbol{W}\boldsymbol{R} + \tau\boldsymbol{E}_n)^{-1}$ by using the inverse matrices of subproblems in a similar way to the computation of $(\boldsymbol{R}^{\top}\boldsymbol{R} + \tau\boldsymbol{E}_n)^{-1}$.

4.3 Generalized Ridge Regression Model

When adding some constraints such as lower and upper bounds to $\boldsymbol{\pi}$ in the L0+L2-norm model (6), we need to apply a quadratic optimization method to (7) in Algorithm 1 instead of computing (8). However, we may avoid adding constraints to (6) by nicely setting the parameter τ. If investors want positive $\boldsymbol{\pi}$, sufficiently large τ makes the solution $\boldsymbol{\pi}$ positive without adding $\boldsymbol{\pi} \geq \boldsymbol{0}$ constraints, since all of selected assets π_i shrink toward $1/C_0$ when $\tau \to \infty$. Suppose that there are some preferred stocks such that π_i is wanted to be almost twice as large as π_j once assets i and j are chosen to be invested, i.e., $\boldsymbol{\pi}_i, \boldsymbol{\pi}_j \neq 0$. Then we can take such a preference into consideration by changing the target to be shrinked.

Let $\boldsymbol{D_\tau}$ be a diagonal matrix with nonnegative diagonal elements $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$. We consider a more general L2-regularization term, $\boldsymbol{\pi}^{\top}\boldsymbol{D_\tau}\boldsymbol{\pi}$, instead of $\tau\boldsymbol{\pi}^{\top}\boldsymbol{\pi}$ in (6). When $\boldsymbol{\tau}$ becomes sufficiently large, all of selected assets π_i shrink toward η/τ_i, where $\eta = \sum_{k\in S}\frac{1}{\tau_k}$ for a subset S of assets.

Remark 4.1. *In the literature of statistical regression analysis, Hoerl and Kennard [12] proposed a generalized ridge regression with the general L2-regularization term of the form $\boldsymbol{\pi}^{\top}\boldsymbol{D_\tau}\boldsymbol{\pi}$ in addition to the original ridge regression. Following their paper, several papers (e.g., [1]) have discussed how to obtain the best parameter that leads to small prediction errors, even though the number of parameters to be determined in advance would increase.*

It should be noted here that in our setting we additionally impose the budget constraint, which does not usually appear in the ridge regression, and the constraint plays a role in limiting the portfolio vector to $(\eta/\tau_i)_{i\in S}$ as a shrinking point.

We can apply the greedy algorithm to the L0+L2-norm model (6) whose L2-regularization term is replaced by $\boldsymbol{\pi}^\top \boldsymbol{D}_{\boldsymbol{\tau}} \boldsymbol{\pi}$. The optimal solution (8) and λ^* can be modified by replacing $\tau \boldsymbol{E}_k$ by $\boldsymbol{D}_{\boldsymbol{\tau}}$. We can easily compute the inverse matrix $(\boldsymbol{R}^\top \boldsymbol{W} \boldsymbol{R} + \boldsymbol{D}_{\boldsymbol{\tau}})^{-1}$ by using the inverse matrices of subproblems in the similar way to the computation of $(\boldsymbol{R}^\top \boldsymbol{R} + \tau \boldsymbol{E}_n)^{-1}$.

5 Numerical Experiments

We report experimental results on real financial market data: *monthly and weekly return data of stocks listed in the Nikkei 225 index.* The monthly data set consists of returns of 182 companies during the 270 consecutive months between May 1987 and October 2009, whereas the weekly data set consists of returns of the 182 companies during the 1178 consecutive weeks from April 12, 1987 to November 1, 2009. These monthly and weekly returns of the Nikkei 225 index are our target to be tracked.

In the first subsection, we investigate properties of the L0+L2-norm model by changing two parameters, C_0 and τ, for the L0 and L2 penalties. Note that the L0+L2-norm model with $\tau = 0$ corresponds to the L0 model (5), that have been studied by [3, 26] and also approximately solved by [9]. In the second subsection, we compare the L0+L2-norm model with existing norm-constraints models with L0-, L1- or L2-norm. We exclude the portfolio optimization model with the elastic net-type regularization [27] here because it would achieve intermediate performance between the L1-norm model and the L2-norm model.

5.1 Properties of L0+L2-norm model

We designed a portfolio as a greedy solution to (6) using historical data $\boldsymbol{R}_t, \ldots, \boldsymbol{R}_{t+T-1}$ and the Nikkei 225 indices $I_t, \ldots, I_{t+T-1}$ for $T = 120$ (10 years) consecutive periods from the monthly data set or for $T = 150$ (almost 3 years) consecutive periods from the weekly data set. Let $\widehat{\boldsymbol{\pi}}_t$ be a decision vector learned from the T training samples. We evaluated the test error $|I_{t+T} - \boldsymbol{R}_{t+T}^\top \widehat{\boldsymbol{\pi}}_t|$ for the next-step sample $(I_{t+T}, \boldsymbol{R}_{t+T})$. This procedure was repeated for $t = 1, \ldots, \overline{T}$ ($\overline{T} = 150$ for the monthly data set and $\overline{T} = 1028$ for the weekly data set). The mean of the training mean squared error (training MSE) over $\overline{T}$ of time was computed as

$$\frac{1}{\overline{T}} \sum_{t=1}^{\overline{T}} \left\{ \frac{1}{T} \sum_{s=0}^{T-1} (I_{t+s} - \boldsymbol{R}_{t+s}^\top \widehat{\boldsymbol{\pi}}_t)^2 \right\}.$$

The mean of test squared errors (test SEs):

$$\frac{1}{\overline{T}} \sum_{t=1}^{\overline{T}} (I_{t+T} - \boldsymbol{R}_{t+T}^\top \widehat{\boldsymbol{\pi}}_t)^2$$

and the variance of test SEs were employed as performance measures. The turnover of the portfolio was also computed by

$$\text{Turnover} = \frac{1}{\overline{T} - 1} \sum_{t=2}^{\overline{T}} \sum_{j=1}^{n} \left| \widehat{\pi}_{t,j} - \frac{1 + R_{t+T,j}}{1 + \boldsymbol{R}_{t+T}^\top \widehat{\boldsymbol{\pi}}_{t-1}} \widehat{\pi}_{t-1,j} \right|,$$

as another performance measure. The portfolio turnover indicates how frequently assets in a portfolio are bought and sold. All of those measures (i.e., the mean of test SEs, its variance and turnover) are preferred to be small.

The L0+L2-norm model has two parameters: C_0 of the L0-norm and τ of the L2-norm. Figures 1 to 6 show numerical results for the L0+L2-norm model with several C_0 and $\tilde{\tau} = \tau/T = \tau/120$ on monthly data set, while Figures 7 to 12 show results for the L0+L2-norm model with several C_0 and $\tilde{\tau} = \tau/T = \tau/150$ on weekly data set.

Figures 1 and 7 show the mean of test SEs, that was computed with the greedy solution $\boldsymbol{\pi}_G^\tau$ of the L0+L2-norm model. Figures 2 and 8 show the variance of test SEs. Note that the L0+L2-norm model with $\tilde{\tau} = 0$ corresponds to the L0-portfolio model with parameter C_0, and the model with $C_0 = \infty$ corresponds to the L2-portfolio model with parameter $\tilde{\tau}$. From these four figures, we see that positive parameter $\tilde{\tau}$ considerably improves test SEs compared to $\tilde{\tau} = 0$, though the L0-portfolio model achieves smaller training MSE than the L0+L2-norm model. These figures also imply that the L0+L2-norm model achieved good performance with small asset size (for example, $C_0 = 20$ in Figures 1 and 2) if we set a small value to $\tilde{\tau}$ (e.g., $\tilde{\tau} = 20$). It is, however, possible to achieve better performance with larger asset size and larger $\tilde{\tau}$ ($C_0 = 50$ and $\tilde{\tau} = 80$ in Figures 1 and 2), while much computation time is needed because of the larger combinatorial number $\binom{182}{C_0}$.

Figures 3 and 9 show the turnover of a portfolio $\boldsymbol{\pi}_G^\tau$ obtained via Algorithm 1. Positive τ has a role to depress the turnover of a portfolio in the L0+L2-norm model. A large enough parameter C_0 also tries to make the turnover small.

Let N_+ be the set of indices of positive elements in $\boldsymbol{\pi}_G^\tau \in \mathbb{R}^{C_0}$ and N_- be the set of indices of negative elements in $\boldsymbol{\pi}_G^\tau$. Since there were no zero elements in $\boldsymbol{\pi}_G^\tau$, one has $|N_+ \cup N_-| = C_0$ and $N_+ \cap N_- = \emptyset$. Figures 4 and 10 show the ratio of the number of positive elements in $\boldsymbol{\pi}_G^\tau$ over C_0, that is, $|N_+|/C_0$. On the other hand, $1 - |N_+|/C_0$ shows the ratio of negative elements in $\boldsymbol{\pi}_G^\tau$. These figures show that the ratio of negative elements becomes large when C_0 becomes large. However, we can make the ratio of negative elements small when we set τ large. This indicates that nicely setting the parameters τ and C_0 can avoid short sale.

Figures 5 and 11 depict the L1-norm of the obtained portfolios, $\|\boldsymbol{\pi}_G^\tau\|_1$. Note that transaction cost for a portfolio $\boldsymbol{\pi}$ consists of the fixed cost related to $\|\boldsymbol{\pi}\|_0$ and variable costs depending on the amount of trading assets, that is, $\|\boldsymbol{\pi}\|_1$. Therefore, small $\|\boldsymbol{\pi}\|_1$ is preferable in order to make the transaction cost small. $\tilde{\tau}$ is a parameter for making $\|\boldsymbol{\pi}\|_2$ small, but these figures imply that large $\tilde{\tau}$ make $\|\boldsymbol{\pi}\|_1$ small as well as $\|\boldsymbol{\pi}\|_2$. Noticing that

$$\|\boldsymbol{\pi}_G^\tau\|_1 = \sum_{i \in N_+} \pi_{Gi}^\tau - \sum_{i \in N_-} \pi_{Gi}^\tau = \sum_{i \in N_+ \cup N_-} \pi_{Gi}^\tau - 2 \sum_{i \in N_-} \pi_{Gi}^\tau = 1 - 2 \sum_{i \in N_-} \pi_{Gi}^\tau,$$

we expect that $\|\boldsymbol{\pi}_G^\tau\|_1$ is close to 1 if the number of negative elements, that is $|N_-|$, is small in $\boldsymbol{\pi}_G^\tau$. Figures 4 and 5 as well as Figures 10 and 11 support the expectation.

L2-regularization in the ridge regression helps to avoid the so-called multicollinearity. Too small or large eigenvalues of $(\widetilde{\boldsymbol{R}}_{S_{C_0}}^\top \widetilde{\boldsymbol{R}}_{S_{C_0}})$ would indicate multicollinearity problems. Therefore, The condition number, defined by the square root of the largest eigenvalue divided by the smallest eigenvalue, i.e, $\kappa = \sqrt{\lambda_{max}/\lambda_{min}}$, is sometimes used to detect whether multicollinearity occurs in the data matrix. Figures 6 and 12 depict the condition number, scaled by C_0. We can make sure that the assets were selected to avoid the multicollinearity when τ becomes positive, though the effect is not so obvious for weekly dataset.

5.2 Comparison to existing norm-constrained portfolio models

We compared performance of the L0+L2-norm model with existing portfolio models. In the L1- and L2-norm models, the parameters (C_1 and C_2) were systematically tuned as follows. Using the first $\frac{5}{6}T$-period of the training samples, we obtained the pair (β, C) that gave the minimum

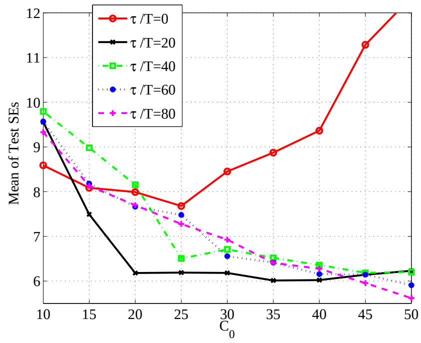

Figure 1: Mean of test SEs (monthly).

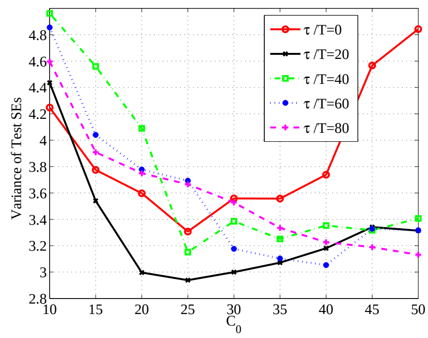

Figure 2: Variance of test SEs (monthly).

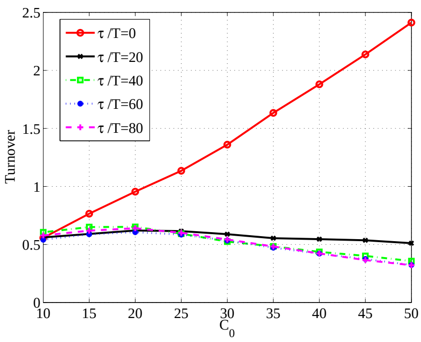

Figure 3: Turnover (monthly).

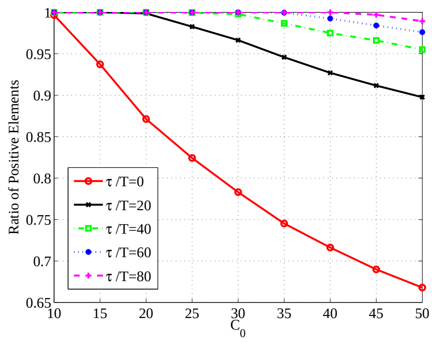

Figure 4: The ratio of the number of positive elements in $\boldsymbol{\pi}_G^\tau$ over C_0 (monthly).

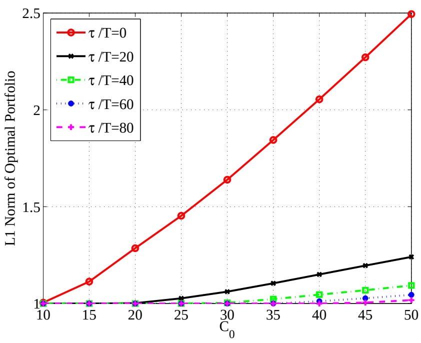

Figure 5: L1-norm of $\boldsymbol{\pi}_G^\tau$ (monthly).

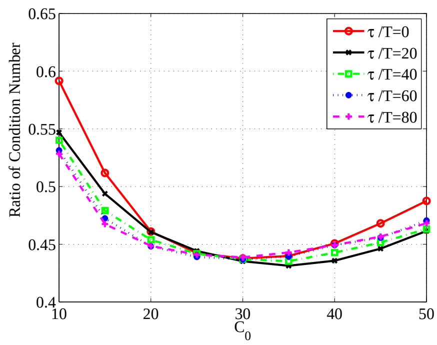

Figure 6: The ratio of the condition number of $(\widetilde{\boldsymbol{R}}_{S_{C_0}}^\top \widetilde{\boldsymbol{R}}_{S_{C_0}})$ over C_0 (monthly).

MSE for the remaining $\frac{1}{6}T$-period. C_1 was chosen from $1.00 + 0.05(k-1)$, $k = 1, \ldots, 5$, and C_2 was chosen from $1/\sqrt{n} + k(1 - 1/\sqrt{n})/(10\sqrt{n})$, $k = 1, \ldots, 5$. The way of parameter tuning for the L2-norm model was proposed in [24].

Table 1 shows the mean of test squared errors (SEs): $(I_{t+T} - \boldsymbol{R}_{t+T}^\top \widehat{\boldsymbol{\pi}}_t)^2$ for $t = 1, \ldots, \overline{T}$, variance of test SEs and average nonzero ratio of optimal solutions, that was computed as the average number of nonzero elements in solutions $\boldsymbol{\pi}_t$, $t = 1, \ldots, \overline{T}$, divided by the number of assets $n = 182$. The L1-norm of each portfolio is also shown. The results of the L0- model and the L0+L2-norm model are taken from numerical experiments with good parameter choices for C_0 and τ, that are shown

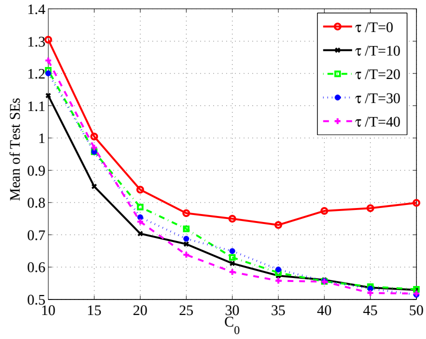

Figure 7: Mean of test SEs (weekly).

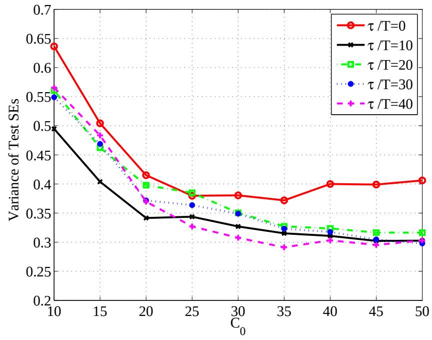

Figure 8: Variance of test SEs (weekly).

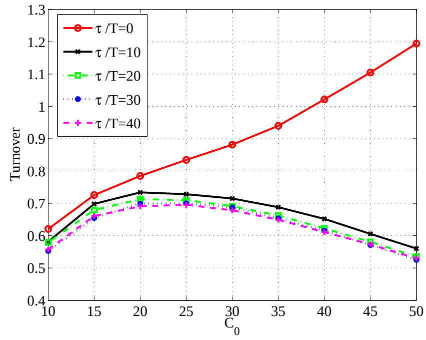

Figure 9: Turnover (weekly).

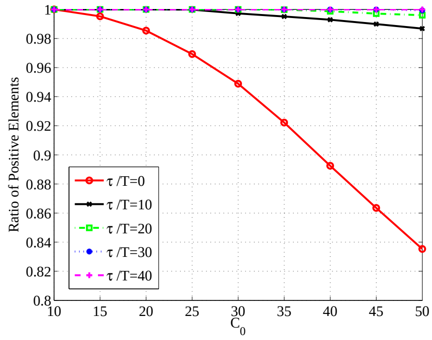

Figure 10: The ratio of the number of positive elements in $\boldsymbol{\pi}_G^\tau$ over C_0 (weekly).

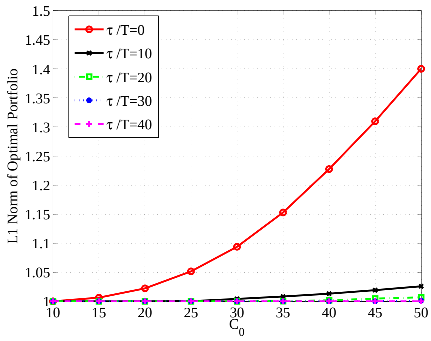

Figure 11: L1-norm of $\boldsymbol{\pi}_G^\tau$ (weekly).

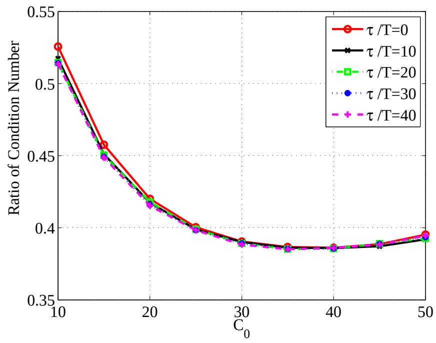

Figure 12: The ratio of the condition number of $(\widetilde{\boldsymbol{R}}_{S_{C_0}}^\top \widetilde{\boldsymbol{R}}_{S_{C_0}})$ over C_0 (weekly).

in the previous section in Figures 1 to 12. The table shows that the L0+L2-norm model competes with the other models in terms of test SEs and their variances, though the resulting portfolios are sparse compared to the L1- and L2-norm models. Especially, even if we restrict the asset number in portfolio to $C_0 = 20$ for monthly data set, the L0+L2-norm model could achieve competitive results to other models.

Finally we show that the L0+L2-norm model still competes with the other models even if we decrease the number of assets in the asset universe. We randomly chose n $(= 20, 40, \dots, 180)$ assets from the 182 assets and investigated the sparsity and prediction performance of optimal portfolios

Table 1: Comparison among L0-norm model (5), L1-norm model (2), L2-norm model (3) and L0+L2-norm model (6) for Nikkei dataset ($n = 182$)

	monthly dataset			
	mean test SEs	variance	nonzero [%]	L1norm
L0+L2-norm model ($C_0 = 50, \tilde{\tau} = 80$)	5.62	3.13	27.5	1.02
L0+L2-norm model ($C_0 = 20, \tilde{\tau} = 20$)	6.18	3.00	11.0	1.00
L0-norm model ($C_0 = 25$)	7.68	3.31	13.7	1.45
L1-norm model	6.28	3.13	30.6	1.04
L2-norm model	5.57	3.24	100	1.05
	weekly dataset			
	mean test SEs	variance	nonzero [%]	L1norm
L0+L2-norm model ($C_0 = 50, \tilde{\tau} = 30$)	0.514	0.298	27.5	1.00
L0-norm model ($C_0 = 35$)	0.731	0.372	19.2	1.15
L1-norm model	0.511	0.308	47.9	1.03
L2-norm model	0.474	0.296	100	1.09

in each model. The parameters C_1 of the L1-norm model and C_2 of the L2-norm model were systematically tuned as above. The L0-norm model was omitted since it performed very poorly.

Figures 13 to 16 show the results for monthly data set and Figures 17 to 20 show the results for weekly data set. The L2-norm model achieved the best performance in terms of the mean of test SEs. However, the optimal portfolio of the L2-norm model is completely dense, that is, there are no zero in the optimal portfolio. On the other hand, we can make sure that the prediction performance of the L0+L2-norm model is not deteriorated so much even if we limit the number of assets in portfolio to $C_0 = 30$ or 40 for monthly data set and $C_0 = 40$ or 50 for weekly data set. In the L0+L2-norm model, we can choose the number of non-zero elements in the portfolio as we like and the resulting portfolio (even if it is a greedy solution) achieves a good prediction performance. In that sense, the L0+L2-norm model is one of realistic tracking portfolio models.

6 Conclusions

We have shown it is possible to effectively combine two different methods of regularization in optimization of portfolios that track an index, with an L0-norm selecting a small subset and the L2-norm offering smoothness, thereby achieving enhanced out-of-sample performance. Our model incorporates these two regularizers simultaneously, thereby drawing on their respective strengths. To mitigate against non-convexity associated with L0-norm, we have adopted a greedy search strategy with analytic treatment of the remainder embedded within it. Empirical work on datasets over two different timescales confirm statistically significant improvements by this combination of regularizers.

We have adopted the combination of L0- and L2-norms in the tracking portfolio model (and also in the standard Markowitz portfolio model (1)) as regularization terms. We are interested in using the combined L0- and L2-norm regularization to the minimum variance portfolio optimizations, to which DeMiguel et al. [7] have added a norm-constraint. We expect that the combined L0- and L2-norm regularization leads to a sparse portfolio achieving good out-of-sample performance.

In this paper we proposed a new tracking portfolio model (6) and made sure that we could get good numerical results from the model. In current work, we are interested in improving the simple greedy algorithm, i.e., Algorithm 1. We could use the idea of orthogonal matching pursuit [6, 19],

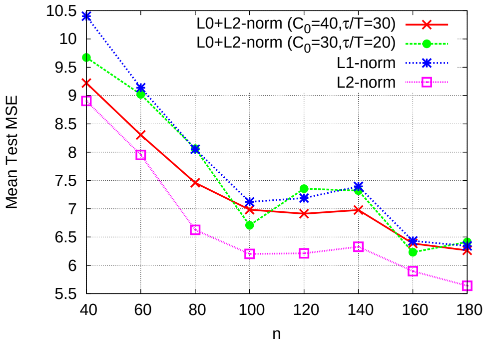

Figure 13: Mean of test SEs (monthly).

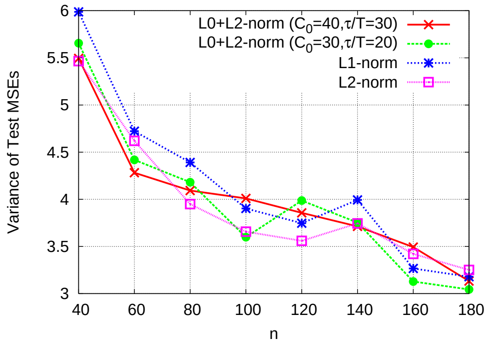

Figure 14: Variance of test SEs (monthly).

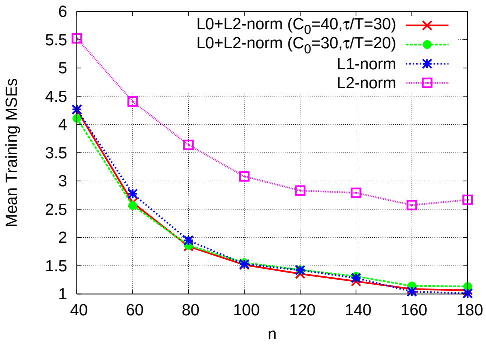

Figure 15: Mean of training MSE (monthly).

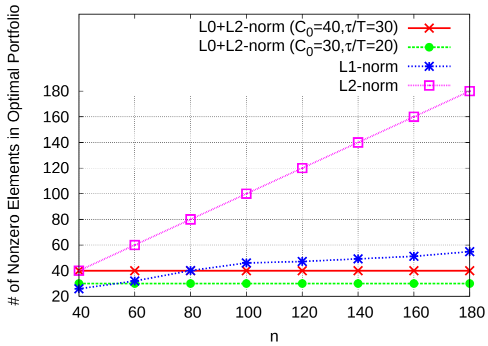

Figure 16: The number of nonzero elements (monthly).

that is used for finding sparse representations for signals in compressive sensing, in our problem setting.

References

[1] P. M. Boer and C. M. Hafner. Ridge regression revisited. *Statistica Neerlandica*, 59(4):498–505, 2005.

[2] J. Brodie, I. Daubechies, C. De Molc, D. Giannone, and I. Loris. Sparse and stable markowitz portfolios. *PNAS*, 106:12267–12272, 2009.

[3] T.J. Chang, N. Meade, J.E. Beasley, and Y.M. Sharaiha. Heuristics for cardinality constrained portfolio optimisation. *Computers & Operations Research*, 27:1271–1302, 2000.

[4] A. Das and D. Kempe. Algorithms for subset selection in linear regression. In *Proceedings of STOC 2008*, pages 45–54, 2008.

[5] A. Das and D. Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of International Conference on Machine Learning*, 2011.

[6] G. Davis, S. Mallat, and Z. Zhang. Adaptive time-frequency decompositions with matching pursuits. *Optical Engineering*, 33(7):2183–2191, 1994.

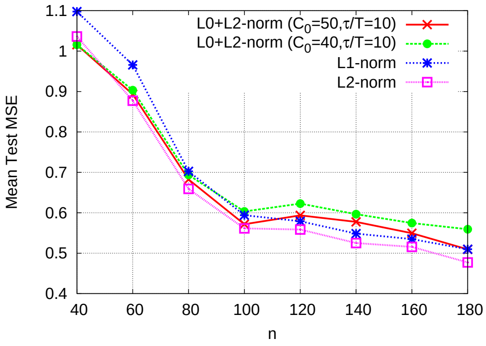

Figure 17: Mean of test SEs (weekly)

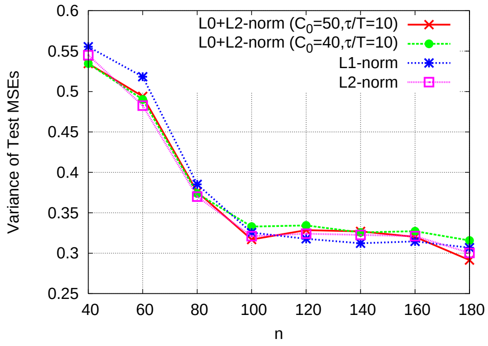

Figure 18: Variance of test SEs (weekly)

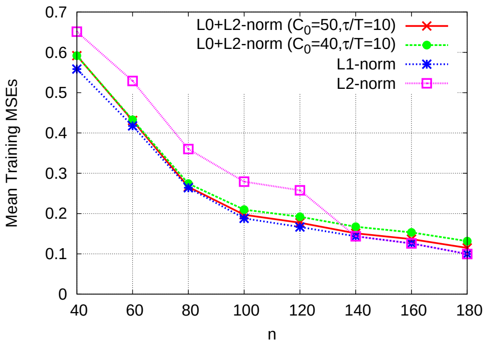

Figure 19: Mean of training MSE (weekly).

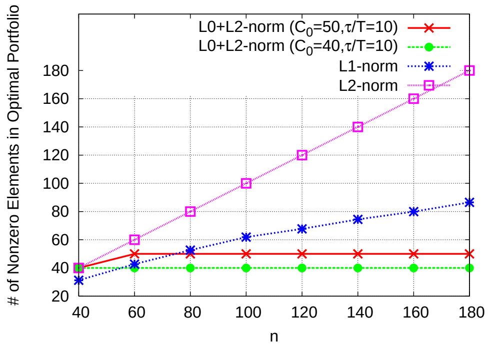

Figure 20: The number of nonzero elements (weekly).

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

[8] V. DeMiguel, L. Garlappi, and R. Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies*, 22(5):1915–1953, 2009.

[9] X. Fengmin, Z. Xu, and H. Xue. Sparse index tracking: An l1/2 regularization based model and solution. *Available at http://gr.xjtu.edu.cn:8080/LiferayFCKeditor/UserFiles/File/NewAOR.pdf*, 2011.

[10] M. Gilli and E. Këllezi. The threshold accepting heuristic for index tracking. In *Financial Engineering, E-Commerce, and Supply Chain*, pages 1–18. Kluwer Academic, Dordrecht, 2002.

[11] J. Gotoh and A. Takeda. On the role of norm constraints in portfolio selection. *Computational Management Science*, 8(4):323–353, 2011.

[12] E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[13] H. Konno and A. Wijayanayake. Portfolio optimization problem under concave transaction costs and minimal transaction unit constraints. *Mathematical Programming*, 89(2):233–250, 2001.

[14] H. Konno and A. Wijayanayake. Portfolio optimization under d.c. transaction costs and minimal transaction unit constraints. *Journal of Global Optimization*, 22(1–4):137–154, 2002.

[15] M. Lobo, M. Fazel, and S. Boyd. Portfolio optimization with linear and fixed transaction costs. *Annals of Operations Research*, 152(1):341–365, 2007.

[16] H. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[17] M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*, pages 234–243, 1978.

[18] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

[19] Y. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit : recursive function approximation with application to wavelet decomposition. In *Asilomar Conf. on Signals, Systems and Comput.*, 1993.

[20] J.-L. Prigent. *Portfolio Optimization and Performance Analysis.* Chapman & Hall/CRC, Boca Raton, 2007.

[21] R. Roll. A mean/variance analysis of tracking error. *Journal of Portfolio Management*, 18(4):13–22, 1992.

[22] M. Rudolf, H. J. Wolter, and H. Zimmermann. A linear model for tracking error minimization. *Journal of Banking & Finance*, 23(1):85–103, 1999.

[23] B. Schölkopf, A. Smola, R. Williamson, and P. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

[24] A. Takeda, J. Gotoh, and M. Sugiyama. Support vector regression as conditional value-at-risk minimization with application to financial time-series analysis. In *Proceedings of 2010 IEEE International Workshop on Machine Learning for Signal Processing*, 2010.

[25] T.Hastie, R.Tibshirani, and J.Friedman. *The Elements of Statistical Learning –Data Mining, Inference, and Prediction.* Springer-Verlag, Ney York, second edition, 2008.

[26] M. Woodside-Oriakhi, C. Lucas, and J.E. Beasley. Heuristic algorithms for the cardinality constrained efficient frontier. *European Journal of Operational Research*, 213:538–550, 2011.

[27] Y. Yen and T. Yen. Solving norm constrained portfolio optimizations via coordinate-wise descent algorithms. Technical report, London School of Economics and Political Science, UK., 2011.

[28] T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. *Advances in Neural Information Processing Systems*, 22, 2008.