# Sparse and robust normal and t- portfolios by penalized Lq-likelihood minimization

Ferrari, Davide

University of Melbourne

davide.ferrari@unimelb.edu.au

Giuzio, Margherita

EBS Universität für Wirtschaft und Recht

margherita.giuzio@ebs.edu

Paterlini, Sandra

EBS Universität für Wirtschaft und Recht

sandra.paterlini@ebs.edu

May 22, 2014

**Abstract**

Two important problems arising in traditional asset allocation methods are the sensitivity to estimation error of portfolio weights and the high dimensionality of the set of candidate assets. In this paper, we address both issues by proposing a new minimum description length criterion for portfolio selection. The new criterion is a two-stage description of the available information, where the $q$-entropy, a generalized measure of information, is used to code the uncertainty of the data given the parametric model and the uncertainty related to the model choice. The information about the model is coded in terms of a prior distribution that promotes asset weights sparsity. Our approach carries out model selection and estimation in a single step, by selecting few assets and estimating their portfolio weights simultaneously. The resulting portfolios are doubly robust, in the sense that they can tolerate deviations from both, assumed data model and prior distribution for model parameters. Empirical results on simulated and real-world data support the validity of our approach in comparison to state-of-art benchmarks.

*Keywords*: $q$-entropy, penalized least squares, sparsity, index tracking

JEL: C15, C61, G11

# 1 Introduction

Asset allocation aims to determine an optimal portfolio from a large set of assets by explicitly considering their contribution in terms of performance and diversification. Markowitz (1952) pioneered modern portfolio theory by introducing the mean-variance optimization framework, which explicitly deals with the trade-off between portfolio risk and return, quantified respectively by the expected covariance matrix and the expected asset returns. It is known, however, that the performance of the mean-variance portfolio is largely affected by the uncertainty about the model parameters and the data distribution, which is traditionally assumed to be Gaussian (Markowitz, 1952; Michaud, 1989; Merton, 1980; Frost and Savarino, 1988). Portfolio weights are very sensitive to errors for the model parameter estimates induced by model misspecification (Best and Grauer, 1991; Jagannathan and Ma, 2003); for example, even a small change in the covariance and mean estimates of the multivariate Gaussian model might cause substantial changes in the optimal portfolio asset allocations (De Miguel and Nogales, 2009). As a result, the out-of-sample performance of the portfolio is often unsatisfactory. Moreover, the effect of estimation bias is frequently enhanced by the large pool of candidate assets and by the strong correlation between asset returns series. In addition, financial data are leptokurtic and contaminated by outliers (Cont, 2001). The misspecification of the underlying Gaussian distribution causes then imprecise estimates, if deviations from the assumed model are not appropriately addressed.

To deal with these issues, two distinct classes of statistical methods have become increasingly popular in the financial literature. To achieve robust estimation of model parameters, various robust procedures to estimate the covariance matrix have been proposed (Ledoit and Wolf, 2004b; Welsch and Zhou, 2007; Vaz de Melo and Camara, 2005; De Miguel and Nogales, 2009). To achieve sparse portfolios with a relatively small number of assets corresponding to active (i.e. non-zero) weights from a large pool of assets, several authors have advocated the use of penalized least squares methods. For instance, the least absolute shrinkage and selector operator (Lasso) (Tibshirani, 1996) has proved to be very useful in asset allocation since it not only increases the stability of portfolio weights, but also enforces sparse solutions by imposing a penalty on the asset weights (De Mol et al., 2008; De Miguel and Nogales, 2009; Fan et al., 2012; Carrasco and Noumon,

2012).

Considering the high-dimensionality of the asset allocation problem and the stylized facts of financial data (Cont, 2001), we address the important question of how to build well-performing and sparse portfolios. Then, we introduce a new class of algorithms for portfolio selection, which merges the strengths of both, robust and penalized estimation methods. Our approach involves the minimization of a description length criterion that accounts for the uncertainty about the data as well as that related to the parametric model structure. Both sources of uncertainty are coded in terms of the $q$-entropy, a generalized information measure introduced by Havrda and Charvát (1967) and studied by Tsallis (1988) in statistical mechanics. The $q$-entropy coding ensures double robustness by protecting against two sources of model misspecification: i) it down-weights observations that diverge from the data model assumed for the assets; ii) it mitigates the effect of parameter estimates that are far from the assumed prior structure for the parameters.

The overall behavior of our criterion-function depends on a tuning parameter $q$, which controls the trade-off between the statistical accuracy and the stability of estimates (Ferrari and La Vecchia, 2012). When $q \to 1$ our procedure is equivalent to maximum a posteriori estimation of the parameters, yielding optimal statistical efficiency but scarce robustness; values of $q < 1$ yield instead robust estimates with negligible loss of efficiency. Importantly, our approach tackles the model selection and estimation in a single step by identifying optimal portfolios with few active positions by means of a penalty function on the asset weights, with size depending on a regularization parameter $\lambda$. Larger penalty values imply sparse portfolios with a relatively small number of assets. The solution to the new criterion can be efficiently solved by iterative algorithms that we develop in the paper.

The remainder of the paper is structured as follows: in Section 2, we describe the general methodology based on the two-stage description length minimization framework. In Section 3, we propose an efficient re-weighting algorithm to compute optimal portfolios. In the same section, we focus on two important cases, when the portfolio returns are assumed to follow a normal or a t-Student distribution, while the prior structure on the parameters is coded by a Laplace distribution. In Section 4, we compare the performance of our method by Monte Carlo simulation

with other benchmark approaches. In Section 5, we illustrate our portfolio selection method by considering real-world data. Section 6 concludes the paper.

# 2 The Methodology

## 2.1 A two-stage description length criterion

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be a $p$-dimensional real-valued random vector of asset returns generated by some unknown multivariate distribution. A financial portfolio return, $Y$, can then be defined by the linear combination $Y = \boldsymbol{\beta}^T \boldsymbol{X}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is a vector of constants playing the role of asset allocation weights. The portfolio expected return and variance are $E(Y) = \mu$ and $Var(Y) = \sigma^2$, respectively. We assume that the standardized portfolio return $Z = (Y - \mu)/\sigma$ has probability distribution $G(z)$ and probability density function $g(z)$. To emphasize possible model misspecifications we distinguish between the true density $g$, and the model density $f$. The latter is a user-specified model chosen to represent the data which might or might not coincide with the true density $g$. For example, $f$ can be the standard normal distribution or the t-Student distribution with $\nu$ degrees of freedom. In our framework, the portfolio expected return is viewed as a fixed target, and the main interest is to estimate the asset weights $\boldsymbol{\beta}$ for $\mu$ set equal to some desirable level $\mu^*$. Although $\sigma$ can be viewed as a nuisance parameter and fixed at some target value, we typically estimate both $\sigma$ and $\boldsymbol{\beta}$.

Given a mean target value $\mu = \mu^*$ and observations $\mathbf{x}_i$, $i = 1, \ldots, n$, we compute portfolio coefficients, $\widehat{\boldsymbol{\beta}}_{q,\lambda}$, by minimizing the following two-stage description length, or generalized description length (GDL) criterion:

$$\widehat{D}_{q,\lambda}(\boldsymbol{\beta}, \sigma) = -\sum_{i=1}^{n} L_q \left\{ f\left( \frac{\mathbf{x}_i^T \boldsymbol{\beta} - \mu^*}{\sigma} \right) \right\} - \sum_{j=1}^{p} L_q \left\{ \pi(\beta_j; \lambda) \right\}, \tag{1}$$

for fixed tuning constants $\lambda \geq 0$ and $q \leq 1$. In (1), $L_q(\cdot)$ is the generalized $q$-logarithm

$$L_q(u) = \begin{cases} (u^{1-q} - 1)/(1-q), & q \neq 1, \\ \log(u), & q = 1, \end{cases} \tag{2}$$

and $\pi(\beta_j; \lambda)$ is a symmetric density function for $\beta_j$ with zero mean and variance depending on $\lambda$. To gain more insight on the proposed optimization task, it is helpful to consider the special case where $q \to 1$, which implies $L_q(\cdot) \to \log(\cdot)$ and

$$-\widehat{D}_{q,\lambda}(\boldsymbol{\beta}, \sigma) \to \log \left\{ \prod_{i=1}^{n} f(\sigma^{-1}(\mathbf{x}_i^T \boldsymbol{\beta} - \mu^*)) \prod_{j=1}^{p} \pi(\beta_j; \lambda) \right\}. \tag{3}$$

This shows that when $q$ is near 1, minimization of (1) is equivalent to maximum a posteriori (MAP) estimation of $\boldsymbol{\beta}$, where $\pi(\beta_j; \lambda)$, $j = 1, \ldots, p$, play the role of prior pdfs for independent coefficients $\beta_j$, $j = 1, \ldots, p$.

Criterion (1) is regarded as a two-stage description of the total information. The first sum in (1) is interpreted as the information provided by the data ($\mathbf{x}_i$, $i = 1, \ldots, n$) given a model indexed by $\boldsymbol{\beta}$, $\sigma$ and $\mu^*$. The second sum encodes the information about the model structure itself through the prior distributions $\pi(\beta_j; \lambda)$, $j = 1, \ldots, p$. The penalty function $\pi(\beta_j; \lambda)$ drives the model selection step and leads to sparse optimal solutions by forcing to zero the components of the $\boldsymbol{\beta}$ vector that are contributing in reaching the mean target value $\mu = \mu^*$. Different pdfs could be chosen as penalty functions $\pi(\beta_j; \lambda)$. For example, Figure 1 (left) shows $L_q(\pi(\beta; \lambda))$ for a single coefficient $\beta_j$ when $\pi(\cdot; \cdot)$ is a Normal, a Laplace or a Double Pareto pdf with $q = 1/2$ and $\lambda = 1$. When comparing the Laplace with the Double Pareto, we notice that the Laplace penalty is weaker when $\beta_j$ is close to zero and much stronger when $\beta_j$ moves further away from zero. In this paper, we focus on considering the Laplace as penalty function $\pi$, but the model can be easily extended to include other penalty without requiring any algorithmic modification. Figure 1 (right) shows the effect of increasing the value of $\lambda$ in the Laplace penalty of the $\beta$ coefficients. The parameter $\lambda$ controls for the size of $\pi(\beta_j; \lambda)$: increasing the $\lambda$ value leads then to identify optimal solutions with a smaller number of active $\beta$ coefficients.

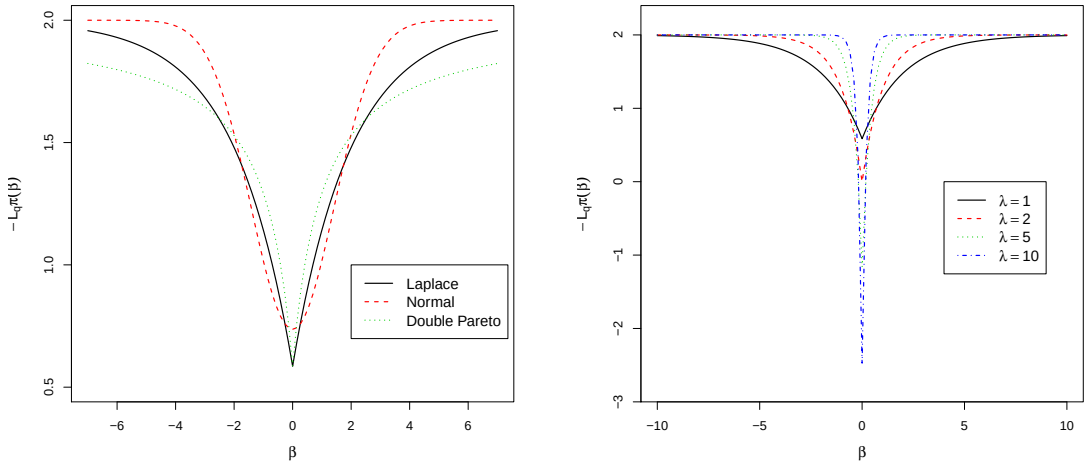



Figure 1: Penalty functions. Left panel: Penalty functions $\pi()$ for a single component $\beta$ corresponding to (standard) Laplace, normal and double Pareto penalties with $q = 1/2$ and $\lambda = 1$. Right panel: Laplace penalties for different values of $\lambda$ ranging from 1 to 10 and $q = 1/2$. Note that $-L_q(\pi(\beta|\lambda)) \to (1-q)^{-1}$ as $|\beta| \to \infty$.

Note that for many choices of $\pi$, our penalty function has the property of being non-convex. This aspect is important to deal with the shortcoming of the convex $\ell_1$-penalty, which is known to produce biased estimates for large (absolute) coefficients (Fan and Li, 2001; Zou, 2006). As a solution, various authors proposed to use penalties that are singular at the origin (just like the $\ell_1$-penalty) in order to promote sparsity, but non-convex, in order to countervail bias (see Gasso et al. (2009) for a discussion of benefits of using non-convex penalties).

## 2.2 Robust Estimating equations

Differentiating the criterion function (1) with respect to parameters $(\boldsymbol{\beta}, \sigma)^T$ results in the following estimating equations:

$$\mathbf{0} = \widehat{\boldsymbol{\Psi}}(\boldsymbol{\beta}, \sigma) := \nabla \hat{D}_{q,\lambda}(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^{n} w_q(\mathbf{x}_i, \boldsymbol{\beta}, \sigma)\boldsymbol{u}(\mathbf{x}_i, \boldsymbol{\beta}, \sigma) + \mathbf{p}'_\lambda(\boldsymbol{\beta}), \tag{4}$$

where "$\nabla$" denotes the gradient operator, so that $\widehat{\boldsymbol{\Psi}}(\boldsymbol{\beta}, \sigma)$ is the $(p+1)$-vector of derivatives with respect to the elements of the parameter vector $(\boldsymbol{\beta}, \sigma)^T$. In the above expression, $\boldsymbol{u}$ and $w_q$ are

the data-dependent score vector and importance weights

$$\boldsymbol{u}(\mathbf{x}_i, \boldsymbol{\beta}, \sigma) = \nabla \log f(\sigma^{-1}(\mathbf{x}_i^T\boldsymbol{\beta} - \mu^*)), \quad w_q(\mathbf{x}_i, \boldsymbol{\beta}, \sigma) = f(\sigma^{-1}(\mathbf{x}_i^T\boldsymbol{\beta} - \mu^*))^{1-q}. \tag{5}$$

The quantity $\mathbf{p}'_\lambda(\boldsymbol{\beta})$ denotes the $(p+1)$-vector of first derivatives with elements $\{\mathbf{p}'_\lambda(\boldsymbol{\beta})\}_j = \nabla L_q(\pi(\beta_j;\lambda)) = v_q(\boldsymbol{\beta}_j, \lambda)s(\boldsymbol{\beta}_j, \lambda)$, $j = 1, \ldots, p$, and $\{\mathbf{p}'_\lambda(\boldsymbol{\beta})\}_{p+1} = 0$, where the prior scores and importance weights are

$$s(\boldsymbol{\beta}_j, \lambda) = \nabla \log \pi(\beta_j;\lambda), \quad v_q(\boldsymbol{\beta}_j, \lambda) = \pi(\beta_j;\lambda)^{1-q}. \tag{6}$$

It is important to note that the above estimating equations imply a double weighting scheme which at once controls for the importance of observations $\mathbf{x}_i$, $i = 1, \ldots, n$, and candidate coefficients $\beta_j$, $j = 1, \ldots, p$. For unusual observations incompatible with the model $f$, the relative importance of the score $\boldsymbol{u}(\mathbf{x}_i, \boldsymbol{\beta}, \sigma)$ is automatically reduced since the corresponding weights $w_q(\mathbf{x}_i, \boldsymbol{\beta}, \sigma)$ are proportional to a power-transformation of the assumed density model $f$. Particularly when $q < 1$, $w_q$ is typically small when $|\mathbf{x}_i^T\boldsymbol{\beta} - \mu^*|$ is large. Therefore, linear combinations $\mathbf{x}_i^T\boldsymbol{\beta}$ occurring far away from the target mean $\mu^*$ receive small weights. For example, this is true in the normal and t-Student models. For the penalty term, an analogous behavior is implied by the weights $v_q(\beta_j, \lambda)$, which are small if $|\beta_j|$ are large. For instance, if $\beta_j \sim N(0, \lambda^{-1})$, and $\pi$ is the normal density function, then $v_j$ is small when $\beta_j$ is far away from 0, so that the penalization is allowed to be larger when the $j$th component is close to zero.

A noteworthy special case of our approach is the popular Lasso method (Tibshirani, 1996) which is obtained when: $q \to 1$, $f(z)$ is the normal density function, and the penalty term uses the Laplace density function $\pi(\beta;\lambda) = \lambda \exp\{-\lambda|\beta|\}/2$. In the Lasso case, the data weights $w_i$, $i = 1, \ldots, n$, and parameters weights $v_j$, $j = 1, \ldots, p$, are all equal to 1 and therefore do not affect the optimization process. The constant weighting scheme implied by the Lasso leads to unstable behavior and inaccurate selections for large coefficients (Fan and Li, 2001). To improve the accuracy of the estimates Zou (2006) proposed to vary the penalty term introducing a weighting scheme, similar to our weights $v_j$, $j = 1, \ldots, p$. Differently from our approach, however, the

weights are based on OLS-estimates and the varying tuning parameter is determined by either a prior distribution or a particular expectation on the market.

# 3 Algorithms

## 3.1 Doubly re-weighted algorithm (2RE)

In this section, we provide a general algorithm for portfolio selection and then derive two important special cases of the algorithm when the working model $f$ for the data is represented by either the normal and t-Student distributions. Computing the optimal portfolios, $\widehat{\boldsymbol{\beta}}_{q,\lambda}$, by direct minimization of (1) is challenging because the terms $L_q\left\{f(\mathbf{x}_i^T\boldsymbol{\beta};\mu,\sigma)\right\}$ and $\sum_{j=1}^p L_q\left\{\pi(\beta_j;\lambda)\right\}$ are typically non-convex in the parameters. However, this issue can be efficiently addressed by noting that optimization of (1) can be divided into a sequence of simpler (convex) optimization steps. When the weights $w_q$ and $v_q$ in (4) are fixed constants, say $w_i$, $i=1,\ldots,n$ and $v_j$, $j=1,\ldots,p$, finding the solution to (4) becomes then a penalized likelihood problem. This suggests an iteratively re-weighted strategy to find the estimates, where we iterate parameter estimation by solving (4) with given weights and then update the weights based on the latest parameter estimates. Since the re-weighting is applied to both, data and penalty scores, we call this algorithm a doubly re-weighted (2RE) algorithm for portfolio selection.

For given tuning constants $q \leq 1$, $\lambda \geq 0$, and a target portfolio return $\mu^*$, the algorithm consists of the following steps:

0. At Step $s=0$, compute initial parameter values $\widehat{\boldsymbol{\beta}}^{(s)}$ and $\widehat{\sigma}^{(s)}$.

1. Set $s=s+1$, and update the data weights $\widehat{w}_i^{(s)} = f((\mathbf{x}_i\widehat{\boldsymbol{\beta}}^{(s-1)} - \mu^*)/\widehat{\sigma}^{(s-1)})^{1-q}$, $i=1,...,n$, and the penalty weights $\widehat{v}_j^{(s)} = \pi(\widehat{\beta}_j^{(s-1)};\lambda)^{1-q}$, $j=1,...,p$.

2. Find the parameter values $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\sigma}$ by minimizing

$$\sum_{i=1}^n \widehat{w}_i \log f((\mathbf{x}_i^T\boldsymbol{\beta}-\mu^*)/\sigma) + \sum_{j=1}^p \widehat{v}_j \log \pi(\beta_j;\lambda). \tag{7}$$

3. Compute $\widehat{\boldsymbol{\beta}}^{(s)}$ and $\widehat{\sigma}^{(s)}$ by solving $f((\mathbf{x}_i^T\boldsymbol{\beta}-\mu^*)/\sigma) \propto f(\mathbf{x}_i^T\widetilde{\boldsymbol{\beta}}-\mu^*)/,\widetilde{\sigma})^q$.

4. Repeat Steps 1 and 2 until a stopping criterion is satisfied.

Next we give some remarks on the 2RE algorithm. Firstly, if the initial estimates are obtained by solving (4) with equal weights $w_q(\mathbf{x}_i,\boldsymbol{\beta},\sigma)=1$, $i=1,\ldots,n$, and $v_q(\boldsymbol{\beta}_j,\lambda)=1$, $j=1,\ldots,p$, this corresponds then to the Lasso approach. However, since the Lasso solution is non-robust and typically sensitive to the presence of outliers, robust approaches for computing stable starting points could be also be considered, for example by trimming out the 10% most extreme observations for each asset before computing the Lasso estimates. Secondly, note that the re-scaling operation to correct for bias in Step 3 is due to $w_q(\cdot,\boldsymbol{\beta},\sigma)$, which arises in location-scale models, but not in pure location models. In fact, for the location-scale models, $E\left[w_q(\boldsymbol{X},\boldsymbol{\beta}_0,\sigma_0)\boldsymbol{u}(\boldsymbol{X},\boldsymbol{\beta}_0,\sigma_0)\right] \neq 0$ and a simple transformation is typically needed to re-center the estimates. For a similar estimator without the penalty term, Ferrari and La Vecchia (2012) show how to correct for such a bias; following their Proposition 1, we re-scale the parameter estimate by solving $f((\mathbf{x}_i^T\boldsymbol{\beta}-\mu^*)/\sigma) \propto f((\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}-\mu^*)/\widehat{\sigma})^q$, in $\boldsymbol{\beta}$ and $\sigma$ where $(\widehat{\boldsymbol{\beta}},\widehat{\sigma})$ are the solution to (4). For example, if $f$ is the normal pdf, then we take $\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}$ and $\sigma=\widehat{\sigma}/\sqrt{q}$ as final estimates. Finally, a second potential source of bias is given by the penalty function $\sum_j L_q(\pi(\beta_j;\lambda))$ in criterion (1); the heuristic derivation in the Appendix illustrates how this bias can be controlled by choosing sufficiently large values of $\lambda$.

## 3.2 Algorithm for normal portfolios

An important model in portfolio theory is when the assets $\boldsymbol{X}$ are assumed to follow a $p$-variate normal distribution. The model for $Y$ is then the univariate normal distribution $N_1(\mu,\sigma^2)$. If we choose $\pi(\beta_j;\lambda)=\lambda\exp\{-\lambda|\beta_j|\}/2$, $j=1,\ldots,p$, Step 2 of the algorithm in Section (3.1) computes the portfolio estimates by solving

$$\widehat{\boldsymbol{\beta}}^{(s)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \widehat{w}_i^{(s-1)} \frac{1}{2} \left( \frac{\mu^* - \mathbf{x}_i^T\boldsymbol{\beta}}{\widehat{\sigma}^{(s-1)}} \right)^2 + \lambda \sum_{j=1}^{p} \widehat{v}_j^{(s-1)} |\beta_j| \right\}, \tag{8}$$

and the weights $\widehat{w}_i$ and $\widehat{v}_j$ are computed using estimates obtained in Step $s-1$ as follows:

$$\widehat{w}_i^{(s-1)} = \left[ \frac{1}{\sqrt{2\pi\widehat{\sigma}^{2(s-1)}}} \exp\left\{ -\frac{\left(\mu^* - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}^{(s-1)}\right)^2}{2\widehat{\sigma}^{2(s-1)}} \right\} \right]^{1-q}, \qquad \widehat{v}_j^{(s-1)} = \left[ \frac{\lambda}{2} \exp\left\{ -\lambda |\widehat{\beta}_j^{(s-1)}| \right\} \right]^{1-q}. \tag{9}$$

The portfolio variance is also updated using estimates from Step $s-1$ as

$$\widehat{\sigma^{2(s)}} = \frac{\sum_{i=1}^n \widehat{w}_i^{(s-1)} \left(\mu^* - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}^{(s-1)}\right)^2}{q \sum_{i=1}^n \widehat{w}_i^{(s-1)}}. \tag{10}$$

When the portfolio variance is not estimated and instead set equal to a fixed target value, say $\sigma^{*2}$, then we have $\widehat{\sigma}^{2(s)} = \sigma^{*2}$, for all $s \geq 0$.

The updating rule in (8) is a weighted $L_1$-penalized quadratic problem (Lasso) and can be solved efficiently using existing algorithms. Tibshirani (1996) and Turlach et al. (2005) proposed to use quadratic optimization to solve the $L_1$-penalized optimization problem, while Friedman et al. (2007) developed a coordinate-wise approach that works efficiently with convex optimization problems. To compute (8), we use the gradient projection algorithm developed by Figueiredo et al. (2007), as Gasso et al. (2009) have shown that such algorithm is more efficient to solve problems as ours compared to quadratic programming and coordinate-wise optimization.

### 3.3 Expectation-Maximization algorithm for t-portfolios

Assume that the portfolio $Y = \boldsymbol{\beta}^T \boldsymbol{X}$ follows a non standardized t-Student distribution with mean $\mu$ and variance $\sigma$ with probability density function

$$f_\nu(y; \mu, \sigma) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}, \tag{11}$$

where $\nu$ is the number of degrees of freedom and $\Gamma$ is the gamma function. The value of $\nu > 1$ and $\mu^*$ are fixed and will not be estimated. Under the t-Student model, for each Step $s \geq 1$ we

compute $\{\widehat{\boldsymbol{\beta}}^{(s)}, \widehat{\sigma}^{(s)}\}$ by solving

$$\underset{\boldsymbol{\beta},\sigma}{\operatorname{argmin}} \left\{ -\left(\frac{\nu+1}{2}\right) \sum_{i=1}^{n} \widehat{w}_i^{(s-1)} \log\left\{1 + \frac{(\mathbf{x}_i\boldsymbol{\beta}^T - \mu^*)^2}{\nu\sigma^2}\right\} + \lambda \sum_{j=1}^{p} \widehat{v}_j^{(s-1)} |\beta_j| \right\}, \tag{12}$$

where $\sigma > 0$ and $\nu > 1$ and the estimation weights $\widehat{w}_i$ is updated from estimates obtained in Step $s-1$ as follows

$$\widehat{w}_i^{(s-1)} = \left[ f_\nu \left( \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}^{(s-1)}; \mu^*, \widehat{\sigma}^{(s-1)} \right) \right]^{1-q}, \quad i = 1, \ldots, n, \tag{13}$$

where $f_\nu$ is the probability density function in (11) and the penalty weights $\widehat{v}_j$ have the same expression as in (9).

In this case, the updating rule (12) (Step 2 of the general algorithm in Section 3.1) represents a non-convex problem. Thus, solving (12) can lead to unreliable estimates, especially when the number of assets $p$ is large. Under the t-Student model, a stable approach for (weighted) likelihood optimization is given by the Expectation-Maximization (EM) algorithm. The EM algorithm exploits the well-known fact that a t-Student distribution can be equivalently represented as a scale mixture of normals; particularly, one observation from the t-Student model can be written as $Y_i \sim N(\mu, \sigma^2 Z_i^{-1})$ where $Z_i$ follows a Gamma distribution $Z_i \sim \text{Ga}(\nu/2, \nu/2)$ (McLachlan and Krishnan, 2007). This considerations motivate the following EM steps, which specify Step 2 of the algorithm in Section (3.1) for the t-Student distribution.

For any $s > 0$, given robust weights $\widehat{w}_i^{(s-1)}$ and $\widehat{v}_i^{(s-1)}$ obtained in Step $s-1$, first set the initial mixture weights $\widehat{z}_i = 1/n$, $i = 1, \ldots, n$. Then, obtain the updated estimates $\widehat{\boldsymbol{\beta}}^{(s)}$ and $\widehat{\sigma}^{(s)}$ by iterating the following EM steps

- *M-Step:* This step computes weighted parameter estimates $\boldsymbol{\beta}$ and $\sigma$ using the mixing con-

stants $\widehat{z}_i$:

$$\boldsymbol{\beta}' = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \widehat{w}_i^{(s-1)} \widehat{z}_i^{(s-1)} \frac{1}{2} \left( \frac{\mathbf{x}_i^T \boldsymbol{\beta} - \mu^*}{\widehat{\sigma}^{(s-1)}} \right)^2 + \lambda \sum_{j=1}^{p} \widehat{v}_j^{(s-1)} |\beta_j| \right\}, \tag{14}$$

$$\sigma'^2 = \frac{\sum_{i=1}^{n} \widehat{w}_i^{(s-1)} \widehat{z}_i^{(s-1)} \left( \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} - \mu \right)^2}{\sum_{i=1}^{n} \widehat{w}_i^{(s-1)} \widehat{z}_i^{(s-1)}} \times \frac{\nu}{(\nu+1)q - 1}. \tag{15}$$

where $\widehat{w}_i^{(s-1)}$ is proportional to the t-Student probability density function as in (13) evaluated in $\widehat{\boldsymbol{\beta}}^{(s-1)}$, $\widehat{\sigma}^{(s-1)}$, the parameter estimates obtained in Step $s-1$.

- *E-Step:* Update the mixing constants as follows:

$$\widehat{z}_i = \frac{(\nu_q + 1)\sigma'^2}{\nu_q \sigma'^2 + \widehat{w}_i^{(s-1)} (\mathbf{x}_i^T \boldsymbol{\beta}' - \mu)^2}, \quad i = 1, \ldots, n, \tag{16}$$

where $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}$ are computed in the M-Step and $\nu_q = (\nu+1)q - 1$.

In the remainder of this section, we derive the above M- and E-Steps. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, with $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}$, be the vector of portfolio returns such that $Y_i \sim N(\mu, \sigma^2/Z_i)$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$ be a random vector such that $Z_i \sim \text{Gamma}(\nu/2, \nu/2)$. We also denote by $w_i$ and $v_j$ the robustness weights in (13) and (9) respectively, which are fixed constants in the EM. Then, the complete (re-weighted and penalized) log-likelihood function is

$$\begin{aligned} \log L(\boldsymbol{Y}, \boldsymbol{Z}; \boldsymbol{\beta}, \sigma) &= \sum_{i=1}^{n} w_i \left\{ \log N\left( Y_i; \frac{\sigma^2}{Z_i} \right) + \log \text{Ga}\left( Z_i; \frac{\nu}{2}, \frac{\nu}{2} \right) \right\} + \sum_{j=1}^{p} v_j \log \pi(\beta_j; \lambda) \\ &\propto -\sum_{i=1}^{n} w_i \frac{\log \sigma^2}{2} - \sum_{i=1}^{n} w_i Z_i \left( \frac{\boldsymbol{X}_i^T \boldsymbol{\beta} - \mu^*}{2\sigma^2} \right)^2 - \lambda \sum_{j=1}^{p} v_j |\beta_j|, \end{aligned} \tag{17}$$

and in the last expression we drop all the terms not depending on $\boldsymbol{\beta}$ or $\sigma$. Next, we compute the expected value of the complete log-likelihood function, with respect to the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{X}$ under the current estimates of the parameters, denoted by $\boldsymbol{\beta}'$ and $\sigma'$:

$$E_{\boldsymbol{Z};\boldsymbol{Y}} \left[ \log L(\boldsymbol{Y}, \boldsymbol{Z}; \boldsymbol{\beta}, \sigma); \boldsymbol{\beta}', \sigma' \right] = \log L\left( \boldsymbol{Y}, E_{\boldsymbol{Z};\boldsymbol{Y}} \left[ \boldsymbol{Z}; \boldsymbol{\beta}', \sigma' \right]; \boldsymbol{\beta}, \sigma \right), \tag{18}$$

where the equality follows from the fact that (17) is linear in $\boldsymbol{Z}$. Since the Gamma distribution is the conjugate prior for the normal scale parameter, we have that the posterior distribution of $Z_i$, given $Y_i' = \boldsymbol{X}_i\boldsymbol{\beta}'$, is

$$Z_i \sim \text{Ga}\left(\frac{\nu}{2}, \frac{\nu}{2} + w_i\left(\frac{Y_i' - \mu^*}{2\sigma'^2}\right)^2\right), \tag{19}$$

which implies

$$\widehat{z}_i = E_{Z_i}\left[Z_i; \boldsymbol{\beta}', \sigma'\right] = \frac{(\nu+1)\sigma'^2}{\nu\sigma'^2 + w_i(\boldsymbol{X}_i^T\boldsymbol{\beta}' - \mu^*)^2}, \quad i = 1, \ldots, n. \tag{20}$$

The last expression is the E-Step in (16). The updating formulas for the parameters (14) and (15) in the M-Step are simply obtained by maximizing the complete log-likelihood (17) with $Z_i$ replaced by its conditional expectation $\widehat{z}_i$.

### 3.4 Robust selection of $\lambda$

The parameter $\lambda$ controls for the size of the penalty. Large $\lambda$ values imply more parsimonious portfolios with a larger number of assets having zero weight. In the literature of penalized regression, similar tuning constants are chosen by information theoretical criteria, such as the AIC and BIC selection criteria yielding good empirical performances (Zhang et al., 2010). We will follow the same approach here. Many authors, including Ronchetti (1997), Shi and Tsai (1998) and Machado (1993), have highlighted the non-robust nature of the traditional information theoretical criteria and stressed instead the importance of robust model selection procedures. Following Ronchetti (1997) and Machado (1993), for a given $q$ we choose optimal values of $\lambda$ by minimizing the following robust Bayesian Information Criterion (BIC):

$$\text{BIC}_q = -2\sum_{i=1}^{n} L_q\left\{f\left(\frac{\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}_{q,\lambda} - \mu^*}{\hat{\sigma}_{q,\lambda}}\right)\right\} + \log(n)k, \tag{21}$$

where $k \leq p$ is the number of active portfolio positions and the traditional BIC is obtained when $q \to 1$. Note that when the penalty $\log(n)k$ is replaced by $2k$ in (21) we obtain the robust AIC

approach by Ronchetti (1997).

# 4 Monte Carlo simulations

We evaluate the performance of the proposed approaches for normal and t-Student portfolios (denoted by $GDL_N$ and $GDL_t$, respectively) and compare our method to other popular penalization schemes. The data are simulated from the following models:

1. *Model 1:* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mu_j = 1$, if $j \leq k$, and $\mu_j = 0$, if $j > k$, and the covariance matrix is such that $\Sigma_{jj} = 1$, $j = 1, \ldots, p$, and off-diagonal elements $\Sigma_{jk} = \rho$, $0 \leq \rho < 1$, $j \neq k$.

2. *Model 2:* Multivariate t-Student distribution $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, with $\nu$ degrees of freedom. The mean-variance structure is as in Model 1.

The above models reflect different levels of correlation for the assets $\boldsymbol{X}$ and possible presence of fat tails. For each model, we generate $B = 250$ samples with $n = 500$, $p = 50$, $k = 10$ and correlation $\rho$ ranging from 0.2 to 0.8. Models 1 and 2 represent the situation in which the assets are characterized by heterogeneous returns and homogeneous risk; choosing only a few assets characterized by positive returns can achieve a positive target return for the overall portfolio. The normality assumption is the standard assumption in the modern portfolio theory introduced by Markowitz (1952). However, as shown in Figure 5, the t-Student distribution can be a more suitable model in real applications due to leptokurtic distribution of the data. Therefore, both normal and t-Student distributed data have to be investigated in our simulations. Given $B$ samples from one of the above models, we assess the overall performance with respect to a target value $\mu^* = k$ by the Monte Carlo mean squared error (MSE)

$$\widehat{MSE} = \frac{1}{B} \sum_{b=1}^{B} \left( \frac{\boldsymbol{\mu}^T \widehat{\boldsymbol{\beta}}_b - \mu^\star}{\sqrt{\widehat{\boldsymbol{\beta}}_b^T Var(\boldsymbol{X}) \widehat{\boldsymbol{\beta}}_b}} \right)^2 , \tag{22}$$

where $\boldsymbol{\mu}^T \widehat{\boldsymbol{\beta}}_b$ and $\widehat{\boldsymbol{\beta}}_b^T Var(\boldsymbol{X}) \widehat{\boldsymbol{\beta}}_b$ are the mean and variance of the $b$-th optimal portfolio in the $bth$ Monte Carlo run. To assess sparsity, we calculate the number of estimated active components as

$\widehat{k} = \sum_{j=1}^{p} I(|\widehat{\beta}_j| > \tau)$, where $I(\cdot)$ is the indicator function and $\tau = 0.005$ is a threshold value, below which asset weights are set equal to zero. To evaluate the model selection properties, we compute the F-measure as follows:

$$\text{F-measure} = 2\frac{|\text{supp}(\boldsymbol{\beta}^*)| \cap |\text{supp}(\widehat{\boldsymbol{\beta}})|}{|\text{supp}(\boldsymbol{\beta}^*)| + |\text{supp}(\widehat{\boldsymbol{\beta}})|}, \tag{23}$$

where the support of a vector $\boldsymbol{\beta}$ is defined as $\text{supp}(\boldsymbol{\beta}) = \{j : |\beta_j| \geq \tau\}$ and $\boldsymbol{\beta}^*$ is the vector with values equal to 1 in the first $k$ positions. The larger the F-measure value, the better the model selection performance. In fact, a F-measure equal to 1 corresponds to the optimal situation in which only the $k$ active assets in $\boldsymbol{\beta}^*$ are selected.

We compare the resulting estimates with analogous estimates obtained by the following methods: Lasso (Tibshirani (1996)), the Zhang-penalty method by Gasso et al. (2009) with $\eta = 0.2$, and the logarithm penalization by Weston et al. (2003) with $\phi = 1.5$. For each method we select the solutions with the lowest BIC, as explained in Section 3.4. All the methods except Lasso involve the solution of non-convex problems, which we solve by the DC-programming approach proposed by Gasso et al. (2009). The optimization relies on iterative primal-dual approach for the non-convex penalty functions considered, where in each iteration a convex primal problem is solved by the gradient projection algorithm of Figueiredo et al. (2007). To initialize all the algorithms we used ordinary least squares estimates, except for the EM algorithm of the $GDL_t$ approach. In our numerical experiments, we noticed that the convergence of the EM algorithm can be sensitive to the initial estimates of the $\beta$ coefficients, as also discussed in the statistical literature, so we use the robust estimates obtained by $GDL_N$ approach as initial values. Both $GDL_t$ and $GDL_N$ are initialized using equal weights $w_i = 1/n,\ i = 1, \ldots, n$, and $v = 1/p,\ j = 1, \ldots, p$. The iterative algorithm for normal portfolios and the EM algorithm converge after a small number of iterations. For Model 1, the Monte Carlo average for the number of iterations based on $B = 150$ runs was 7.11 (se=0.17) and 3.91 (se=0.32) for $GDL_N$ and $GDL_t$, respectively.

## 4.1 Empirical Results

**Sparsity and accuracy** Figure 2 and Figure 3 show boxplots for the number of active positions, F-measure and MSE, obtained under mildly and strongly correlated data ($\rho = 0.2, 0.8$ and $q = 0.9$). In terms of sparsity, the GDL algorithms perform considerably better than all other approaches in the first two models (panels (a) and (d)), regardless of the correlation level. The number of active regressors is close to the optimal value $k = 10$. The GDL always outperforms the Lasso and Zhang penalties, which give over-fitting solutions with too many coefficients different from zero. The Log penalty shows a good performance when $\rho = 0.8$. This is not surprising as previous studies have already shown the considerable model selection accuracy of the Log penalty in the presence of highly correlated data (Gasso et al., 2009). Panels (b) and (e) show that the GDL algorithms not only select the correct number of active regressors, but also tend to select the right assets more frequently compared to the other methods. Particularly, the boxplots of the F-measure are centered on larger values than those of other approaches. In terms of MSE, the Lasso has the worst overall performance, due to selecting too many unnecessary active weights (panels (c) and (f)). The good model selection performance of the GDL approaches comes with a relatively small price in terms of MSE compared to the Zhang and Log penalties. The $GDL_t$ outperforms the $GDL_N$, due to the advantage in terms of good initialization and the larger penalization for the observations in the tails.

**Effect of correlation** Comparing Figure 2 and Figure 3 shows that the GDL approaches work well regardless the amount of correlation, while the other approaches tend to keep too many variables, with the Log penalty being the only exception for $\rho = 0.8$. Clearly, this leads to smaller values of MSE, while the model selection performance, as quantified by the F-measures, is still inferior to the GDL approaches.

**Role of the tuning constant** $q$ Figure 4 shows the boxplots for the $GDL_t$ for $q = 0.5, 0.7, 0.9$ when $\rho = 0.2$ and 0.6. For Model 1, smaller values of $q < 1$ lead to a more parsimonious selection with fewer active components, while better F-measure values are reached for $\rho = 0.2$ and $q = 0.9$. In terms of MSE, in Model 1 we expect the $GDL_t$ approach with $q = 0.9$ to perform best, as there

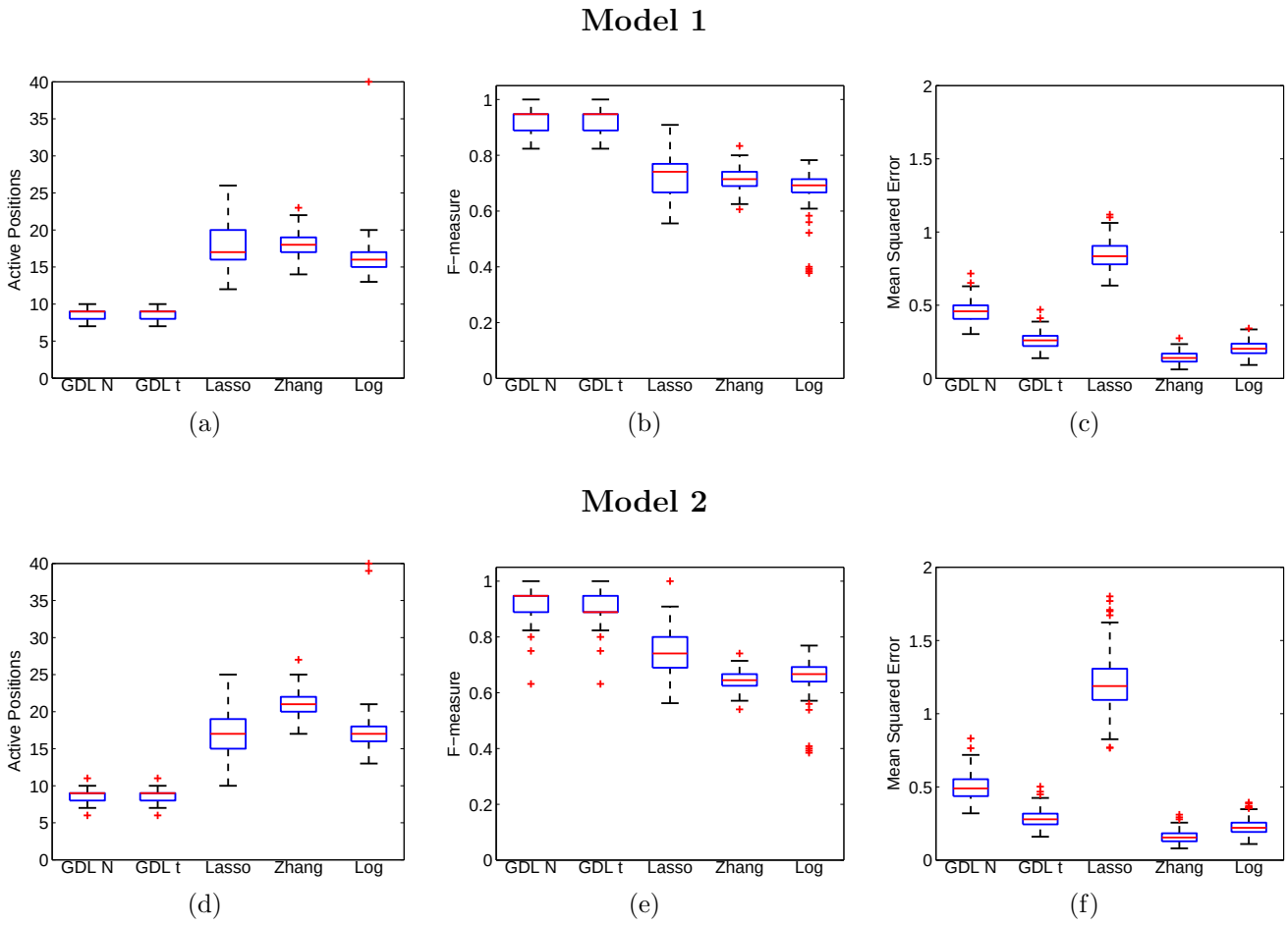



Figure 2: Monte Carlo simulations - Mildly correlated data. From left to right: Number of estimated active positions, F-measure and MSE for different selection methods: GDL for Normal and t-Student, Lasso, Zhang and Log penalty. The boxplots are based on 250 samples of size $n = 500$ from Model 1 (first row) and Model 2 (second row) with $\rho = 0.2$, $q = 0.9$, $p = 50$ and $k = 10$.

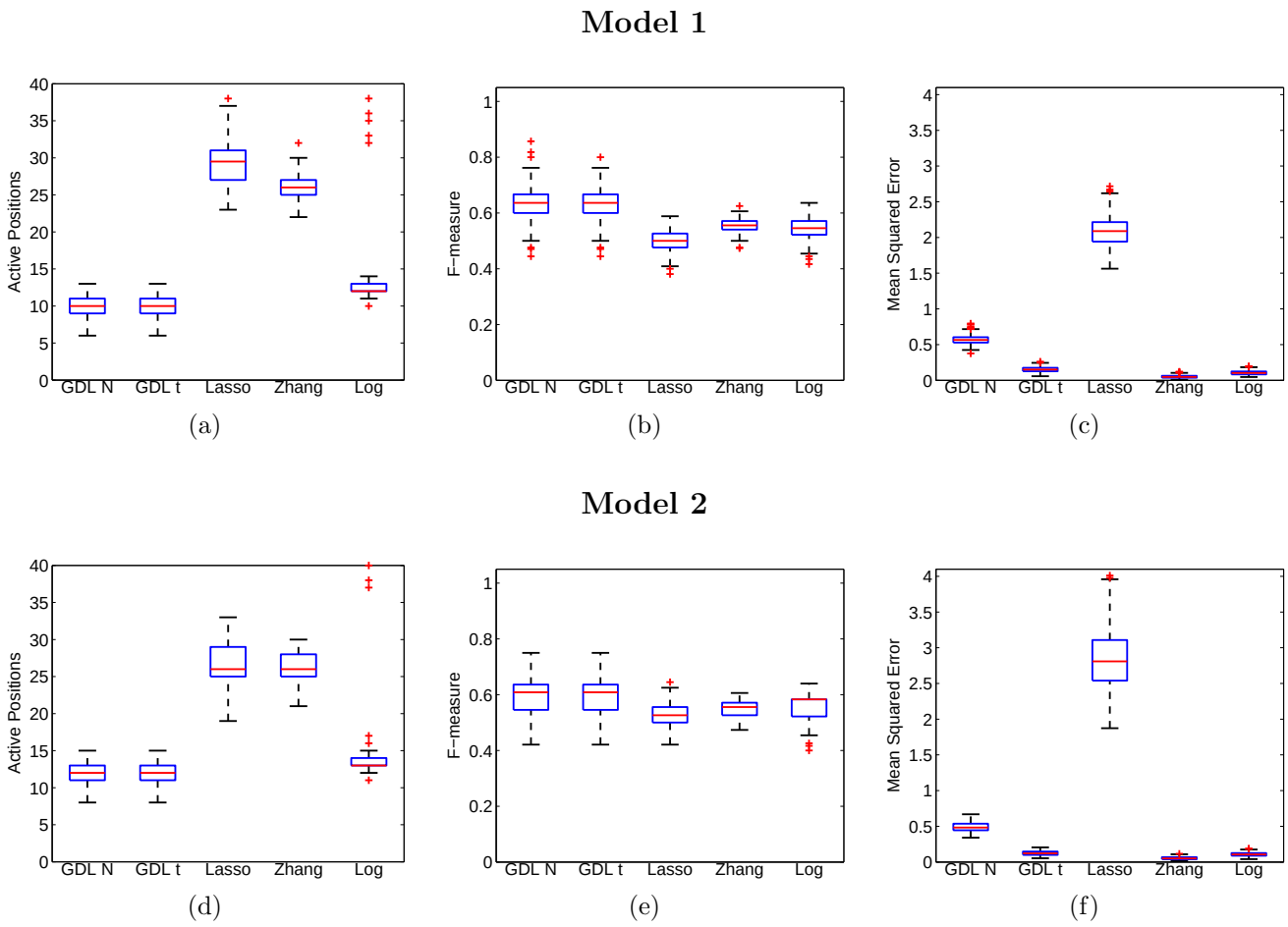



Figure 3: Monte Carlo simulations - Highly correlated data. From left to right: Number of estimated active positions, F-measure and MSE for different selection methods: GDL for Normal and t-Student, Lasso, Zhang and Log penalty. The boxplots are based on 250 samples of size $n = 500$ from Model 1 (first row) and Model 2 (second row) with $\rho = 0.8$, $q = 0.9$, $p = 50$ and $k = 10$.

is no need to robustify the estimates with respect to potential extreme values. For Model 2, using the $GDL_t$ accounts automatically for the possible presence of fat tails, leading to satisfactory results for all $q$ levels. However, when the correlation level increases, the degree of robustness given by $q = 0.9$ does not give sufficiently sparse models.

Summing up, the optimal choice of $q$ depends on the distribution of the data. A larger value of $q$ should be preferred when data do not exhibit fat tails, since there is no need to trade off bias for variance to improve the quality of the estimates. On the other hand, smaller values of $q$ help to reduce the number of active beta in presence of high correlation and fat tails, despite paying a small price in terms of accuracy. Clearly, while we currently set a priori the value of $q$, it is high on our agenda to develop an automatic data-driven procedure for its optimal choice.

# 5 Index Tracking

One of the most challenging problems in asset allocation is the so-called index-tracking problem. The aim of index tracking is to replicate the performance of a financial index by using a relatively small subset of its constituents, with the advantage of managing transaction and monitoring costs. Typically, the problem can be set-up as a regression problem with a so-called budget constraint (i.e. the sum of the coefficients of the $\boldsymbol{\beta}$ vector has to sum to 1) and a constraint on the 0-norm of the $\boldsymbol{\beta}$ vector, by imposing a maximum number of active positions. The latter constraint and the typical large dimensionality of the problem makes then the optimization NP-hard and different solutions have been proposed in the literature (see Beasley et al. (2003) for a review). Recently, Giamouridis and Paterlini (2010) and Fastrich et al. (forthcoming) have shown that using a Lasso or a non-convex approach can be beneficial in hedge fund replication and index tracking, by allowing to automatically select sparse solutions with good out-of-sample properties.

In this section, we compute the optimal tracking sparse portfolios for a set of financial indexes with a different number of constituents. We consider daily observations from 23.08.2002 to 27.03.2008 of three financial indexes, Fama & French 100, S&P 200 and S&P 500, with number of constituents $p$ equal to 100, 200 and 500, respectively. In Figure 5, we show the log-returns for all indexes along with the fitted normal and t-Student distributions, obtained using maximum likelihood

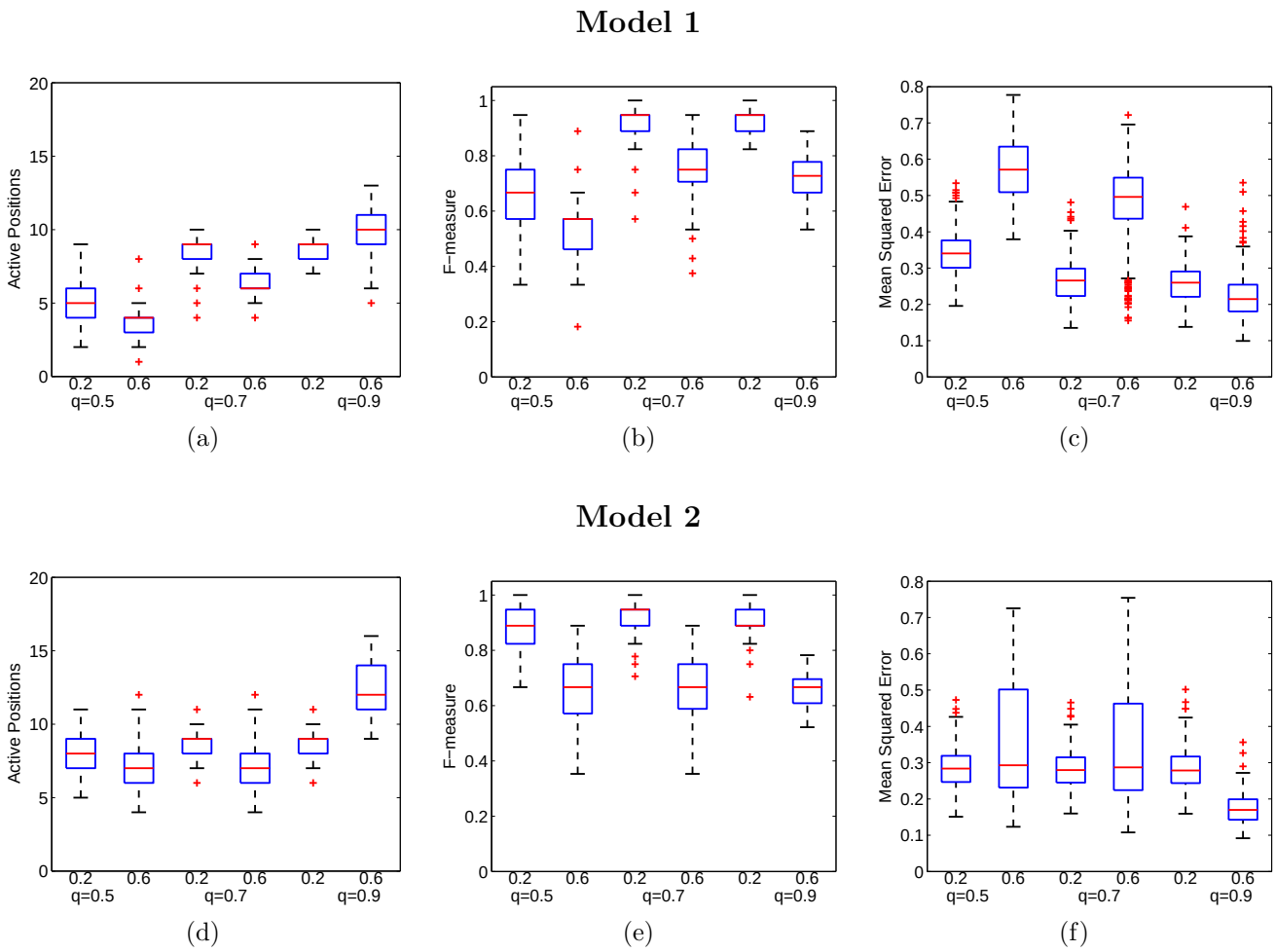
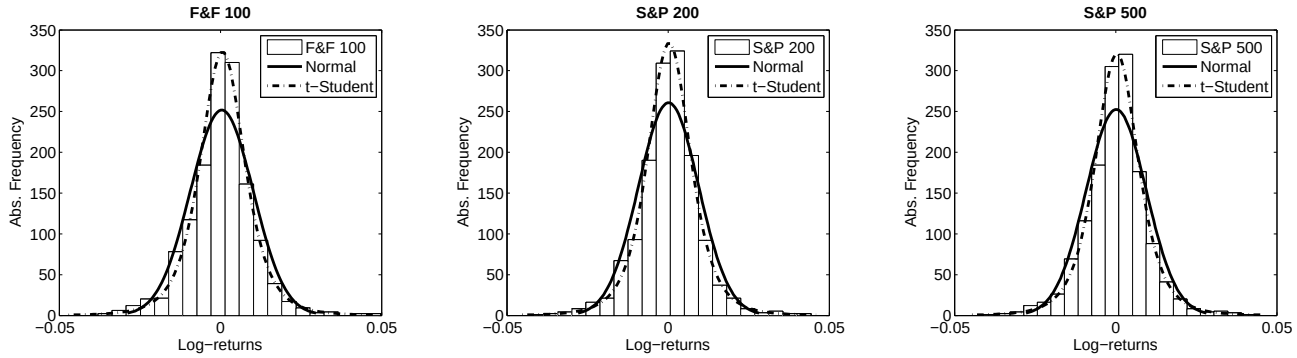



Figure 4: Monte Carlo simulations - Role of $q$. From left to right: Number of estimated active positions, F-measure and MSE for $GDL_t$. The boxplots are based on 250 samples of size $n = 500$ from Model 1 (first row) and Model 2 (second row) with different levels of correlation $\rho = 0.2, 0.6$, $q = 0.5, 0.7, 0.9$, $p = 50$ and $k = 10$.

estimation. The data are slightly asymmetric and leptokurtic and a satisfactory fit is given by the t-Student distribution with approximately 4 degrees of freedom. The F&F 100 is a capitalization-weighted index where 8.16% of the assets have weight larger than 5%, 11.23% have weights values between 1% and 5%, and the remaining 80.61% have weights smaller than 1%. The S&P indexes are market-value weighted indexes, characterized by much smaller positions in the constituents: no weight is larger than 5% and the majority of weights are smaller than 1% (89% for S&P200, 95.80% for S&P 500).




(a) (b) (c)

Figure 5: Fitted Normal and t-Student curves on histograms of F&F 100 (a), S&P 200 (b) and S&P 500 (c)

Our aim is to build sparse portfolios that are able to track as closely as possible the index performance over time. For each dataset, we compare six tracking techniques: the $GDL_N$ and the $GDL_t$ approaches with $q = 0.9$ and $q = 0.5$, the Lasso, and the equally weighted strategy $(1/N)$. We test the out-of-sample performance of these methods adjusting the asset weights periodically based on a rolling window of 250 in-sample observations, moving each time ahead by one observation and then discarding the oldest data point. For each window, we set $\mu^*$ equal to the mean of the index in-sample returns and then select the optimal portfolios according to the BIC criterion, as described in Section (3.4). The out-of-sample excess returns over the index is then computed with respect to the next observation in time. The rational of our study is to take the viewpoint of an investor who decides his optimal allocation by exploiting the information in the in-sample window and then holds the portfolio for one day, before revising it. We assess the portfolio selection by looking at different performance measures. The risk/return performance is

quantified by the Information Ratio (IR), which corresponds to the ratio of the mean excess return (ER) and the tracking error volatility (TEV) relative to the index. The excess return time series is determined by the difference between the index and the tracking portfolio on the entire out-of-sample period, while the tracking error volatility is the standard deviation of the out-of-sample excess return. To assess the tracking ability, we compute the portfolio out-of-sample correlation (Cor) with respect to the index. Finally, to evaluate sparsity and stability of portfolios, which have a direct impact on transaction costs, we compute the number of active positions and the average turnover (TO).

| Strategy | ER (%) | TEV (%) | IR | $\bar{k}$ | TO | Cor |
|---|---|---|---|---|---|---|
| **PANEL A: F&F 100** | | | | | | |
| $GDL_N$ $q = 0.9$ | 0.338 | 0.624 | 0.542 | 37.749 | 0.068 | 0.999 |
| $GDL_N$ $q = 0.5$ | 0.302 | 0.659 | 0.458 | 38.244 | 0.060 | 0.999 |
| $GDL_t$ $q = 0.9$ | 0.170 | 0.492 | 0.346 | 32.241 | 0.066 | 0.999 |
| $GDL_t$ $q = 0.5$ | 0.186 | 0.527 | 0.352 | 32.121 | 0.064 | 0.999 |
| Lasso | 1.030 | 2.117 | 0.486 | 65.939 | 0.017 | 0.990 |
| 1/N | 0.929 | 3.854 | 0.241 | 98 | 0 | 0.960 |
| **PANEL B: S&P 200** | | | | | | |
| $GDL_N$ $q = 0.9$ | 0.319 | 4.500 | 0.071 | 36.950 | 0.399 | 0.963 |
| $GDL_N$ $q = 0.5$ | 0.639 | 4.961 | 0.129 | 43.627 | 0.346 | 0.967 |
| $GDL_t$ $q = 0.9$ | -2.421 | 4.897 | -0.494 | 28.431 | 0.520 | 0.933 |
| $GDL_t$ $q = 0.5$ | -0.196 | 4.614 | -0.042 | 28.011 | 0.481 | 0.960 |
| Lasso | 4.760 | 7.267 | 0.655 | 66.532 | 0.037 | 0.950 |
| 1/N | 5.937 | 2.518 | 2.357 | 200 | 0 | 0.971 |
| **PANEL C: S&P 500** | | | | | | |
| $GDL_N$ $q = 0.9$ | 2.906 | 6.966 | 0.417 | 44.564 | 0.605 | 0.932 |
| $GDL_N$ $q = 0.5$ | -1.989 | 8.100 | -0.245 | 45.736 | 0.322 | 0.947 |
| $GDL_t$ $q = 0.9$ | 1.192 | 9.018 | 0.132 | 27.770 | 0.811 | 0.872 |
| $GDL_t$ $q = 0.5$ | 2.287 | 8.859 | 0.258 | 26.944 | 0.801 | 0.902 |
| Lasso | 2.986 | 10.315 | 0.289 | 66.407 | 0.053 | 0.926 |
| 1/N | 5.113 | 3.107 | 1.646 | 500 | 0 | 0.962 |

Table 1: Out-of-sample statistics of each tracking portfolio computed by different strategies: annualized excess return ER in percentage, tracking error volatility TEV, Information Ratio IR, average number of active components $\bar{k}$, turnover TO, correlation w.r.t. index Cor.

Table 1 reports the out-of-sample statistics of the six portfolios for each dataset. As expected, in terms of risk/return performance, the $1/N$ portfolio is a tough benchmark to beat, especially for S&P 200 and S&P 500, which are market weighted indexes with about 90% of weights smaller

than 1%, while F&F 100 is a value-weighted index. However, investing in the equally weighted portfolio would imply having a position in each constituents, which could be undesirable due to monitoring and transaction costs. Comparing the IR results of the five remaining strategies (Column 4) shows that for F&F 100 and S&P 500 the $GDL_N$ ($q$=0.9) has the largest IR (even larger than the equally weighted for the first index), while the Lasso outperforms the other methods for S&P 200. Since the target here is to track an index as closely as possible and not to beat it, a good tracking portfolio should have excess returns close to zero, as well as ideally the smallest out-of-sample tracking volatility. We observe that the GDL strategies always have a lower out-of-sample TEV than Lasso (Column 3), which is consistent with our simulation analysis, showing that the GDL approaches usually identify solutions with smaller MSE and still sparser than the Lasso. According to the average number of active positions $\bar{k}$ in Column 5, the smaller levels of TEVs of the GDL approaches are obtained by using approximately 35% of the available positions for F&F 100 (Panel A), less than 25% for S&P 200 (Panel B) and less than 10% for S&P 500 (Panel C). Lasso identifies optimal tracking portfolios with always more than 65 constituents for each index. However, investing in a larger number of constituents gives a turnover rate lower than the GDL, as shown in Column 6. The GDL portfolios have values of correlation with respect to the index very close to 1. In particular, our method outperforms the Lasso portfolio reaching better out-of-sample tracking volatility. This aspect is illustrated in Figure 6, where we show the out-of-sample returns, rebased to 100, for the $1/N$, Lasso and the GDL approaches. A portfolio close to the index indicates a good out-of-sample tracking performance. The GDL approach provides satisfactory replications of the index by using a relatively small subset of its components. Instead, 1/N and Lasso show returns patterns which, despite outperforming the indexes in the selected period, are often far away from the indexes.

# 6 Conclusion

This work proposes a new objective function and then develops the related algorithm for financial portfolio selection. The main rationale of our approach is to select asset weights that minimize two

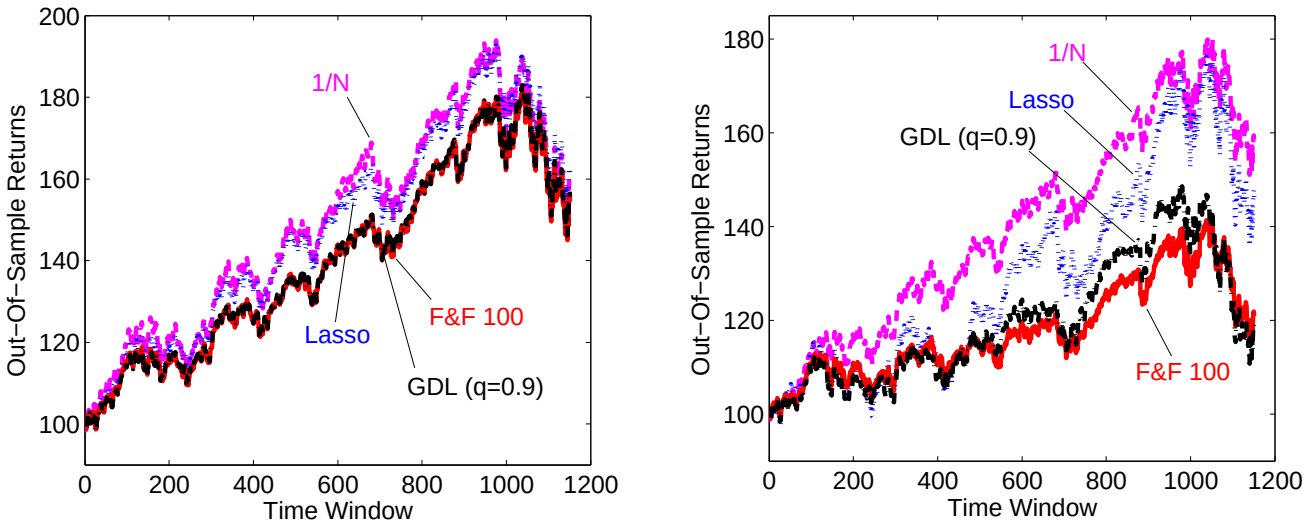



Figure 6: Normalized trends of the tracking portfolios with respect to the corresponding indexes: F&F 100 (a) and S&P 200 (b).

sources of uncertainty. One source is related to the choice of the model for portfolio returns (e.g., normal or t-Student models); while the other pertains to the choice of the model structure itself. Differently from other criteria for portfolio selection, the information is coded by the $q$-entropy, a generalized information measure, which has been previously shown to be useful for inference due to its robustness properties in the presence of model misspecification (Ferrari and La Vecchia, 2012). The resulting optimization task may be regarded as the optimization of a robust penalized likelihood function, subject to a constraint depending on the prior density model for the portfolio coefficients or asset weights. To compute the estimates, we developed an iteratively re-weighted algorithm when the underlying models for the asset returns are the normal or t-Student models, which are typical choices in the asset allocation literature. In this work we explored the use of a Laplace density as a prior density model for the coefficient weights; such a choice is shown to be useful when the goal is to obtain sparse portfolios (i.e. portfolios with a relatively small number of active weights).

The empirical findings in Section 4 show a good performance of the proposed method in most of our simulation settings when compared to other state-of-the-art approaches, pointing out that

the GDL criterion can be a useful tool for portfolio selection. Besides the robustness to possible misspecifications of the normal or t-Student models for the data, another important aspect emerging from our empirical study is the robustness to the presence of strong correlation. This feature is not shared by other popular approaches such as the Lasso method (Tibshirani, 1996), whose performance is instead negatively influenced by high levels of correlation. The robustness to model misspecification is achieved through the well-established properties of the first term of our objective function in Equation (1), which are already well-studied in the literature of M-estimation. The robustness to strong correlation, however, is related to the form of the second term of criterion (1). Our empirical findings suggest that further research on the theoretical properties of the proposed estimator in the context of a more general regression setting may be fruitful. Moreover, given that the overall behavior of the GDL criterion in terms of sparsity and robustness is based on the choice of the tuning parameters $\lambda$ and $q$, it is a priority in our agenda to develop automatic methods for the optimal selection of the two parameters.

In Section 5, we consider the index-tracking problem. Ideally, a tracking strategy should select portfolios with good out-of-sample replicating performance and a small number of constituents to limit transaction costs. Indeed, the GDL approach selects sparse portfolios with remarkable out-of-sample performance. In our empirical study we found that the optimal GDL portfolios outperform the Lasso portfolios, by achieving a smaller out-of-sample tracking error volatility despite investing in a lower number of assets. The proposed approach also outperforms the equally-weighted strategy in terms of sparsity and tracking ability. Only for the S&P 500 dataset, the equally-weighted portfolio shows a lower out-of-sample tracking volatility. The large diversification in the S&P 500 index implies that actual proportions of its assets are not much different from those of the $1/N$ portfolios.

# References


J.E. Beasley, N. Meade, and T.-J. Chang. An evolutionary heuristic for the index tracking problem. *European Journal of Operational Research*, 148:621–643, 2003.

M. J. Best and R. R. Grauer. On the sensitivity of mean-variance efficient portfolios to changes in asset means: some analytical and computational results. *The Review of Financial Studies*, 4(2):315342, 1991.

M. Carrasco and N. Noumon. Optimal portfolio selection using regularization. Working Paper University of Montreal, 2012.

R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1:223–236, 2001.

V. De Miguel and F. J. Nogales. Portfolio selection with robust estimation. *Operations Research*, 57:560–577, 2009.

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

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, 96:1348–1360, 2001.

J. Fan, J. Zhang, and K. Yu. Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107:498:592–606, 2012.

B. Fastrich, S. Paterlini, and P. Winker. Cardinality versus q-norm constraints for index tracking. *Quantitative Finance*, ISSN: 1469-7696, doi: 10.1080/14697688.2012.691986, forthcoming.

D. Ferrari and D. La Vecchia. On robust estimation via pseudo-additive information. *Biometrika*, 99(1):238–244, 2012.

M. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Convex Optimization Methods for Signal Processing*, 1, 4: 586–598, 2007.

J. Friedman, T. Hastie, H. Hfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1, 2:302–332, 2007.

P. Frost and J.E. Savarino. For better performance: Constrain portfolio weights. *Journal of Portfolio Management*, 15(1):2934, 1988.

G. Gasso, A. Rakotomamonjy, and S. Canu. Recovering sparse signals with a certain family of nonconvex penalties and dc programming. *IEEE Transactions on Signal Processing*, 57, 12: 4686–4698, 2009.

D. Giamouridis and S. Paterlini. Regular(ized) hedge funds clones. *Journal of Financial Research*, 33(3):223–247, 2010.

J. Havrda and F. Charvát. Quantification method of classification processes: Concept of structural entropy. *Kibernetika*, 3:30–35, 1967.

R. Jagannathan and T. Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, LVIII:1651–1683, 2003.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004b.

J.A.F. Machado. Robust model selection and m-estimation. *Econometric Theory*, 9:478–478, 1993.

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

Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

R.C. Merton. On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, 8(4):323361, 1980.

R.O. Michaud. The markowitz optimization enigma: Is optimized optimal? *Financial Analysts Journal*, 45(1):3145, 1989.

E. Ronchetti. Robustness aspects of model choice. *Statistica Sinica*, 7:327–338, 1997.

P. Shi and C.L. Tsai. A note on the unification of the akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):551–558, 1998.

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

C. Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal od Statistical Physics*, 52:479–487, 1988.

B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 27:349–363, 2005.

B. Vaz de Melo and R. P. Camara. Robust multivariate modeling in finance. *International Journal of Managerial Finance*, 4:12–23, 2005.

R.E. Welsch and X. Zhou. Application of robust statistics to asset allocation models. *RevStat*, 5: 97–114, 2007.

J. Weston, A. Elisseeff, and B. Schlkopf. Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3:14391461, 2003.

Y. Zhang, R. Li, and C.L. Tsai. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105:312–323, 2010.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 476:1418–1429, 2006.

# A Sources of bias and corrections

Suppose $\boldsymbol{\beta}_0$ and $\sigma_0$ are optimal parameter values such that the standardized variable $Z_0 = (\boldsymbol{X}^T\boldsymbol{\beta}_0 - \mu^*)/\sigma_0$ follows exactly the assumed model $f(z)$ (normal or t-Student model). The solution of equations (4) are generally biased for $\boldsymbol{\beta}_0$ and $\sigma_0$ for large $n$ ($n \to \infty$), which is important to discuss. The following heuristic derivation illustrates how the bias can be controlled by appropriate selection of $q$ and $\lambda$. Let $\lambda_n$ be a sequence such that $\lambda_n \to \infty$ as $n \to \infty$. Appropriate regularity conditions on $f$ and $\pi$ yield the first-order Taylor expansion $\mathbf{0} = n^{-1}\widehat{\boldsymbol{\Psi}}(\boldsymbol{\beta}_0) + n^{-1}\nabla\widehat{\boldsymbol{\Psi}}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}}_{q,\lambda_n} - \boldsymbol{\beta}_0) + o_p(1)$. Therefore, for large $n$, the Law of Large Numbers gives the following approximation:

$$\text{Bias}(\widehat{\boldsymbol{\beta}}_{q,\lambda_n}) = \widehat{\boldsymbol{\beta}}_{q,\lambda_n} - \boldsymbol{\beta}_0 \approx \mathbf{A}_n^{-1} E\widehat{\boldsymbol{\Psi}}(\boldsymbol{\beta}_0)/n = \mathbf{A}_n^{-1}\left\{ E\left[w_q(\boldsymbol{X}, \boldsymbol{\beta}_0, \sigma_0)\boldsymbol{u}(\boldsymbol{X}, \boldsymbol{\beta}_0, \sigma_0)\right] + \frac{1}{n}\mathbf{p}'_{\lambda_n}(\boldsymbol{\beta}) \right\}, \tag{24}$$

where $\mathbf{A}_n = E\left[\nabla w_q(\boldsymbol{X}, \boldsymbol{\beta}_0, \sigma_0)\boldsymbol{u}(\boldsymbol{X}\boldsymbol{\beta}_0, \sigma_0)\right] + n^{-1}\nabla\mathbf{p}'_{\lambda_n}(\boldsymbol{\beta})$ is typically a positive definite matrix. By looking at (24), we can distinguish between two sources of bias: the first source of bias is due to $w_q(\cdot, \boldsymbol{\beta}, \sigma)$ and arises in location-scale models, but not in pure location models. The second source of bias arises from the term $\mathbf{p}'_\lambda(\boldsymbol{\beta})$ implied by the penalty term and affects the coefficients.

For the location-scale models, $E\left[w_q(\boldsymbol{X}, \boldsymbol{\beta}_0, \sigma_0)\boldsymbol{u}(\boldsymbol{X}, \boldsymbol{\beta}_0, \sigma_0)\right] \neq 0$ and a simple transformation is typically needed to re-center the estimates. For a similar estimator without the penalty term, Ferrari and La Vecchia (2012) show how to correct for such a bias; following their Proposition 1, we rescale the parameter estimate by solving $f((\mathbf{x}_i^T\boldsymbol{\beta} - \mu^*)/\sigma) \propto f((\mathbf{x}_i^T\widehat{\boldsymbol{\beta}} - \mu^*)/\widehat{\sigma})^q$, in $\boldsymbol{\beta}$ and $\sigma$ where $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})$ are the solution to (4). For example, if $f$ is the normal pdf, then we take $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ and $\sigma = \widehat{\sigma}/\sqrt{q}$ as final estimates. This transformation implies that the first summand in (4) is equal to zero when $n$ is large.

The bias arising from the penalty term $\mathbf{p}'_{\lambda_n}(\boldsymbol{\beta})$ is generally necessary to prevent overfitting in over-parametrized problems. However, it is easy to see that the bias is generally small when the tuning parameter $\lambda$ is sufficiently large. If $\pi(\beta; \lambda)$ is the Laplace density function $\pi(\beta; \lambda) =$

$2^{-1}\lambda \exp\{-\lambda|\beta|\}$, we have

$$n^{-1}\left\{\mathbf{p}_{\lambda_n}(\boldsymbol{\beta})\right\}_j = n^{-1}\{v_q(\beta_j,\lambda_n)\}s(\beta_j,\lambda_n)\} = \left(\frac{\lambda_n^{2-q}}{n2^{1-q}}\right) e^{-\lambda_n|\beta_{0,j}|(1-q)}\text{sign}(\beta_{0,j}), \tag{25}$$

where $\text{sign}(\beta_{0,j}) = 1$ if $\beta_{0,j} > 1$, $\text{sign}(\beta_{0,j}) = 0$ if $\beta_{0,j} = 0$, and $\text{sign}(\beta_{0,j}) = -1$ if $\beta_{0,j} < 0$ . The above expression shows that $\|\text{Bias}(\widehat{\boldsymbol{\beta}}_{q,\lambda_n})\|$ is close to zero if the tuning constant $\lambda_n$ is sufficiently large and $q < 1$, since the term $e^{-\lambda_n|\beta_{0,j}|(1-q)}$ dominates the other terms in (25). Finally, Eq. (25) helps illustrate another important point: when $q < 1$ and $\min_j |\beta_{0,j}|$ is large, $\|\text{Bias}(\widehat{\boldsymbol{\beta}}_{q,\lambda_n})\|$ tends to be small. The same behavior, however, is not observed for the Lasso, since $q = 1$ implies the presence of the biasing term $\lambda_n\text{sign}(\beta_{0,j})/n$ in (25), which does not vanish as $n \to \infty$ (unless $\lambda_n$ goes to zero at an appropriate rate, which would then imply absence of regularization).