Journal Pre-proof

Sparse Portfolio Selection via the sorted ‘$_1$ - Norm

Philipp J. Kremer, Sangkyun Lee, Ma Igorzata Bogdan,
Sandra Paterlini

PII: S0378-4266(19)30261-4
DOI: https://doi.org/10.1016/j.jbankfin.2019.105687
Reference: JBF 105687

To appear in: *Journal of Banking and Finance*

Received date: 26 September 2018
Accepted date: 1 November 2019

Please cite this article as: Philipp J. Kremer, Sangkyun Lee, Ma Igorzata Bogdan, Sandra Paterlini, Sparse Portfolio Selection via the sorted ‘$_1$ - Norm, *Journal of Banking and Finance* (2019), doi: https://doi.org/10.1016/j.jbankfin.2019.105687

Sparse Portfolio Selection via the sorted ℓ_1 - Norm

Philipp J. Kremer[a,b,*], Sangkyun Lee[c,*], Małgorzata Bogdan[d], Sandra Paterlini[e]

[a]Prime Capital AG, Frankfurt am Main, Germany
[b]EBS Universität für Wirtschaft und Recht, Wiesbaden, Germany
[c]Hanyang University ERICA, Ansan, Republic of Korea
[d]University of Wroclaw, Wrocław, Poland
[e]University of Trento, Italy

Abstract

We introduce a financial portfolio optimization framework that allows to automatically select the relevant assets and estimate their weights by relying on a sorted ℓ_1-Norm penalization, henceforth SLOPE. To solve the optimization problem, we develop a new efficient algorithm, based on the Alternating Direction Method of Multipliers. SLOPE is able to group constituents with similar correlation properties, and with the same underlying risk factor exposures. Depending on the choice of the penalty sequence, our approach can span the entire set of optimal portfolios on the risk-diversification frontier, from minimum variance to the equally weighted. Our empirical analysis shows that SLOPE yields optimal portfolios with good out-of-sample risk and return performance properties, by reducing the overall turnover, through more stable asset weight estimates. Moreover, using the automatic grouping property of SLOPE, new portfolio strategies, such as sparse equally weighted portfolios, can be developed to exploit the data-driven detected similarities across

*Corresponding authors
Email addresses: philipp.kremer@ebs.edu (Philipp J. Kremer), sangkyun@hanyang.ac.kr (Sangkyun Lee)

[1]The views expressed in this article are those of the authors, and do not necessarily reflect the views of Prime Capital AG.

assets.

Keywords: Portfolio Management, Markowitz Model, Sorted ℓ_1-Norm Regularization, Alternating Direction Method of Multipliers

1. Introduction

The development of successful asset allocation strategies requires the construction of portfolios that perform well out-of-sample, provide diversification benefits, and are cheap to maintain and monitor. The problem is then one of statistical model selection and estimation, i.e. the identification of the assets in which to invest and the determination of the optimal weight for each asset.[2] In 1952, Harry Markowitz laid the foundation for the modern portfolio theory by introducing the mean-variance optimization framework. Assuming that asset returns are normally distributed, such model requires only two input estimates: the vector of expected returns and the expected covariance matrix of the assets. Solving the quadratic optimization problem, by minimizing the portfolio expected risk, for a given level of expected return, the investor can then find the optimal portfolio allocation. Although Markowitz's model has been widely criticized, it is the backbone of the vast majority of portfolio optimization frameworks and is still largely used in practice, especially in fintech companies as part of their robo-advisory (see e.g. Kolm et al. (2014)).

One of the major shortcomings of the mean-variance approach is the fact that opti-

[2]Another stream of literature investigates the utilization of norm penalties in portfolio selection from a behavioral perspective, in which the investor tries to model a simplified version of a complex investment processes. In such context, sparsity allows to simplify the model at hand by focusing the attention on the relevant variables and thereby taking into account the mental cost of processing data. Using the LASSO or the SLOPE penalty still results in tractable models, not NP-Hard ones, which allow a natural way to model investors preference for simpler representation of the world, in which many features are eliminated and sparsity can model dynamic attention to features of the environment. For the behavioral perspective, we refer the interested reader to Gabaix (2014) and references therein.

mized weights are highly sensitive to estimation errors and to the presence of multicollinearity in the inputs. In particular, it is acknowledged that estimating the expected returns is more challenging, than just focusing on risk minimization and thereby looking for the portfolios with minimum risk, i.e. the so-called global minimum variance portfolios (GMV) (Merton, 1980; Chopra and Ziemba, 1993; Jagannathan and Ma, 2003). But even in the GMV set-up, the sample covariance matrix might exhibit estimation error that can easily accumulate, especially when dealing with a large number of assets (Michaud, 1989; Ledoit and Wolff, 2003; DeMiguel and Nogales, 2009; Fan et al., 2012). Furthermore, multicollinearity and extreme observations often leads to undesirable and unrealistic extreme long and short positions, which can hardly be implemented in practice, due to regulatory and short selling constraints (Shefrin and Statman, 2000; DeMiguel et al., 2009b; Boyle et al., 2012; Roncalli, 2013). An ideal portfolio then has: a) conservative asset weights, which are stable in time, to avoid high turnover and transaction costs, and b) still promotes the right amount of diversification, while being able to control the total amount of shorting.

A natural approach to solve this problem is to extend the Markowitz optimization framework, by using a penalty function on the weight vector, typically given by the norm, and whose intensity is controlled by a tuning parameter λ. Probably, the most recent successful approach, using convex penalty functions, is the Least Absolute Shrinkage and Selection Operator (LASSO) introduced by Tibshirani (1996). The LASSO framework typically relies on adding to the Markowitz formulation a penalty proportional to the ℓ_1-Norm[3] on the asset weight vector (Brodie et al., 2009;

[3]Let $\boldsymbol{w} = [w_1, w_2, ..., w_k]'$ be the portfolio weight vector, then the ℓ_q-Norm is defined as: $||\boldsymbol{w}||_q = (\sum_{i=1}^{k} |w_i|^q)^{\frac{1}{q}}$, with $0 < q < \infty$. If $q = 1$, then $\ell_1 = \sum_{i=1}^{k} |w_i|$ (LASSO), while for $q = 2$ we have $||\boldsymbol{w}||_2 = (\sum_{i=1}^{k} w_i^2)^{1/2}$ (RIDGE). Note that ℓ_q with $0 < q < 1$ is not a norm but a quasi norm.

DeMiguel et al., 2009a; Carrasco and Noumon, 2012; Fan et al., 2012). DeMiguel et al. (2009a) provide a general framework that nests regularized portfolio strategies based on the ℓ_1-Norm with the approaches introduced by Ledoit and Wolff (2003) and Jagannathan and Ma (2003). Furthermore, the authors advocate their superior performance in an out-of-sample setting. Brodie et al. (2009) and Fan et al. (2012) show that the LASSO (a) results in constraining the gross exposures, (b) can be used to implicitly account for transaction costs, and (c) sets an upper bound on the portfolio risk depending just on the maximum estimation error of the covariance matrix. Moreover, the shrinkage covariance estimation of Jagannathan and Ma (2003), obtained by adding a no-short sale constraint (the so-called GMV long-only (GMV-LO)), can be considered a special case of the LASSO.

Despite its appealing properties, the LASSO has reported shortcomings of (a) large biased coefficient values (Gasso et al., 2010; Fastrich et al., 2015), of (b) reduced recovery of sparse signals when applied to highly dependent data, like crisis periods (Giuzio and Paterlini, 2016), and of (c) randomly selecting among equally correlated coefficients (Bondell and Reich, 2008). Moreover, it is ineffective in the presence of no short selling (i.e. $w_i \geq 0$) and an imposed budget constraint (i.e., $\sum_{i=1}^{k} w_i = 1$), as the ℓ_1-Norm is then just equal to 1.

To overcome these limitations, we extend the literature on convex regularization methods in various ways: First, we introduce the *Sorted ℓ_1 Penalized Estimator* (SLOPE), as a new penalty function within the mean-variance portfolio optimization framework. The SLOPE penalty takes the form of a sorted ℓ_1 - Norm, in which each asset weight is penalized individually using a vector of tuning parameters, $\boldsymbol{\lambda}_{SLOPE} = (\lambda_1, \lambda_2, \ldots, \lambda_k)$, with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k \geq 0$ and whereas $\boldsymbol{\lambda}_{SLOPE}$ is decreasing, attributing the largest weight to the highest regularization parameter, such that SLOPE penalizes the weights according to their rank magnitude. This leads to an octagonal shape of the penalty in a 2-dimensional setting (see Figure 1) that combines the two favorable properties of the ℓ_∞-Norm and the ℓ_1-Norm[4], which is the grouping of variables (i.e. some asset weights are assigned the same coefficient value), and the singularity at the origin, respectively. Our work shows that, opposed to the LASSO, in portfolio optimization and together with an added budget constraint (i.e. $\sum_{i=1}^{k} w_i = 1$), SLOPE continues to shrink the active weights, even when short sales are restricted (i.e. $w_i \geq 0, \ \forall \ i = 1, ..., k$). Consequently, it spans the diversification frontier from the GMV-LO up to the equally weighted (EW) portfolio, as $\boldsymbol{\lambda}_{SLOPE}$ goes to infinity. Together with the feature of grouping equally correlated assets, the penalty provides an increased flexibility for the investor in creating individual trading strategies, as opposed to state-of-the-art shrinkage methods.

Second, we introduce a new optimization algorithm to solve the mean-variance portfolio problem with the sorted ℓ_1 regularization and linear constraints on the asset weights. The algorithm uses the ideas of variable splitting and the Alternating Direction Method of Multipliers (ADMM) framework (Powell, 1969; Hestenes, 1969; Boyd et al., 2011). Using a mathematically equivalent reformulation of the original problem, the algorithm can use existing implementations of proximal operators (Parikh and Boyd, 2014), associated with the ℓ_1, the sorted ℓ_1, and even other regularizers. Furthermore, Appendix C shows that the ADMM provides a more efficient alternative for solving the LASSO optimization problem, than the state-of-art Cyclic Coordinate Descend (CyCoDe) algorithm.

Third, we are, to our knowledge, the first to investigate the properties of SLOPE under a realistic factor model, which assumes that all assets can be represented as linear combination of a small number of hidden risk factors, as e.g. in Fan et al. (2008).

[4]Given a weight vector $\boldsymbol{w}$ with k elements, the $\ell_\infty = ||\boldsymbol{w}||_\infty = max(w_1, ..., w_k)$.

In the set-up of classical multiple regression, in which the explanatory variables are assumed independent, Bogdan et al. (2013, 2015) and Su and Candès (2016) provide extensive evidence of SLOPE's superior model selection and estimation properties. Further evidence for these properties are provided by the results of Bellec et al. (2016a) and Bellec et al. (2016b), which show that contrary to LASSO, SLOPE is asymptotically optimal for the general class of design matrices satisfying the modified Restricted Eigenvalue condition.

Moreover, Bondell and Reich (2008) and Figueiredo and Nowak (2014) investigate the properties of SLOPE and its predecessor OSCAR (Octagonal Shrinkage and Selection Operator, Bondell and Reich (2008)) in the situation, when regressors are strongly correlated. Bondell and Reich (2008) apply OSCAR to agricultural data, showing that the method successfully forms predictive clusters, which can then be analyzed according to their individual characteristics. Figueiredo and Nowak (2014) illustrate the "clustering" properties of the *ordered weighted* ℓ_1 *- Norm* (OWL) in the linear regression framework with strongly correlated predictors, providing further simulation and theoretical results. However, none of these works addresses the interesting situation, in which the correlation structure results from the dependency of the explanatory variables on a few hidden factors and on financial real-world data. Recently, Xing et al. (2014) applied the OSCAR to the mean-variance portfolio optimization, together with a linear combination of the ℓ_1- and the ℓ_∞-Norms. They advocate the method for its ability to identify portfolios that attain higher Sharpe Ratios and lower turnovers compared to those resulting from traditional approaches like the GMV and the GMV-LO portfolios. However, they do not point out the clumping property of the OSCAR. With SLOPE, we consider a generalized framework that nests the GMV, the GMV-LO, the LASSO, the ℓ_∞ - Norm and the approach of Xing et al. (2014).

In this paper, we analyze the properties of SLOPE, with both simulated and real world data. The simulations show that SLOPE reduces the estimation errors in the portfolio weights and groups assets depending on the same risk factors together. This grouping behavior then allows the investor to select individual constituents from the clusters, for example based on her preferences and asset-specific properties, enabling her to develop new investment strategies such as SLOPE-EW, which we introduce in Section 4.1.

For the real world data analysis, we use monthly returns of the 10- and 30-Industry portfolios (Ind), as well as the 100 Fama French (FF) portfolios formed on Size and Book-to-Market, covering the period from 1970 to 2017. Furthermore, we consider daily returns of the S&P 500 (SP500) from 2004 to 2016. Our results show that the risk of the SLOPE portfolio is comparable to or smaller than the risk of the LASSO portfolio. Also, we observe that SLOPE outperforms the LASSO, yielding better risk- and weight diversification measures. In fact, the sorted ℓ_1- Norm is able to span the entire risk-diversification frontier, starting from the GMV, via the GMV-LO up to the EW. The investor can then select the portfolio with the risk-diversification trade-off that best fits her preferences.

The above mentioned characteristics establish SLOPE as a new attractive portfolio construction alternative, capable of controlling short sales and identifying groups of assets. It thereby offers the possibility to implement individual views, which goes beyond the standard statistical shrinkage or regularization approaches.

The paper is structured as follows: Section 2 introduces our methodology and discusses the properties of SLOPE. Section 3 analyses the behavior of SLOPE in simulated environments, while Section 4 focuses on the empirical results. Section 5 concludes.

2. Sparse Portfolio Selection via the Sorted ℓ_1-Norm

Given k jointly normally distributed asset returns $R_1, \ldots, R_k$, with expected value vector $\boldsymbol{\mu} = [\mu_1, ..., \mu_k]'$ and covariance matrix $\boldsymbol{\Sigma}$, a generalization to the Markowitz (1952) portfolio selection problem can be stated as the following optimization:

$$\min_{\boldsymbol{w} \in \mathbb{R}^k} \frac{\phi}{2} \boldsymbol{w}' \boldsymbol{\Sigma} \boldsymbol{w} - \boldsymbol{\mu}' \boldsymbol{w}, \text{ subject to } \sum_{i=1}^{k} w_i = 1 \tag{1}$$

where $\sigma_p^2 = \boldsymbol{w}' \boldsymbol{\Sigma} \boldsymbol{w}$ is the portfolio risk, $\boldsymbol{\mu}' \boldsymbol{w}$ is the portfolio return and $\phi > 0$ is the coefficient of relative risk aversion (Markowitz, 1952; Fan et al., 2012; Li, 2015). Despite the advantage of being a quadratic optimization problem, the generalized Markowitz model is often criticized, as it leads to extreme and unstable optimal portfolio weights.

One approach to circumvent instability and extreme estimates is to modify the optimization problem (1), by adding a penalty function, $\rho_\lambda(\boldsymbol{w})$.[5] In typical applications, the penalty is a non-decreasing function of $\boldsymbol{w}$ and leads to shrinking an estimate of the weight vector towards zero. The shrinkage stabilizes the weight estimates and, in case when $\rho_\lambda(\boldsymbol{w})$ has a singularity at zero, promotes the sparsity by shrinking some coordinates of the estimated vector to 0. An additional parameter λ controls the impact of the penalty and thereby the amount of shrinkage applied to the weights vector and the level of sparsity. Additionally, the penalty function allows to take into account a prior knowledge of an investor, who can assign smaller penalty terms

[5]Another stream of literature focusses on directly shrinking the moments of the distribution, as opposed to adding a norm penalty on the weight vector to the model in (1). However, as have been shown by DeMiguel et al. (2009a), adding a norm constraint on the weight vector to the portfolio optimization is equal to shrinking the extreme estimates in the covariance matrix. For elaborations on directly improving the inputs to the Markowitz optimization, the interested reader is referred to Ledoit and Wolff (2003, 2004a), Jorion (1986) and references therein.

to selected important assets. The optimization problem can be stated as:

$$\min_{\boldsymbol{w}\in\mathbb{R}^k} \frac{\phi}{2}\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} - \boldsymbol{\mu}'\boldsymbol{w} + \rho_\lambda(\boldsymbol{w}) \text{ s.t. } \sum_{i=1}^{k} w_i = 1 \tag{2}$$

The simplest approach is the LASSO, which considers as a penalty function the ℓ_1-Norm of the asset weights vector ($\rho_\lambda(\boldsymbol{w}) = \lambda \times \sum_{i=1}^{k} |w_i|$, with λ being a scalar). The resulting optimization problem is still convex, while promoting model selection and estimation in a single step. From a financial perspective, LASSO is interpreted as a gross exposure constraint (i.e. a constraint on the total amount of shorting) or a way to account for transaction costs (Brodie et al., 2009). However, it is not effective in the presence of both a budget ($\sum_{i=1}^{k} w_i = 1$), and a no-short selling (i.e., $w_i \geq 0$) constraint, as the ℓ_1-norm is then simply equal to 1.

Following, we propose a more general approach that within a single optimization algorithm allows us to encompass the original LASSO, the OSCAR of Bondell and Reich (2008), and the combination of ℓ_1 and ℓ_∞ penalties, as proposed in Xing et al. (2014).

In fact, we penalize the weights vector by considering as $\rho_\lambda(\boldsymbol{w})$ the sorted ℓ_1-Norm, defined as:

$$\begin{gathered} \rho_\lambda(\boldsymbol{w}) := \sum_{i=1}^{k} \lambda_i |w|_{(i)} = \lambda_1 |w|_{(1)} + \lambda_2 |w|_{(2)} + ... + \lambda_k |w|_{(k)} \\ \text{s.t. } \lambda_1 \geq \lambda_2 \geq \dots \lambda_k \geq 0 \text{ and } |w|_{(1)} \geq |w|_{(2)} \geq \dots |w|_{(k)} \ , \end{gathered} \tag{3}$$

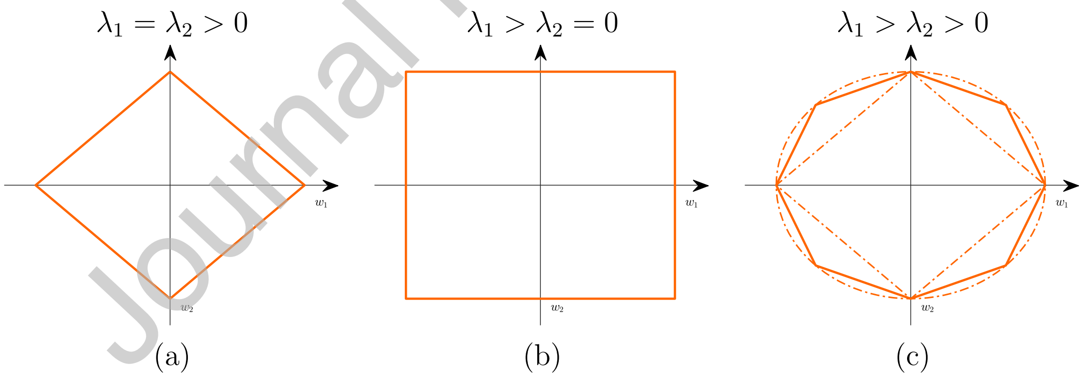
where $|\boldsymbol{w}|_{(i)}$ denotes the ith largest element in absolute value of the vector $\boldsymbol{w}$. The sorted ℓ_1-Norm was originally introduced in Bogdan et al. (2013, 2015) to construct the Sorted ℓ_1 Penalized Estimator for the selection of explanatory variables in the multiple regression model. It was also developed independently by Zeng and Figueiredo (2014) as Ordered Weighted ℓ_1 Norm (OWL). To our knowledge, this is the first work in financial portfolio selection that applies SLOPE and discusses its grouping properties, while also introducing a new optimization algorithm.

2.1. Geometric Interpretation

Compared to most of the other regularization methods, SLOPE does not rely on a single tuning parameter λ, but rather on a non-increasing sequence $\boldsymbol{\lambda}_{SLOPE} = (\lambda_1, \lambda_2, \ldots, \lambda_k)$, with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq 0$. This sequence is aligned to the sorted weight vector, such that the largest absolute weight is penalized with the largest tuning parameter. Consequently, the sequence of λ parameters gives a natural interpretation of importance to the asset weights, besides providing full flexibility in recapturing the profiles of the ℓ_1- and ℓ_∞- Norms, as well as of their linear combinations. Figure 1 shows a simple set-up with two assets and the respective shapes of spheres (i.e. the set points for which $\rho_\lambda(\boldsymbol{w}) = c$) that we obtain, depending on how the sequence $\boldsymbol{\lambda}_{SLOPE} = (\lambda_1, \lambda_2)$ is chosen. As shown in Panel (a), if $\lambda_1 = \lambda_2 > 0$ the SLOPE sphere coincides with the well studied diamond shape of the LASSO penalty. Through its singularity at the origin, the LASSO promotes sparse solutions that set one of the two assets' weights exactly equal to zero. On the other hand, choosing $\lambda_2 = 0$ and $\lambda_1 > 0$, yields the regularization term of the ℓ_∞-Norm. The respective shape, as shown in Panel (b), takes the form of a square and promotes the grouping of variables, i.e. it encourages solutions under which both asset weights are assigned exactly the same value.

Given these two extreme cases, Panel (c) of Figure 1 shows the octagonal shape of SLOPE, obtained by using a decreasing sequence of lambda values, with $\lambda_1 > \lambda_2 > 0$. By choosing different tuning parameter sequences, the penalty allows to approximate a variety of norms between the ℓ_1 and the ℓ_∞, combining the properties of the Lasso and the ℓ_∞ penalties and due to its singularity, is either able to set some weights exactly equal to zero, and/or to assign the same value to some of the other weights. Furthermore, it approximates the shape of the already well studied RIDGE penalty, which corresponds to a circle in the 2-dimensional set-up, and is even able to reach one of RIDGE's special solutions, i.e. the equally weighted portfolio, which is obtained, when RIDGE's penalty parameter approaches infinity. Although RIDGE is still convex, the shape of the penalty does not promote sparsity among the coefficients, leading to undesirable portfolios with a large number of active positions (Carrasco and Noumon, 2012; DeMiguel et al., 2009a). Thus, the choice of the lambda sequence for SLOPE provides the investor with the flexibility to choose any of these shapes of the unit sphere and of the corresponding mode of shrinking the dimension of the weight vector.

Figure 1: Geometric Representation of Penalty Functions

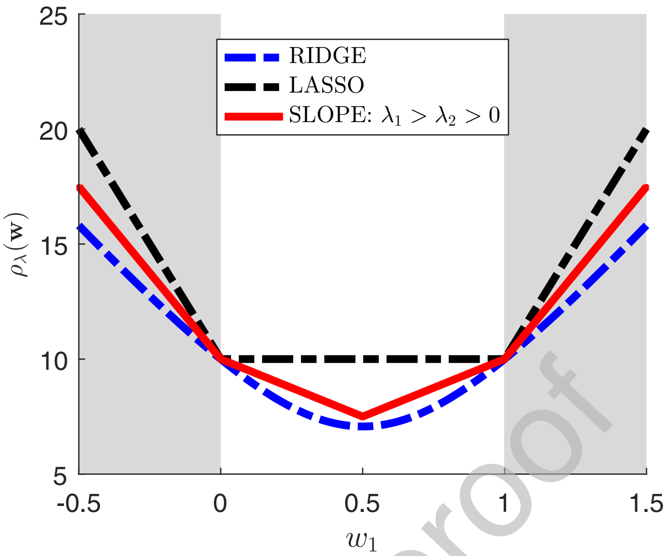


For two asset weights $\boldsymbol{w} = [w_1\ w_2]'$, the figure shows the unit spheres for different SLOPE sequences: (a) the LASSO ℓ_1 sphere, when $\lambda_1 = \lambda_2 > 0$, (b) the ℓ_∞ sphere, when $\lambda_1 > \lambda_2 = 0$ and (c) the SLOPE sphere, when $\lambda_1 > \lambda_2 > 0$. The dashed lines in (c) represent the diamond shape of the LASSO and the RIDGE ℓ_2-balls, respectively.

In portfolio optimization, a budget constraint that requires the weights of the portfolio to sum to one, is imposed. Consequently, we discuss how the penalties behave in the presence of such an additional constraint. Figure 2 plots the SLOPE penalty, together with the LASSO and the RIDGE penalty for a universe of two assets and under the condition that $w_1 + w_2 = 1$. Furthermore, we consider the penalty functions in the presence of short sales (gray area) and no short sales (white area). In Figure 2, we can see that the LASSO (shown in black) is only effective when short sales are permitted, while the presence of the budget constraint makes it ineffective in the long-only area. In contrast, the RIDGE attains its minimum for an equally weighted portfolio, and when short sales are restricted. Similarly, the SLOPE penalty (shown in red) also reaches its minimum at the equally weighted solution (i.e., $w_1 = w_2 = 0.5$). Still, to control for monitoring and transaction costs of financial assets, we prefer SLOPE over the RIDGE estimator, because it can promote sparsity by exploiting the singularities.

Figure 3 plots the contours of the objective function, together with those of the SLOPE spheres for the two asset case, and when we do not impose a budget constraint (i.e. $\sum_{i=1}^{k} w_i = 1$), as well as considering orthogonal and correlated designs. As noted before, if only $\lambda_2 > 0$, SLOPE always has singular points when one of the asset weights is equal to zero, thereby promoting sparsity. When $\lambda_1 > \lambda_2 > 0$, that is, the sequence is monotonically decreasing, then SLOPE has additional singular points, which correspond to $|w_1| = |w_2|$. This is an appealing property in the presence of correlated data. Specifically, as Panel (b) shows, strong correlation between assets lead to the same weights and thereby grouping. This is consistent with portfolio theory, as it is known that, if assets have all the same correlation coefficients, as well as identical means and variances, the EW is the unique optimal portfolio. SLOPE then allows us to automatically group assets with similar correlation.

Figure 2: Penalty Functions in a Two Asset Universe with Budget Constraint

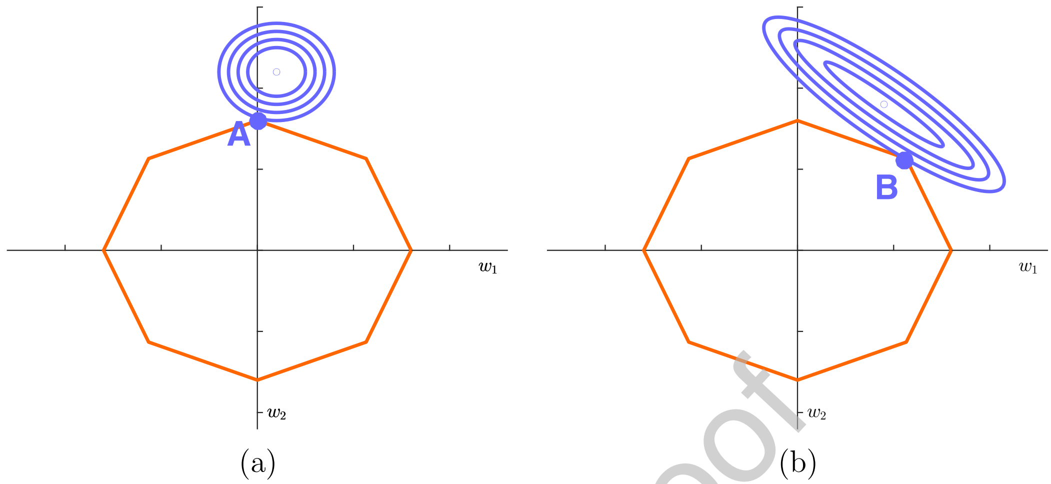


The figure plots the SLOPE coefficient alongside the LASSO ($\ell_1 - Norm$) and the RIDGE penalty ($\ell_2 - Norm$), for a two asset case and under the condition that $w_1 + w_2 = 1$.

For our simulation analysis and empirical investigations, SLOPE requires us to define a specific form of the sequence of $\boldsymbol{\lambda}_{SLOPE} = (\lambda_1, \lambda_2, \ldots, \lambda_k)$. For that, we use the decreasing sequence of quantiles of the standard normal distribution, as in Bogdan et al. (2013) and Bogdan et al. (2015), with $\lambda_i = \alpha\Phi^{-1}(1 - q_i)$, $\forall i = 1, ..., k$, where Φ is the cumulative distribution function of the standard normal distribution and $q_i = i \times \theta/2k$, and in which $\theta = 0.01$, regulates how fast the sequence of lambda parameters is decreasing. Bogdan et al. (2013) and Bogdan et al. (2015) have shown that in orthogonal design this sequence controls the False Discovery Rate in a multiple testing framework.[6]

[6]We investigated different sequences of lambda parameters, including changing values of θ, as well as a linear decreasing sequence and obtained qualitative similar results. Consequently, we

Figure 3: Sorted ℓ_1-Norm Penalty without Budget Constraint



The figure plots in Panel (a) and (b), the Sorted ℓ_1-Norm Penalty (SLOPE) in a 2-dimensional setting, considering orthogonal design and correlated design, respectively.

2.2. Optimization Algorithm

In this section, we describe our solution algorithm, which is based on equivalent reformulations of the Alternating Direction Method of Multipliers (ADMM) approach (see Appendix A for details).

choose the exponentially decreasing sequence, as proposed by Bogdan et al. (2013) and Bogdan et al. (2015). Still, future research on how to choose the sequence of lambda parameters is currently high on our agenda.

Algorithm 1 ADMM Algorithm

1: Input: Expected value vector $\mu \in \mathbb{R}^k$, covariance matrix $\Sigma \in \mathbb{R}^{k\times k}$.

2: Initialize $\boldsymbol{w}^0 \in \mathbb{R}^k$, $\boldsymbol{v}^0 \in \mathbb{R}^k$, $\alpha^{\mathbf{0}} \in \mathbb{R}^{\mathbf{k}}$, $\beta^0 \in \mathbb{R}$ and $j = 0$.

3: Given a stopping threshold value $\tau > 0$.

4: **while** $\mathcal{G}(\boldsymbol{w}^j, \boldsymbol{v}^j, \boldsymbol{\alpha}^j, \beta^j) > \tau$ **do**

5: Update $\boldsymbol{w}^j$, $\boldsymbol{v}^j$, $\boldsymbol{\alpha}^j$, and β^j as follows,

$$\begin{cases} \boldsymbol{w}^{j+1} & = \arg\min_{\boldsymbol{w}} \ \mathcal{L}_\eta(\boldsymbol{w}, \boldsymbol{v}^j; \boldsymbol{\alpha}^j, \beta^j) = (\phi\boldsymbol{\Sigma} + \eta(I + \boldsymbol{e}\boldsymbol{e}'))^{-1}(\boldsymbol{\mu} - \boldsymbol{\alpha}^j - \beta^j\boldsymbol{e} + \eta(\boldsymbol{v}^j + \boldsymbol{e})) \\ \boldsymbol{v}^{j+1} & = \arg\min_{\boldsymbol{v}} \ \mathcal{L}_\eta(\boldsymbol{w}^{j+1}, \boldsymbol{v}; \boldsymbol{\alpha}^j, \beta^j) = \mathrm{prox}_{\lambda/\eta}(\boldsymbol{w}^{j+1} + (1/\eta)\alpha^j) \\ \boldsymbol{\alpha}^{j+1} & = \boldsymbol{\alpha}^j + \eta(\boldsymbol{w}^{j+1} - \boldsymbol{v}^{j+1}) \\ \beta^{j+1} & = \beta^j + \eta(\boldsymbol{e}'\boldsymbol{w}^{j+1} - 1) \ , \end{cases} \tag{4}$$

6: $j = j + 1$

7: **end while**

In Algorithm 1, the quantity $G(\boldsymbol{w}^j, \boldsymbol{v}^j, \boldsymbol{\alpha}^j, \beta^j)$ represents the primal-dual gap which converges to the zero value when the iterates $\boldsymbol{w}^j, \boldsymbol{v}^j, \boldsymbol{\alpha}^j, \beta^j$ approaches the optimal quantities (see Appendix B for details). In the presence of the no-short selling constraint, we consider a slightly different formulation to (2) with an extra constraint that $\boldsymbol{w} \geq 0$. In this case, we can use the Algorithm 1 almost as it is, except that the w update in (4) is modified as follows,

$$\boldsymbol{w}^{j+1} = \arg\min_{\boldsymbol{w}\geq 0} \ \mathcal{L}_\eta(\boldsymbol{w}, \boldsymbol{v}^j; \boldsymbol{\alpha}^j, \beta^j) = \max\left\{(\phi\boldsymbol{\Sigma} + \eta(I + \boldsymbol{e}\boldsymbol{e}'))^{-1}(\boldsymbol{\mu} - \boldsymbol{\alpha}^j - \beta^j\boldsymbol{e} + \eta(\boldsymbol{v}^j + \boldsymbol{e})), \ 0\right\} \ , \tag{5}$$

where the minimizer is obtained by adding a simple clipping operation, since $L_\eta(\boldsymbol{w}, \boldsymbol{v}^j; \boldsymbol{\alpha}^j, \beta^j)$ is a convex function in $\boldsymbol{w}$.

Our algorithm can also be used to solve the LASSO optimization problem, which is a specific instance of SLOPE. In Appendix C, we provide a direct comparison of our algorithm to the state-of-art Cyclic Coordinate Descent (CyCoDe) for LASSO, considering a simulated constant correlation model.

Bounds on the Objective Function. To solve the mean-variance problem, as stated in (2), the investor needs to provide an estimate of the true covariance matrix of asset returns $\boldsymbol{\Sigma}$ and of the true mean $\boldsymbol{\mu}$, which are in the most simplest form given by the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ and the sample mean $\hat{\boldsymbol{\mu}}$, respectively. However, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\mu}}$ might be prone to substantial estimation errors and highly sensitive to outliers. Let us define $M(\boldsymbol{\Sigma}, \boldsymbol{\mu}) = \frac{\phi}{2}\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} - \boldsymbol{w}'\boldsymbol{\mu}$, where $\boldsymbol{w}$ is the vector of weights returned by SLOPE. Now, observe that the Sorted ℓ_1-Norm satisfies $\rho_\lambda(\boldsymbol{w}) \geq \lambda_k ||\boldsymbol{w}||_1$. Thus, as $\lambda_k > 0$, we have $||\boldsymbol{w}||_1 \leq c$, with $c = \frac{\rho_\lambda(\boldsymbol{w})}{\lambda_k}$, and simple calculations following the results of Fan et al. (2012) for LASSO, yield:

$$|M(\widehat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}}) - M(\boldsymbol{\Sigma}, \boldsymbol{\mu})| \leq \frac{\phi}{2}||\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_\infty \rho_\lambda^2(\boldsymbol{w})/\lambda_k^2 + ||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||_\infty \rho_\lambda(\boldsymbol{w})/\lambda_k \tag{6}$$

where $||\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_\infty$ and $||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||_\infty$ are the maximum component-wise estimation errors for the covariance matrix and the expected return.

This result implies that the difference between the objective functions for the estimated and true vector of parameters decreases, as we restrict the Sorted ℓ_1-Norm of the weight vector. It is also important to observe that, due to the budget constraint, a higher weight on the penalty sets an upper bound on the total amount of short sales in the portfolio, as $\rho_\lambda(\mathbf{w}) \geq \lambda_\mathbf{k}||\boldsymbol{w}||_\mathbf{1} = \lambda_\mathbf{k}(\boldsymbol{w}^+ + \boldsymbol{w}^-)$, with $\boldsymbol{w}^+ - \boldsymbol{w}^- = 1$, where $\boldsymbol{w}^+ = \sum\limits_{w_i \geq 0} w_i$ and $\boldsymbol{w}^- = \sum\limits_{w_i < 0} w_i$ are the gross amount of long and short positions, respectively.

3. Simulation Analysis

This section investigates the effect of SLOPE on the model risk, the sparsity and the grouping properties, by considering simulated data. The purpose of the simulations is to investigate the properties of our new penalty, when the data generating mechanism is completely known, so that the results can be compared to the so-called *oracle* solution. Furthermore, and as is it widely acknowledged that the estimation errors in $\boldsymbol{\mu}$ are much larger than in $\boldsymbol{\Sigma}$, we focus on a risk minimization framework. Assuming $\boldsymbol{\Sigma}$ to be known, we can use the alternative formulation of SLOPE and define:

$$\boldsymbol{w}_{opt} = \underset{\boldsymbol{w}:\sum_{i=1}^{k} w_i=1,\ \rho_{\lambda(\boldsymbol{w})}\leq c}{\arg\min} \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} \quad \text{and} \quad \hat{\boldsymbol{w}}_{opt} = \underset{\boldsymbol{w}:\sum_{i=1}^{k} w_i=1,\ \rho_{\lambda(\boldsymbol{w})}\leq c}{\arg\min} \boldsymbol{w}'\hat{\boldsymbol{\Sigma}}\boldsymbol{w} \tag{7}$$

whereas $\boldsymbol{w}_{opt}$ and $\hat{\boldsymbol{w}}_{opt}$ are the theoretical optimal and the empirical optimal weights vector, respectively. We then define the *empirical* portfolio risk as $\widehat{Risk}(\hat{\boldsymbol{w}}_{opt}) = \hat{\boldsymbol{w}}'_{opt}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{w}}_{opt}$, the *actual* portfolio risk as $Risk(\hat{\boldsymbol{w}}_{opt}) = \hat{\boldsymbol{w}}'_{opt}\boldsymbol{\Sigma}\hat{\boldsymbol{w}}_{opt}$ and the *oracle* portfolio risk as $Risk(\boldsymbol{w}_{opt}) = \boldsymbol{w}'_{opt}\boldsymbol{\Sigma}\boldsymbol{w}_{opt}$, respectively. Following the proof of Theorem 1 of Fan et al. (2012), we can easily show that when $\lambda_k > 0$, the pair differences between the three measures are upper bounded by:

$$|Risk(\hat{\boldsymbol{w}}_{opt}) - Risk(\boldsymbol{w}_{opt})| \leq 2c^2||\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_\infty, \tag{8}$$

$$|Risk(\hat{\boldsymbol{w}}_{opt}) - \widehat{Risk}(\hat{\boldsymbol{w}}_{opt})| \leq c^2||\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_\infty, \tag{9}$$

$$|Risk(\boldsymbol{w}_{opt}) - \widehat{Risk}(\hat{\boldsymbol{w}}_{opt})| \leq c^2||\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_\infty \tag{10}$$

The three risk measures then allow us to extract different information: The empirical risk is the only one that is known, as it is estimated from our in-sample data. The actual risk is the one, to which the investor is truly exposed to, when using the estimated optimal weights ($\hat{\boldsymbol{w}}_{opt}$). Finally, the oracle risk is the risk the investor could only obtain, if $\boldsymbol{\Sigma}$ is known. As the SLOPE penalty becomes more binding, when $\boldsymbol{\lambda}$ "increases", the three risk measures align. In the following section, we investigate how increasing the SLOPE penalty allows to reduce the estimation error and to avoid its accumulation in the portfolio risk.

Assume that the return of an asset is represented by a linear combination of r risk factors. Furthermore, let t be the number of observations, k be the number of assets, and $\boldsymbol{F}_{t\times r} = [\boldsymbol{f}_1 \ \boldsymbol{f}_2 \ ... \ \boldsymbol{f}_r]$, where $\boldsymbol{f}_i$ is the $t \times 1$ vector of returns of the i^{th} risk factor. Moreover, let $\boldsymbol{B}_{r\times k}$ be the loading matrix for the individual risk factors. Then, the $t \times k$ matrix of asset returns from the Hidden Factor Model (i.e. $\boldsymbol{R}_{HF}$) can be represented as:

$$\boldsymbol{R}_{HF} = \boldsymbol{F} \times \boldsymbol{B} + \boldsymbol{\epsilon} \tag{11}$$

where $\boldsymbol{\epsilon}$ is a $t \times k$ matrix of error terms.

For our first illustration of the performance of SLOPE, we generate the data using the following simplified scenario:

- $t = 50$, $k = 12$, $r = 3$,
- the risk factors $f_1, \ \ldots, \ f_3$ are independent from the multivariate standard normal $N(0, I_{r\times r})$ distribution, with $I_{r\times r}$ being an identity matrix,
- the vectors of error terms ϵ_i, $i = 1, \ldots, k$, for each asset are independent from each other, as well as from each of the risk factors and come from the multivariate normal distribution $N(0, 0.05 \times I_{r\times r})$
- the loadings matrix $B_{r\times k}$ is made of exactly four copies of each of the following

columns: $[0.77\ 0.64\ \ 0]'$, $[0.9\ \ 0\ \ -0.42]'$ and $[0\ \ 0.31\ 0.64]'$.

In this way, we generate three different groups that have the same exposure to the same two risk factors and are thus strongly correlated.[7]

Finally, given (11), the covariance matrix of the assets $\boldsymbol{\Sigma}_{HF}$ is given by:

$$\boldsymbol{\Sigma}_{HF} = \boldsymbol{B}'\boldsymbol{B} + 0.05 \times I_{k\times k}. \tag{12}$$

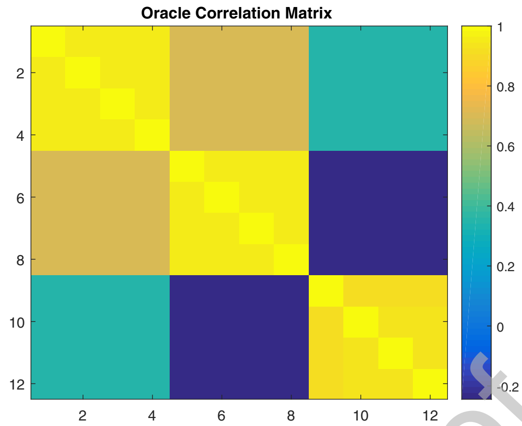
After generating our $t\times k$ matrix $\boldsymbol{R}_{HF}$ of asset returns from (11), we can then estimate $\boldsymbol{\Sigma}_{HF}$, using the sample covariance estimate $\hat{\boldsymbol{\Sigma}}_{HF}$.[8] Figure 4 shows the correlation matrix resulting from (12), illustrating that our simulation scenario explicitly models a block correlation environment, with strong correlation among each of the assets having the same underlying risk factor exposures, and low to negative correlations between the assets with a different underlying factor structure. Following, we investigate the behavior of SLOPE and the LASSO with respect to portfolio risk, and when we increase the value of the tuning parameter.

Unlike the LASSO, SLOPE requires us to define a decreasing sequence of $\boldsymbol{\lambda}_{SLOPE} = (\lambda_1, \lambda_2, \ldots, \lambda_k)$. As pointed out in Section 2.1, we use the decreasing sequence of quantiles of the standard normal distribution, as in Bogdan et al. (2013) and Bogdan et al. (2015), with $\lambda_i = \alpha\Phi^{-1}(1 - q_i)$, $\forall i = 1, ..., k$, where Φ is the cumulative distribution function of the standard normal distribution and $q_i = i \times \theta/2k$, and in which $\theta = 0.01$, regulates how fast the sequence of lambda parameters is decreasing.

[7]For the robustness of our results, we tested SLOPE in various set-ups, with qualitatively similar results. Due to space limitations, we report only the most interesting one. The results of the remaining simulations are available from the authors upon request.

[8]We explicitly restrict us to use the sample covariance estimate, as opposed to an alternative shrinkage or factor based estimate, to investigate SLOPE's ability to account for estimation errors in the optimization.

Figure 4: Hidden Factors Correlation Matrix

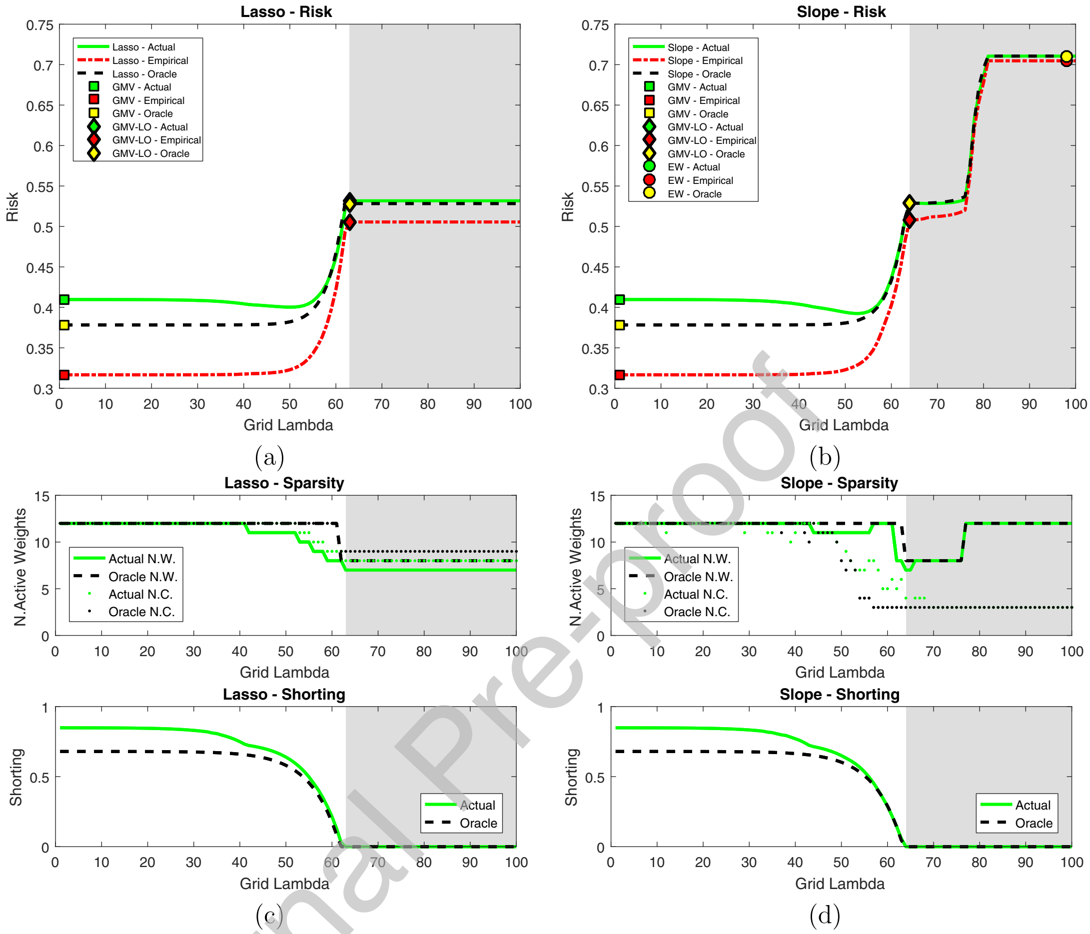


The figure plots the correlation matrix, based on the modeled Hidden Factor Structure, considering a universe of 12 assets, of which 4 are always exposed to exactly two out of the three risk factors in the market.

In our simulations, we vary the scaling parameter α so that the first element of the sequence $\lambda_1 = \alpha\Phi^{-1}(1-q_1)$ is equal to a grid of 100 log-spaced values between 10^{-5} and 10^2. Note that in the case of the LASSO, we only choose one lambda parameter, which then remains constant for all assets. Throughout the paper, we always choose $\lambda_{LASSO} = \lambda_1$. This choice favors sparser solutions for the LASSO, since for the remaining $k-1$ assets its penalty is larger than that of SLOPE.

Figure 5 plots the resulting risk and weight profile for the minimum variance optimization, when we solve (2) separately with the LASSO and the SLOPE penalties for the grid of 100 lambda parameters, and considering $\mathbf{\Sigma}_{HF}$ and the sample covariance estimate $\widehat{\mathbf{\Sigma}}_{HF}$, respectively. In particular, Panels (a) and (b) show the risk profile of the LASSO and SLOPE, i.e. the actual, the oracle, and the empirical risk, together with the results of the GMV, the GMV-LO and the EW portfolios. For both, the oracle and the actual solution, Panels (c) and (d) display on top the number of active weights together with the number of groups, that is the number of distinct coeffi-

Figure 5: Hidden Factors Minimum-Variance Profile



The figure shows the Hidden Factor minimum-variance risk profile for the LASSO and the SLOPE, including in Panel (a) and (b) their actual, empirical and oracle risk profiles, together with that of the GMV, the GMV-LO and the EW solutions. Furthermore, Panel (c) and (d) display the number of active weights, together with the grouping profile (top) and the total amount of shorting (bottom). All values are computed based on a Hidden Factor Structure, with three risk factors and considering for the exponentially decreasing sequence of lambda parameters, a grid of 100 log spaced starting points for λ_1 from 10^{-5} (i.e. x-value = 1) to 10^2 (i.e. x-value = 100).

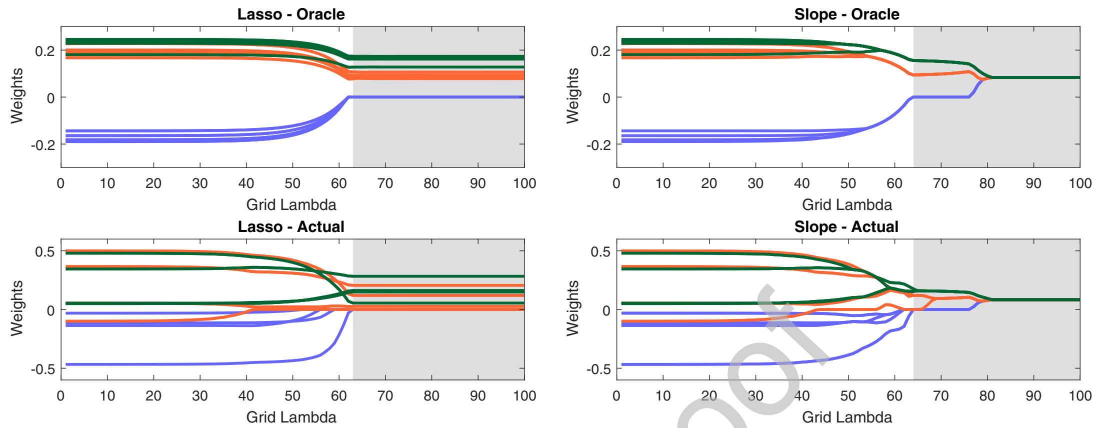
cients, while on the bottom, it shows the amount of shorting (i.e. $\boldsymbol{w}^-$). The grey surface indicates the no-short-sale-area (i.e. $w_i \geq 0 \ \forall \ i = 1, .., k$). Figure 5 shows that for a tuning parameter equal to zero, which corresponds to the GMV solution, the empirical risk is about 1.3 times lower than the actual risk (Panels (a) and (b)), with 12 active positions (Panel (c)) and slightly under 100% short sales (Panel (d)). This can be interpreted as evidence that in over-fitted models the estimation error in $\widehat{\boldsymbol{\Sigma}}_{HF}$ strongly affects the estimation of the asset weights. As here neither the LASSO nor the SLOPE penalty are binding, estimation errors can enter unhindered into the optimization. Michaud (1989) describes this phenomenon as "error maximization", in which the ill-conditioned covariance estimates are amplified through the optimization, leading to extreme long and short portfolio weights. Moving along the grid of $\boldsymbol{\lambda}$ parameters from the left to the right, Panels (c) and (d) show that the two penalties reduce the total amount of shorting in the oracle and the actual portfolio.

As we move from the GMV towards the GMV-LO, the actual, oracle, and empirical risk of the LASSO and the SLOPE align. This effect was first observed and theoretically motivated by Fan et al. (2012), showing that the portfolio risk evolves in a U-shape, in which risk first decreases before increasing again, due to the restriction of short sales. With the observations above, we extend the results of Fan et al. (2012), showing that the U-shaped behavior of the portfolio risk is not the only possible one. Especially when the dependence among the assets is positive, the tighter constraint in terms of short sales shrinks the optimization search space of feasible solutions, making it impossible to exploit the optimal diversification benefits. This leads to a higher portfolio risk when reaching the GMV-LO. The investor also reaches the maximum sparsity, that is the maximum number of coefficients equal to zero, at this point. For the LASSO, increasing the tuning parameter beyond this point does not alter the allocation any further, as the regularization penalty is constant and equal to 1.

This is different for SLOPE: in fact, Figure 6 shows the evolution of the portfolio weights for both the oracle and the actual solution, considering both the LASSO and the SLOPE penalty. As before, the grey surface indicates the no-short-sale-area.

From Figure 6, we can observe two important characteristics of SLOPE: First,

Figure 6: Hidden Factors Minimum-Variance Weight Profiles



(a) (b)

The figure shows the weight profile of the oracle (top) and actual (bottom) solution of the LASSO and the SLOPE penalty, considering a minimum variance setup. All values are computed based on a Hidden Factor Structure, with three risk factors and considering for the exponentially decreasing sequence of lambda parameters, a grid of 100 log spaced starting points for λ_1 from 10^{-5} (i.e. x-value = 1) to 10^{2} (i.e. x-value = 100). Equally colored weights characterize assets with the same underlying factor exposure.

while the LASSO shrinks the weights up until the no short sale area, all non-zero coefficients still receive a different weight, independent of their underlying factor exposures. SLOPE, on the other hand, is able to identify the three distinct types of securities, consistent with the true model, and groups them together, by assigning the same coefficient values to them. This provides information about the dependence structure among the assets, and gives the investor the flexibility to select from the groups the assets, which best fit her individual preferences. Not surprisingly, the oracle risk starts to form groups among the securities even before entering into the no short sale area, while the actual weights can only capture the underlying structure much later, and when we already impose a larger tuning parameter value. Second, and different to the LASSO, increasing the lambda parameters past the point of the GMV-LO, the octagonal shape of the penalty pushes the solution towards the equally weighted portfolio. That is, the aforementioned grouping effect increases, and all weights - even those that were shrunken towards zero - are assigned the same coefficient value of $\frac{1}{k}$. Given that the equally weighted portfolio is only optimal when all assets have the same risk and return characteristics, in our example, this allocation results in higher portfolio risk when compared to the GMV-LO or GMV portfolios.

Summing up, SLOPEs properties allows investors to set up sophisticated asset allocation strategies, exploiting its grouping property, like SLOPE-EW, which we introduce in Section 4.

4. Empirical Analysis

4.1. Set up and Data

This section studies the out-of-sample performance of the SLOPE procedure in a minimum variance framework (see i.e. Jagannathan and Ma (2003); Brodie et al. (2009); DeMiguel et al. (2009a); Giuzio and Paterlini (2016)) and compare it with state-of-the-art portfolio selection methods, such as the EW, the GMV, the GMV-LO, the equal risk contribution (ERC), the RIDGE and the LASSO portfolio. Furthermore, we examine two extensions to our standard SLOPE procedure: (1) SLOPE with an added long-only constraint (SLOPE-LO) and (2) a portfolio in which we utilize SLOPE-LO's selection and grouping ability (SLOPE-EW). For the latter, the portfolio is initialized in $t = 1$, keeping for each data set only the first G groups with the largest estimated parameter values active, while setting the remaining weights equal to zero. Afterwards, the portfolio is re-scaled, such that the weights sum again to one. At each subsequent t, we then rebalance the portfolio, if there is a statistically significant difference in the covariance matrices, at $\alpha = 0.1$, to the last re-balance date.[9]

In the following analysis, we consider four data sets, including the monthly log-return observations for the 10- and 30 Industry Portfolios (Ind), the 100 Fama French (FF) portfolios, formed on Size and Book-to-Market, as well as the daily returns of the SP500. The monthly portfolio values are taken from Kenneth French's Homepage[10] and span the period from January 1970 to January 2017 ($T = 565$ monthly observations). The daily return data are obtained from Datastream, covering the period from 31.12.2004 to 31.01.2016 ($T = 2890$ daily observations). Table 1 reports the descriptive statistics. As shown by the skewness and the kurtosis values, all of them exhibit the typical return time series characteristics, including fat tails and slight asymmetry.

Table 1: Descriptive Statistics of the Dataset

Dataset	T	k	$\hat{\mu}$	$\hat{\sigma}$	$\widehat{med}$	$\widehat{min}$	$\widehat{max}$	$\widehat{skew}$	$\widehat{kurt}$	*period*	*freq.*
10Ind	565	10	0.099	0.043	0.012	-0.211	0.156	-0.476	5.077	01/1970 - 01/2017	Monthly
30Ind	565	30	0.010	0.048	0.012	-0.255	0.179	-0.507	5.749	01/1970 - 01/2017	Monthly
100FF	565	100	0.011	0.052	0.015	-0.262	0.241	-0.551	5.600	01/1970 - 01/2017	Monthly
SP500	2890	443	0.000	0.014	0.000	-0.107	0.109	-0.418	13.234	12/2004 - 01/2016	Daily

The table reports descriptive summary statistics for the 10 Industry Portfolios, the 30 Industry Portfolios, the 100 Fama French Portfolios and the S&P 500, respectively. Reported are for the daily (monthly) data: the number of observations (T), the number of constituents(k), the mean ($\hat{\mu}$), the standard deviation ($\hat{\sigma}$), the median ($\widehat{med}$), the minimum ($\widehat{min}$), the maximum ($\widehat{max}$), the skewness ($\widehat{skew}$), the kurtosis ($\widehat{kurt}$), the period that the data set covers (*period*) and the frequency (*freq.*).

To evaluate the portfolios in an out-of-sampling setting, we rely on a rolling window approach with a window size of $\tau = 120$ monthly observations for the 10Ind, the

[9]We use the MBox test (see e.g. Stevens (1992)) to test, if there is a significant difference in the covariance matrices. Furthermore, we perform robustness tests for $\alpha = 0.01, 0.05$, which are available from the authors upon request.

[10]`http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/`

30Ind, and the 100FF, as well as $\tau = 500$ daily observations for the SP500.[11] All portfolios are re-balanced monthly, discarding the oldest and including the most recent observations, allowing for a total of $t = 445$ ($t = 115$) out-of-sample returns for the monthly (daily) data.

The rolling window approach for the daily data works as follows: the first τ return observations are used to estimate $\hat{\boldsymbol{\Sigma}}_t$, according to the shrinkage approach by Ledoit and Wolff (2004b). Then, $\hat{\boldsymbol{\Sigma}}$ is used as the input to compute the optimal weight vector $\hat{\boldsymbol{w}}_t$. The resulting portfolio is assumed to be held for the following 21 days. At $t+1$, the k constituents' returns over this monthly period, $\boldsymbol{R}_{t+1}$, are used to compute the out-of-sample portfolio return as: $R_{p,t+1} = \hat{w}_t \boldsymbol{R}_{t+1}$. In the next step, we roll the data window forward, dropping the last and adding the most recent 21 observations to our training set. We then estimate a new weight vector, which determines our portfolio holdings and the out-of-sample return for the next month. This process is repeated until the end of the data set is reached. The same procedure is applied to the Industry and Fama French portfolios, though the window is rolled forward by one monthly observation instead of 21 daily observations.

Finally, and given that each data set consists of a different number of assets, we keep for our trading strategy, SLOPE-EW, and depending on the respective data set the following first G groups active: 10Ind - the first 4 groups; 30Ind - the first 2 groups; 100Ind - the first 5 groups; SP500 - the first 5 groups.[12]

[11] To test the robustness of our results, we account for different window sizes of $\tau = 250, 750$ and 1000 daily observations, and make the results available upon request. The obtained results are qualitative similar.

[12] Note: The stated number of active groups G for each data set are selected, such that we always obtain the portfolios with the lowest out-of-sample variance, given $\alpha = 0.1$ and when compared to portfolios that would include more or less groups, respectively. Results for different number of included groups G, as well as different significance levels, i.e. $\alpha = 0.05$ or 0.01, are available from the author upon request.

For all portfolios, the optimal weights vector, $\hat{\boldsymbol{w}}_t$, depends on the choice of the optimal λ parameter value. To select the optimal tuning parameter, we consider a grid of 100 log-spaced values of λ between $10^{-7.5}$ and 10^1, from which we choose $\lambda_{RIDGE} = \lambda_{LASSO} = \lambda_1 = \alpha\Phi^{-1}\left(1 - \frac{0.01}{2k}\right)$. The remaining elements $i = 2, ..., k$ of the λ sequence for SLOPE are as before, equal to $\lambda_i = \alpha\Phi^{-1}\left(1 - \frac{0.01i}{2k}\right)$.

Among the 100 lambda values, we select the optimal tuning parameter for the RIDGE, the LASSO and the SLOPE, such that we obtain a portfolio with approximately 10% of the GMV's short positions. Note that for SLOPE, as we increase the tuning parameter, beyond the GMV-LO solution, we would move along the no-short sale area towards the EW portfolio (see Figure 5). Therefore, we also compute SLOPE-LO to explicitly exploit the grouping feature that predominates in the long-only area, and select the lambda value, which provides us with a portfolio that has the largest number of groups. To guarantee that all our portfolios can also be implemented in practice, all weights that are smaller in absolute value than the threshold of 0.05% are set to zero. Furthermore, we incorporate a transaction cost (TC) regime of TC = 10bps[13], whereas the costs are proportional to the turnover and considered to be the same for selling and buying securities.[14]

Given the optimal portfolio vector $\hat{\boldsymbol{w}}_t$ at time t, we compute the out-of-sample mean

[13] 1 bps = 0.01% = 0.0001

[14] For the robustness of our results, we also consider regimes of no (TC = 0bps) and high transaction costs (TC = 50bps). They show that naturally, high turnover strategies like the GMV suffer with regard to returns and the SR in the higher cost regimes. On the other hand, SLOPE portfolios show a nearly steady performance for all data sets and when considering the different TC regimes. All results are available upon request from the authors.

and the out-of-sample standard deviation, defined as:

$$\hat{\mu}_p = \frac{1}{t}\sum_{i=1}^{t} \hat{\boldsymbol{w}}_t \boldsymbol{R}_{t+1} \tag{13}$$

$$\hat{\sigma}_p = \sqrt{\frac{1}{t-1}\sum_{i=1}^{t} (\hat{\boldsymbol{w}}_t \boldsymbol{R}_{t+1} - \hat{\mu}_p)^2} \tag{14}$$

from which we construct the out-of-sample Sharpe Ratio (SR) as:

$$\widehat{SR} = \frac{\hat{\mu}_p}{\hat{\sigma}_p} \tag{15}$$

To evaluate whether the $\widehat{SR}$ and $\hat{\sigma}_p^2$ of any portfolio is statistically different from our SLOPE procedure, we use the tests developed by Ledoit and Wolf (2008) and Ledoit and Wolf (2011), respectively.

As frequent re-balancing of a portfolio is costly, we complement our analysis by computing the turnover of each portfolio, defined as:

$$\widehat{TO} = \frac{1}{t}\sum_{i=1}^{t} ||\hat{\boldsymbol{w}}_{t+1} - \hat{\boldsymbol{w}}_t^{+}||_1, \tag{16}$$

whereas $\hat{\boldsymbol{w}}_t^{+}$ is the weight vector right before rebalancing at $t+1$ and considering the changes in the assets prices. Consequently, the TO for the EW can be non-zero, as $\hat{\boldsymbol{w}}_t^{+} \neq \hat{\boldsymbol{w}}_{t+1} = 1/k$ (DeMiguel et al., 2009a).

Furthermore, we include the following diversification measures: the Diversification Ratio (DR), the weight (WDiv) and the risk diversification (RDiv) measures. The DR is defined as the ratio of the weighted asset volatility to the overall portfolio volatility:

$$\widehat{\mathrm{DR}} = \frac{\sum_{i=1}^{k} \hat{w}_i \hat{\sigma}_i}{\hat{\sigma}_p} \quad , \tag{17}$$

where $\hat{\sigma}_i$ is the i-th asset's estimated volatility, $\hat{\sigma}_p$ is the estimated portfolio volatility, for which the investor typically prefers a higher value (Choueifaty and Coignard, 2008).

Finally, both the WDiv and RDiv measure the concentration of the portfolio in terms of weights and risk (Maillard et al., 2010; Roncalli, 2013). The WDiv ranges from $\frac{1}{k}$ for a perfectly concentrated portfolio up to 1 for the equally weighted portfolio. It is computed according to:

$$\widehat{\mathrm{WDiv}} = \frac{1}{k \times \sum_{i=1}^{k} \hat{w}_i^2} \tag{18}$$

On the other hand, we obtain the RDiv by substituting the weights for the risk contribution, defined as $\widehat{RC}_i = \hat{w}_i \times \partial_{w_i} \sigma(\hat{\boldsymbol{w}}_i)$, where $\partial_{w_i} \sigma(\hat{\boldsymbol{w}}_i)$, defines the marginal contribution to risk (MRC) of asset i, that is the first derivative of the portfolio variance with respect to portfolio weight w_i. The MRC measures the sensitivity of the portfolio variance, given a change in the i-th asset. The RDiv takes a value of 1 for the equally-weighted risk contributions (ERC) portfolio, which is least concentrated in terms of risk contributions and $\frac{1}{k}$ for a portfolio which is fully concentrated on one asset:

$$\widehat{\mathrm{RDiv}} = \frac{1}{k \times \sum_{i=1}^{k} \widehat{RC}_i^2} \tag{19}$$

Summing up, we prefer values close to one for the WDiv and the RDiv (Cazalet et al., 2014).

4.2. Empirical Results

Table 2 reports the annualized out-of-sample volatility, the annualized out-of-sample SR, the number of active positions, and the turnover for the 10Ind, 30Ind, the 100FF, and the SP500, using a window size of $\tau = 120$ ($\tau = 500$) observations with monthly re-balancing and $TC = 10$bps. We indicate portfolios that are statistically different from our SLOPE procedure at the 10%, 5% and 1% level, given the test for the difference in the SR and the volatility, following Ledoit and Wolf (2008) and Ledoit and Wolf (2011).

Table 2: Risk- and Return Measures

	Vol. (in %)				Sharpe Ratio				AP				Turnover			
	10Ind	30Ind	100FF	SP500	10Ind	30Ind	100FF	SP500	10Ind	30Ind	100FF	SP500	10Ind	30Ind	100FF	SP500
EW	14.491	16.257	17.509	20.238	0.776**	0.656***	0.677***	0.205	10.000	30.000	100.000	443.000	0.049	0.057	0.056	0.077
GMV	10.910	9.152	6.058	11.497	1.102*	1.283***	3.124***	0.057**	9.982	29.885	99.469	434.377	0.125	0.273	0.852	2.748
GMV-LO	11.473	11.214	13.134	10.825	1.012	0.998***	0.954***	0.389	5.371	8.562	9.220	27.553	0.064	0.074	0.101	0.238
ERC	13.578	15.029	16.906	17.948	0.840*	0.730***	0.714***	0.233	10.000	30.000	100.000	443.000	0.048	0.054	0.055	0.076
RIDGE	11.907	12.241	13.219	11.393	0.978	0.955	1.069***	0.490	9.989	29.824	98.171	408.474	0.061	0.078	0.109	0.212
LASSO	11.364	10.781	10.853	9.505	1.028	1.046	1.421***	0.572	6.755	12.301	18.371	130.211	0.079	0.104	0.184	0.434
SLOPE	11.352	10.822	10.977	9.643	1.024	1.047	1.382	0.534	7.027	13.231	22.598	145.552	0.078	0.101	0.172	0.409
SLOPE - LO	11.689	11.865	13.709	11.981	0.950**	0.946**	0.917***	0.344	7.299	18.465	34.616	129.632	0.119	0.217	0.409	0.590
SLOPE - EW	12.539	12.904	14.390	11.017	0.908**	0.929*	0.858***	0.645	6.330	6.128	17.022	76.386	0.052	0.055	0.057	0.284

The table reports the out-of-sample Risk and Return Measures for the 10-, 30-, and 100-Portfolios (SP500), considering a windowsize of $\tau = 120$ monthly ($\tau = 500$ daily) observations and re-balancing the portfolio every month over the period from 01/1970 to 01/2017 (from 12/2004 to 01/2016). Furthermore, we consider a transaction cost of 10bps, which is proportional to the turnover and is assumed to be the same for selling and buying securities. Reported are: The annualized out-of-sample volatility, the annualized out-of-sample Sharpe Ratio, the number of active positions (AP), and the average total turnover. Furthermore, we report the significance for the test of the difference in the volatility and the SRs with regard to SLOPE, at the 10%, 5% and 1% level with *, ** and ***, respectively.

Looking at the values for the out-of-sample volatility in Table 2, we observe that no portfolio is statistically different from our new SLOPE procedure, across any of the data sets. Still, SLOPE yields consistently lower variance than any of the EW, ERC, RIDGE or GMV-LO portfolios. Furthermore, for the SP500, SLOPE and LASSO perform best, reporting the smallest variance among all strategies. Especially for the SP500, the number of observations in the window size is only marginally bigger than the size of our investment universe, the estimated covariance matrix is degenerated and our estimates are very prone to estimation error. Therefore, and even using the shrunken covariance matrix, SLOPE and LASSO are still able to reduce extreme weight estimates. Simultaneously, we explicitly select for the LASSO and the SLOPE, a portfolio with a moderate amount of short sales, making it possible to still exploit diversification benefits. Hence, the resulting allocation has a smaller variance, as compared to the GMV-LO.

Furthermore, the values for the out-of-sample SR, establish SLOPE among the best performing portfolios, across all datasets, with some results being statistically significant. For example, SLOPE is able to statistically significantly outperform the EW, challenging its widely reported characteristic of a tough benchmark to beat (DeMiguel et al., 2009b).

Beside reducing the overall portfolio variance, our goal is to construct sparse portfolios with a low turnover. For that, reconsider that the EW always invests naively in all constituents and thus has the highest possible number of active positions. Similar values are obtained for the ERC, which aims at equalizing the risk contribution of each asset to the overall portfolio risk. The GMV, as being highly sensitive to even small changes in the underlying data structure, typically resulting in extreme positions (see i.e. Michaud (1989)), has the highest turnover values among the non-regularization strategies. The RIDGE, on the other hand, results in more stable asset allocations, despite not setting any asset weight exactly equal to zero. Although both strategies should invest in all assets, the reported number of active positions are slightly reduced, due to our imposed threshold of 5%.

Compared to the strategies above, our new SLOPE procedure is able to promote sparse solutions and to reduce the overall portfolio turnover, consistently reporting lower turnover values than the LASSO.

Of special interest is also the performance of SLOPE-EW. In general, our new SLOPE procedure provides the investor with a large amount of flexibility, as with an increased lambda value the penalty starts to form groups among assets, assigning to them the same coefficient value. This is of special interest for investors, who want to move beyond the property of statistical shrinkage, and who want to include in their portfolio construction process any form of financial indicator, like among others fundamental multiples (i.e. Price/Earnings, Dividends/Earnings), accounting values (i.e., Net Income, Free Cash Flow) or other quantitative measures (i.e., Value-at-Risk or Expected Shortfall). With SLOPE-EW, we construct a simple strategy that selects, out of the formed groups, those which carry assets that are the most important with regard to minimizing the overall portfolio variance. Still, other strategies could be easily developed.

Table 2 shows that SLOPE-EW performs best in reducing the variance for large asset universes, i.e. for the SP500, even outperforming the initial SLOPE-LO portfolio. This result provides two insights: First, using SLOPE-EW, we can eliminate assets from the portfolio that rather increase the portfolio variance, as opposed to reducing it, and second, by eliminating the groups according to the weight magnitude, we might conclude that SLOPE assigns assets to the groups according to their importance in reducing the overall variance. Finally, SLOPE-EW ranks among the portfolios with the smallest number of active positions, and reports the lowest turnover value, across all sparse portfolio methods.

Table 3 complements our risk and return analysis, reporting the DR, the WDiv, and

Table 3: Diversification Measures

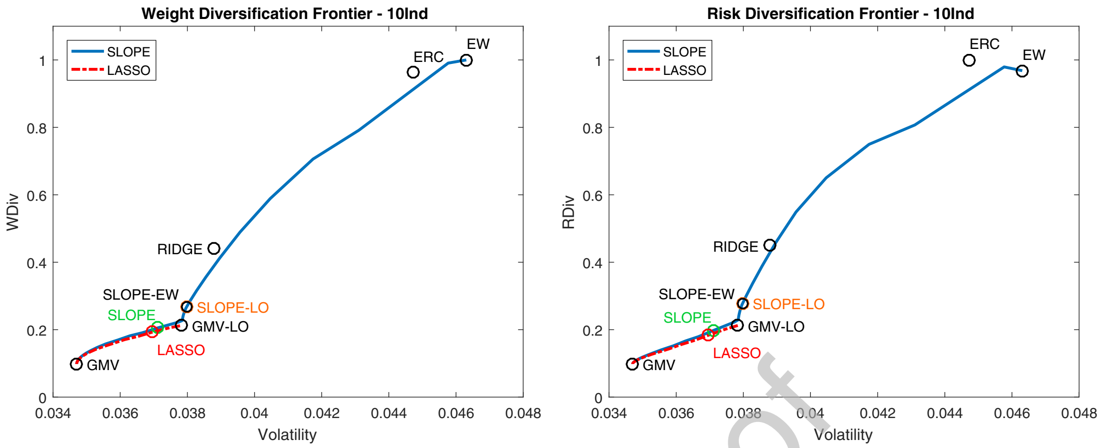
	DR				WDiv				RDiv			
	10Ind	30Ind	100FF	SP500	10Ind	30Ind	100FF	SP500	10Ind	30Ind	100FF	SP500
EW	1.270	1.343	1.212	1.675	1.000	1.000	1.000	1.000	0.933	0.935	0.958	0.894
GMV	1.255	1.362	0.958	3.147	0.197	0.078	0.013	0.012	0.197	0.078	0.013	0.012
GMV-LO	1.289	1.414	1.299	1.944	0.320	0.150	0.062	0.032	0.320	0.150	0.062	0.032
ERC	1.300	1.382	1.225	1.728	0.935	0.914	0.963	0.880	1.000	1.000	1.000	1.000
RIDGE	1.330	1.457	1.256	1.920	0.540	0.430	0.262	0.248	0.577	0.440	0.188	0.120
LASSO	1.289	1.415	1.237	2.221	0.309	0.143	0.044	0.062	0.301	0.132	0.030	0.029
SLOPE	1.295	1.426	1.247	2.213	0.319	0.155	0.054	0.071	0.312	0.144	0.056	0.032
SLOPE - LO	1.315	1.457	1.295	1.936	0.417	0.287	0.209	0.206	0.437	0.319	0.221	0.219
SLOPE - EW	1.289	1.314	1.294	1.808	0.403	0.181	0.118	0.170	0.408	0.182	0.120	0.166

The table reports the diversification measures for the 10-, 30-, and 100- Portfolios (SP500 Portfolios), considering a windowsize of $\tau = 120$ monthly ($\tau = 500$ daily) observations and re-balancing the portfolio every month over the period from 01/1970 to 01/2017 (from 12/2004 to 01/2016). Reported are: The Diversification Ratio (DR), the Weight Diversification (WDiv) and the Risk Diversification (RDiv) measures.

the RDiv. As the EW invests equally in all assets, it achieves, by definition, the best values for the WDiv, with similar values reported for the ERC. As the ERC aims to equalize the contribution to portfolio risk from each asset, it also reports the highest values for the RDiv. SLOPE-LO and SLOPE consistently outperform the LASSO across all datasets for the WDiv and the RDiv. Except for the SP500, this is also true for the DR, while a higher value for the LASSO only results due to the lower variance, as reported in Table 2. It should be pointed out that SLOPE does not only frequently outperform the LASSO, but also provides flexibility with regard to the diversification measures. For that, Figure 7 plots the weight- and risk diversification measure against the attainable portfolio volatility for the LASSO and the SLOPE, together with the other portfolio strategies and considering the first window size of $\tau = 120$ observations for the 10Ind.

For both frontiers, the full grid of lambda parameters for the LASSO enables the investor to select only a combination between the GMV and the GMV-LO solution. SLOPE, on the other hand, is able to span a much larger set of portfolios, beginning

Figure 7: Risk and Weight Diversification Frontier



The figure shows on the left the weight diversification and on the right the risk diversification frontier, both reporting on the x-axis the portfolio volatility and on the y-axis the risk and weight diversification measure, respectively. Considered are the first window size of $\tau = 120$ months for the 10Ind. Plotted are the resulting combinations for the GMV, the GMV-LO, the EW, the ERC, as well as the different combinations for the LASSO and the SLOPE procedure, considering a range of lambda values from $10^{-7.5}$ to 10^{1}.

from the GMV, via the GMV-LO up to the EW. The investor can thus control the trade-off between diversification and volatility out of a much larger set of portfolios, to find the allocation that best suits her individual preferences.

5. Conclusion

This paper extends the literature on financial regularization by introducing SLOPE to the Markowitz portfolio optimization, discussing its properties and testing its performance with regard to risk and return on simulated and real world data.
SLOPE relies on a sorted ℓ_1-Norm, whose intensity is controlled by a decreasing sequence of $\boldsymbol{\lambda}$ parameters and which penalizes the assets by their rank, providing a natural interpretation of importance. To solve the penalized mean-variance optimization, we propose a novel algorithm based on the Alternating Direction

Method of Multipliers (ADMM). When applied to the LASSO, which is a specific case of SLOPE, this algorithm provides the same accuracy as the state-of-the-art CyCoDe, but is superior with regard to computing time, especially when the asset universe is large.

The simulated hidden risk factor analysis shows that SLOPE has the advantage of still being active in the no short sales area and given an imposed budget constraint. Furthermore, SLOPE can automatically identify assets with the same underlying risk factor exposure and group them together, by assigning the same coefficient value to them. This property is especially desirable for investor planning to incorporate their individual views into the optimization, by selecting assets from these groups according to a specific financial characteristic or individual preferences. We exploit such property by building a simple investment strategy, SLOPE-EW.

Moreover, we investigate the performance of SLOPE for four major data sets to other state-of-art portfolio methods in an out-of-sample setting, considering a rolling window approach, and re-balancing the portfolio every month.

Our results show that SLOPE is able to achieve equal and even better out-of-sample portfolio volatilities and SR, when compared to the LASSO. Although, only part of the differences are statistically significant, SLOPE is able to construct sparse portfolios with reduced turnover. This especially applies to situations with a large amount of estimation error, for example when considering the SP500. Furthermore, our SLOPE-EW portfolio results in very sparse portfolios with even lower turnover than state-of-the-art methods and at the same time maintains a comparable performance. Additionally, SLOPE reports improved values for the DR, the WDiv and the RDiv, while the shape of the penalty extends the frontier of attainable portfolios, ranging from the GMV via the GMV-LO, up to the EW portfolio. This enables the investor to select among them the one that provides her with the desired volatility- and diversification trade-off.

The results establish SLOPE as a valid alternative to state-of-art methods by creating sparse portfolios with a reduced turnover rate, improved risk- and weight diversification, and a high degree of flexibility in the portfolio construction process.

A natural extension to our study is to investigate, how different sequences of lambda parameters would impact the risk and portfolio allocation, and whether the investor should choose them according to the underlying correlation regime of the stock market or his own prior beliefs on the assets.

Appendix A. Derivation of the ADMM Algorithm

In order to facilitate the application of proximal operators involving ρ_λ, we first reformulate (2) - (3) into the following form:

$$\min_{\boldsymbol{w}\in\mathbb{R}^k, \boldsymbol{v}\in\mathbb{R}^k} \frac{\phi}{2}\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} - \boldsymbol{\mu}'\boldsymbol{w} + \rho_\lambda(\boldsymbol{v}) \text{ s.t. } \boldsymbol{w} = \boldsymbol{v}, \ \sum_{i=1}^k w_i = 1 \ , \tag{A.1}$$

where $\rho_\lambda(\boldsymbol{w}) := \sum_{i=1}^k \lambda_i |w|_{(i)}$ is the sorted ℓ_1-Norm corresponding to the sequence $\boldsymbol{\lambda}_{SLOPE} = (\lambda_1, \ldots, \lambda_k)'$ satisfying $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k \geq 0$. To solve (A.1), we design an ADMM (for details, see e.g. Boyd et al. (2011)) algorithm, which is based on using the augmented Lagrangian function of (A.1) and on partial updates for the primal variables. In our case the associated augmented Lagrangian is given as:

$$\begin{aligned} \mathcal{L}_\eta(\boldsymbol{w}, \boldsymbol{v}; \boldsymbol{\alpha}, \beta) = & \frac{\phi}{2}\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} - \boldsymbol{\mu}'\boldsymbol{w} + \rho_\lambda(\boldsymbol{v}) + \boldsymbol{\alpha}'(\boldsymbol{w} - \boldsymbol{v}) + \beta(\boldsymbol{e}'\boldsymbol{w} - 1) \\ & + \frac{\eta}{2}\left\{\|\boldsymbol{w} - \boldsymbol{v}\|^2 + (\boldsymbol{e}'\boldsymbol{w} - 1)^2\right\}, \end{aligned} \tag{A.2}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^k$, $\beta \in \mathbb{R}$, $\boldsymbol{e}_{k\times 1} = (1, ..., 1)'$, $\boldsymbol{I}_{k\times k}$ is the identity matrix, and $\eta > 0$ is a penalty parameter. Compared to the Lagrangian L_0 without the penalty term, the augmented Lagrangian $\mathcal{L}_\eta$ with $\eta > 0$ brings the benefit that the dual objective $g_\eta(\boldsymbol{\alpha}, \beta) := \inf_{\boldsymbol{w},\boldsymbol{v}} L_\eta(\boldsymbol{w}, \boldsymbol{v}; \boldsymbol{\alpha}, \beta)$ becomes differentiable without requiring further assumptions on the primal objective (e.g., strict convexity).

The ADMM algorithm consists of the updates:

$$\begin{cases} w^{j+1} & = \arg\min_{\boldsymbol{w}} \; \mathcal{L}_\eta(\boldsymbol{w}, \boldsymbol{v}^j; \boldsymbol{\alpha}^j, \beta^j) = (\phi\boldsymbol{\Sigma} + \eta(I + \boldsymbol{e}\boldsymbol{e}'))^{-1}(\boldsymbol{\mu} - \boldsymbol{\alpha}^j - \beta^j\boldsymbol{e} + \eta(\boldsymbol{v}^j + \boldsymbol{e})) \\ v^{j+1} & = \arg\min_{\boldsymbol{v}} \; \mathcal{L}_\eta(\boldsymbol{w}^{j+1}, \boldsymbol{v}; \boldsymbol{\alpha}^j, \beta^j) = \mathrm{prox}_{\lambda/\eta}(w^{j+1} + (1/\eta)\alpha^j) \\ \alpha^{j+1} & = \alpha^j + \eta(w^{j+1} - v^{j+1}) \\ \beta^{j+1} & = \beta^j + \eta(\mathbf{e}'w^{j+1} - 1) \;\; , \end{cases} \tag{A.3}$$

where $\mathrm{prox}_{\lambda/\eta}(\boldsymbol{z}) := \arg\min_{\boldsymbol{v}} \; \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{z}\|_2^2 + \rho_{\lambda/\eta}(\boldsymbol{v})$ is the proximal operator of the Sorted ℓ_1-Norm, corresponding to the sequence λ/η, provided e.g. in Bogdan et al. (2013, 2015). The updates regarding α and β are due to the gradient ascent applied to the dual objective $g_\eta(\boldsymbol{\alpha}, \beta) := \inf_{\boldsymbol{w},\boldsymbol{v}} L_\eta(\boldsymbol{w}, \boldsymbol{v}; \boldsymbol{\alpha}, \beta)$, where $\nabla_{\boldsymbol{\alpha}} g_\eta(\boldsymbol{\alpha}, \beta) = w^{j+1} - v^{j+1}$ and $\nabla_\beta g_\eta(\boldsymbol{\alpha}, \beta) = \boldsymbol{e}'w^{j+1} - 1$. The first iterates $\boldsymbol{w}^0$, $\boldsymbol{v}^0$, $\boldsymbol{\alpha}^0$, β^0 of the procedure (4) are typically initialized as the zero vectors.

Appendix B. Primal-Dual Gap

The stopping criterion for our algorithm is based on the Primal-Dual Gap, which we estimate as follows. First, taking the infimum over $(\boldsymbol{w}, \boldsymbol{v})$ of the Lagrangian, we get the dual objective,

$$g(\boldsymbol{\alpha}, \beta) = \inf_{\boldsymbol{w}} \frac{\phi}{2}\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} - (\boldsymbol{\mu} - \boldsymbol{\alpha} - \beta\boldsymbol{e})'\boldsymbol{w} - \beta - \rho^*_{\boldsymbol{\lambda}}(\boldsymbol{\alpha}). \tag{B.1}$$

From the optimality condition for the infimum over $\boldsymbol{w}$, we have

$$\boldsymbol{w}^* = \phi^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\alpha} - \beta \boldsymbol{e}). \tag{B.2}$$

Also,

$$\rho^*_{\boldsymbol{\lambda}}(\boldsymbol{\alpha}) = \sup_{\boldsymbol{v}}\{\boldsymbol{\alpha}^T\boldsymbol{v} - \rho_{\boldsymbol{\lambda}}(\boldsymbol{v})\} = \begin{cases} 0 & \text{if } \boldsymbol{\alpha} \in C_{\boldsymbol{\lambda}} \\ +\infty & \text{o.w.} \end{cases} \tag{B.3}$$

where $C_{\boldsymbol{\lambda}} := \{\boldsymbol{v} : \mathbb{R}^k : \rho^D_{\boldsymbol{\lambda}}(\boldsymbol{v}) \leq 1\}$ is the unit sphere defined in the dual norm $\rho^D_{\boldsymbol{\lambda}}(\cdot)$ of $\rho_{\boldsymbol{\lambda}}(\cdot)$. Plugging-in these, we get the dual problem

$$\max_{\boldsymbol{\alpha},\beta} \; -\frac{1}{2\phi}(\boldsymbol{\mu} - \boldsymbol{\alpha} - \beta\boldsymbol{e})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\alpha} - \beta\boldsymbol{e}) - \beta \quad \text{s.t. } \boldsymbol{\alpha} \in C_{\boldsymbol{\lambda}}. \tag{B.4}$$

Then we can estimate the primal-dual gap as follows using (B.2),

$$\begin{aligned} \mathcal{G}(\boldsymbol{w}^*, \boldsymbol{v}^*, \boldsymbol{\alpha}^*, \beta^*) &= \frac{1}{2}\phi(\boldsymbol{w}^*)\boldsymbol{\Sigma}\boldsymbol{w}^* - \boldsymbol{\mu}'\boldsymbol{w}^* + \rho_{\boldsymbol{\lambda}}(\boldsymbol{w}^*) + \frac{1}{2\phi}(\boldsymbol{\mu} - \boldsymbol{\alpha}^* - \beta^*\boldsymbol{e})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\alpha}^* - \beta^*\boldsymbol{e}) + \beta^* \\ &= -(\boldsymbol{\alpha}^* + \beta^*\boldsymbol{e})'\boldsymbol{w}^* + \beta^* + \rho_{\boldsymbol{\lambda}}(\boldsymbol{v}^*) \end{aligned} \tag{B.5}$$

given the dual feasibility of $\boldsymbol{\alpha}^*$, i.e., $\rho^D_{\boldsymbol{\lambda}}(\boldsymbol{\alpha}^*) \leq 1$. Here, $\boldsymbol{w}^*$, $\boldsymbol{v}^*$, $\boldsymbol{\alpha}^*$, β^* can be generated from the procedure (4), and, due to strong duality, the duality gap becomes zero when these iterates are optimal to the problem (A.1). Therefore we can stop our algorithms when the duality gap of the current iterates becomes sufficiently small.

Appendix C. ADMM vs. Cyclic Coordinate Descend

In this section, we use the ADMM algorithm to solve the minimum-variance optimization with an ℓ_1 Norm (which is a specific instance of our new SLOPE penalty) and compare its performance to the the Cyclic Coordinate Descend algorithm (CyCoDe).

The CyCoDe algorithm is considered state-of-art and has found various applications in solving norm constrained optimization problems (see i.e. Fastrich et al. (2014), Yen (2015)). The algorithm works by optimizing the weights along one coordinate direction, while holding all other weights constant. Although there is no general rule on how the CyCoDe updates the weight vector, we follow the procedure of Yen (2015) and update the weights cyclical, that is we first fix $w_i,\ i = 2, ..., k$ and find a new solution for w_1 that is closer to its optimal solution w^*. In a next step, we fix $w_i,\ i = 1, 3, ..., k$ and find a value for w_2 that is again closer to the optimal one w^*. Given a starting criteria $\boldsymbol{w}^0$ for the weight vector, the Lagrange parameter, γ, for the budget constraint and a trade-off parameter, θ, for μ and σ^2, Algorithm 2 shows the pseudo code for the CyCoDe.

Algorithm 2 Cyclic Coordinate Descend

1: Initialize $\boldsymbol{w}^0$ and $j = 0$
2: **while** convergence criteria is not met **do**
3:     **for** $i = 1$ to k **do**
4:         $w_i = ST(\gamma - z_i, \lambda) \times (2 \times \sigma_i^2)^{-1}$
5:         where ST is the soft-thresholding function and $z_i = 2\sum_{j\neq i}^{k} w_j\sigma_{ij} - \theta\mu_i$
6:     **end for**
7:     $j = j + 1$
8: **end while**

To evaluate the performance of the two algorithms, we first draw a random sample of size n for k assets from a multivariate normal $\boldsymbol{X} \sim MVN(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$:

$$\Sigma_{ij} = \begin{cases} 1, & i = j, \\ \rho, & i \neq j, \end{cases} \tag{C.1}$$

and for which we choose $\rho = 0.2$ and 0.8, respectively. Then, we solve the minimum variance problem given in (2) and subject to the ℓ_1- Norm on the weight vector, using as an input for $\boldsymbol{\Sigma}$ the shrunken covariance matrix, introduced by Ledoit and Wolff (2004b).

We initialize both algorithms with a soft starting point $\boldsymbol{w}^0$, that is (1) $\boldsymbol{w}_i^0 = \frac{1}{k}\ \forall\ i = 1, ..., k$, and (2) $\boldsymbol{w}_i^0 = \frac{a_i}{\sum_{i=1}^{k} a_i}$, with $a_i \sim U(0,1)\ \forall\ i = 1, ..., k$, and repeat the above procedure 100 times, using for both algorithms a tolerance stopping point of 10^{-7}. All computations are performed in Matlab 2016a on a Lenovo T430, with Windows 7, an Intel i7-3520M with 2.90 GHZ and 8 GB of RAM.

Table C.4 and C.5 display the minimum and the median of the objective function values, together with the median amount of shorting, the median time in seconds used for each algorithm to solve the 100 simulations and the median absolute weight difference[15], considering as soft starting criteria an equally weighted and a random portfolio weight vector, respectively.[16]

The tables show that both algorithms reach the same global minimum and median objective function value and the same amount of shorting for the low correlation environment, regardless of the chosen lambda value and whether we consider the

[15]The difference in the weights is computed as: $\sum |\boldsymbol{w}^{ADMM} - \boldsymbol{w}^{CyCoDe}|$, where $\boldsymbol{w}^{ADMM}$ and $\boldsymbol{w}^{CyCoDe}$ are the optimal weights obtained with the ADMM and the CyCoDe algorithm, respectively.

[16]Due to space limitations, we have restricted ourselves to report the above mentioned measures. Further results, including the standard deviation of the objective function value and the median number of active positions are available upon request to the authors.

Table C.4: Simulation Results - Equal Weights

				$\lambda = 4.03 \times 10^{-6}$					$\lambda = 5.65 \times 10^{-4}$					$\lambda = 7.91 \times 10^{-2}$				
ρ	n	p	Algo	Min	Med	Short	Time	W.Diff.	Min	Med	Short	Time	W.Diff.	Min	Med	Short	Time	W.Diff.
0.2	500	100	CyCoDe	0.14	0.16	0.51	0.66	5×10^{-7}	0.14	0.16	0.49	0.62	5×10^{-7}	0.23	0.25	0.00	0.18	7×10^{-8}
			ADMM	0.14	0.16	0.51	0.01		0.14	0.16	0.49	0.01		0.23	0.25	0.00	0.01	
	500	250	CyCoDe	0.09	0.11	2.13	13.63	8×10^{-6}	0.09	0.11	2.02	12.87	6×10^{-6}	0.21	0.24	0.00	0.94	8×10^{-8}
			ADMM	0.09	0.11	2.13	0.09		0.09	0.11	2.02	0.09		0.21	0.24	0.00	0.03	
	1000	500	CyCoDe	0.09	0.10	3.46	117.69	3×10^{-5}	0.09	0.11	3.23	116.29	2×10^{-5}	0.22	0.24	0.00	5.58	1×10^{-7}
			ADMM	0.09	0.10	3.46	0.66		0.09	0.11	3.23	0.64		0.22	0.24	0.00	0.17	
0.8	500	100	CyCoDe	0.55	0.64	3.39	11.67	2×10^{-3}	0.55	0.65	3.30	11.23	2×10^{-3}	0.73	0.83	0.00	1.37	8×10^{-7}
			ADMM	0.55	0.64	3.39	0.06		0.55	0.65	3.30	0.05		0.73	0.83	0.00	0.03	
	500	250	CyCoDe	0.34	0.42	10.98	35.33	8×10^{-1}	0.35	0.43	10.46	34.75	8×10^{-1}	0.67	0.82	0.00	6.03	1×10^{-6}
			ADMM	0.34	0.42	10.94	0.58		0.35	0.43	10.47	0.56		0.67	0.82	0.00	0.11	
	1000	500	CyCoDe	0.36	0.42	16.49	109.37	2.1	0.38	0.44	15.44	107.64	1.8	0.75	0.83	0.00	37.20	2×10^{-6}
			ADMM	0.36	0.42	16.34	3.96		0.38	0.43	15.33	3.76		0.75	0.83	0.00	0.61	

The table reports, for the Cyclic Coordinate Descend (CyCoDe) and the Alternating Direction Method of Multipliers (ADMM), the simulation results to the penalized minimum variance problem given in (2), considering six data sets drawn from a multivariate normal distribution, with $\rho = 0.2$ and $\rho = 0.8$, respectively, and using the equally weighted portfolio as a soft starting point. Stated are across all 100 simulations: the minimum (Min) and the median (Med) value of the objective function, the median value of the total amount of shorting (Short) the median time in seconds needed to compute the solution (Time) and the average weight difference (W.Diff.).

equally weighted or the random weight vector as the soft starting point. This also applies to the low dimensional data set, when the correlation is set to $\rho = 0.8$. When $p = 500$ for $\rho = 0.8$, the ADMM reports a lower amount of shorting for the first two lambda values. This holds regardless of how we choose the soft starting point. This difference might also explain the discrepancy in the weight vectors, which is reported to be the highest for these two data sets. Still, the difference in the resulting weight vectors is modest and amounts to an average of 10^{-6} for both low correlation environments, and to 10^{-4}, for the first two high correlation environments and regardless on how we choose the soft starting point.

Most notably, the ADMM outperforms the CyCoDe, with regard to the median time in seconds used to compute the solution for all six data sets. This difference is not negligible: the ADMM uses on average 0.265 seconds in the low correlation environment across all lambdas and all starting criteria, while the CyCoDe is slower by a factor of more than 100, using on average 28.88 seconds. This also applies to the high correlation environment, with the ADMM finding the solution, by taking on average 2.65 seconds and the CyCoDe using 38.98 second. Finally, and for both algorithms, selecting the random weight vector as a starting point results in longer computing times, as opposed to using the equally weighted solution.

Figure C.8 plots the computing times needed for the CyCoDe and the ADMM for

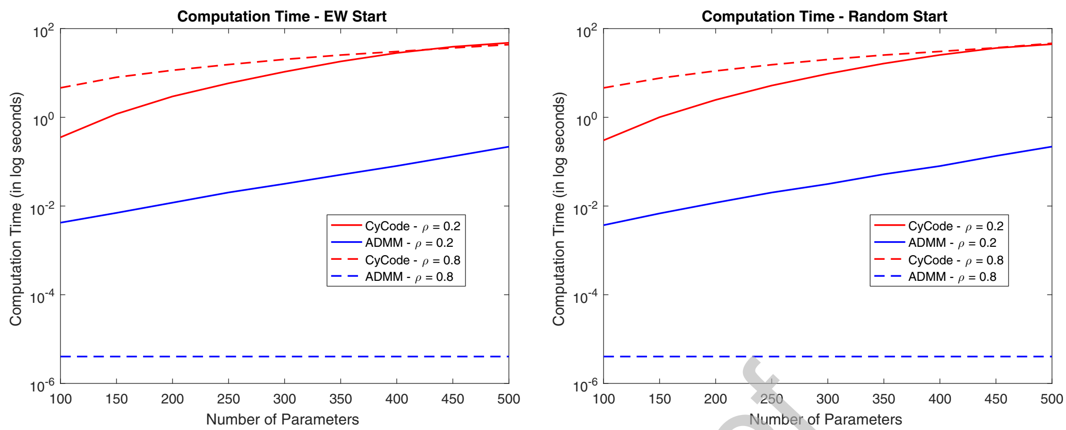
Table C.5: Simulation Results - Random Weights

				$\lambda = 4.03 \times 10^{-6}$					$\lambda = 5.65 \times 10^{-4}$					$\lambda = 7.91 \times 10^{-2}$				
ρ	n	p	Algo	Min	Med	Short	Time	W.Diff	Min	Med	Short	Time	W.Diff	Min	Med	Short	Time	W.Diff
0.2	500	100	CyCoDe	0.13	0.16	0.49	0.46	5×10^{-7}	0.13	0.16	0.47	0.44	4×10^{-6}	0.22	0.25	0.00	0.13	7×10^{-8}
			ADMM	0.13	0.16	0.49	0.01		0.13	0.16	0.47	0.01		0.23	0.25	0.00	0.01	
	500	250	CyCoDe	0.08	0.10	2.12	10.26	8×10^{-6}	0.08	0.10	2.02	10.02	6×10^{-6}	0.19	0.23	0.00	0.74	8×10^{-8}
			ADMM	0.08	0.10	2.11	0.07		0.08	0.10	2.02	0.07		0.19	0.23	0.00	0.02	
	1000	500	CyCoDe	0.08	0.10	3.50	111.66	3×10^{-5}	0.09	0.10	3.28	112.50	2×10^{-5}	0.22	0.24	0.00	5.31	1×10^{-7}
			ADMM	0.08	0.10	3.50	0.52		0.09	0.10	3.28	0.51		0.22	0.24	0.00	0.15	
0.8	500	100	CyCoDe	0.55	0.64	3.30	8.02	2×10^{-3}	0.56	0.64	3.21	7.86	2×10^{-3}	0.72	0.82	0.00	0.89	8×10^{-7}
			ADMM	0.55	0.63	3.30	0.03		0.55	0.64	3.21	0.03		0.72	0.82	0.00	0.02	
	500	250	CyCoDe	0.33	0.41	10.77	31.54	8×10^{-1}	0.35	0.42	10.34	32.05	8×10^{-1}	0.68	0.81	0.00	5.35	1×10^{-6}
			ADMM	0.33	0.41	10.75	0.55		0.35	0.42	10.33	0.53		0.68	0.81	0.00	0.10	
	1000	500	CyCoDe	0.36	0.40	16.42	111.10	2.193	0.37	0.42	15.38	111.70	1.99	0.76	0.82	0.00	38.7	1.81
			ADMM	0.36	0.40	16.37	3.89		0.37	0.42	15.36	3.69		0.76	0.82	0.00	0.60	

The table reports, for the Cyclic Coordinate Descend (CyCoDe) and the Alternating Direction Method of Multipliers (ADMM), the simulation results to the penalized minimum variance problem given in (2), considering six data sets drawn from a multivariate normal distribution, with $\rho = 0.2$ and $\rho = 0.8$, respectively, and using the equally weighted portfolio as a soft starting point. Stated are across all 100 simulations: the minimum (Min) and median (Med) value of the objective function, the median value of the total amount of shorting (Short) the median time in seconds needed to compute the solution (Time) and the average weight difference (W.Diff.).

both the EW and Random weight vector initialization, considering the two correlation regimes and varying the number of parameters that have to be estimated. Clearly the ADMM consistently shows a superior performance, by only using a fraction of the time of the CyCoDe. Furthermore, we can observe that both algorithms are also invariant to the selection of the soft starting point. Only the CyCoDe shows a slight difference for parameter values above $k = 450$, signaling that for the CyCoDe an EW portfolio results in finding the optimal solution faster.

Figure C.8: Computation Times for CyCoDe and ADMM



The figure shows the average computation times needed for the CyCoDe and ADMM algorithm, depending on the correlation regime, the number of parameters and the soft start criterion. All values are based on 100 simulations, considering a constant correlation set-up.

Appendix D. Portfolio Selection Models

Equally Weighted Portfolio. The equally weighted portfolio is considered as one of the toughest benchmarks to beat (see, i.e. DeMiguel et al. (2009b)), and naively distributes the wealth equally among all constituents, such that with k assets:

$$w_i = \frac{1}{k} \ \forall \ i = \{1, ..., k\}, \tag{D.1}$$

where w_i is the weight of asset i. The EW ignores both the variances, the covariances and the return of the assets, and is the optimal portfolio on the mean-variance efficient frontier, when we assume that all three are the same.

Norm-Constrained Minimum Variance Portfolio. Reconsider the formulation of the mean-variance problem in (1). By disregarding the mean in the optimization, we obtain the Global Minimum Variance Portfolio (GMV), given by:

$$\min_{\boldsymbol{w}\in\mathbb{R}^k} \sigma_p^2 = \boldsymbol{w}'\Sigma\boldsymbol{w} \quad s.t. \sum_{i=1}^{k} w_i = 1,\ \forall\ i = \{1, ..., k\}, \tag{D.2}$$

However, this formulation is prone to estimation errors, and unstable portfolio weights. To circumvent these problems, we extend the framework in (D.2) by adding a penalty function $\rho_\lambda(\boldsymbol{w})$ on the weight vector. For LASSO, we add a ℓ_1 - Norm to the formulation in (D.2), such that:

$$\rho_\lambda(\boldsymbol{w}) = \lambda \times \sum_{i=1}^{k} |w_i| \tag{D.3}$$

where λ is a regularization parameter that controls the intensity of the penalty. Besides LASSO, we also consider the RIDGE penalty, which adds an ℓ_2-Norm on the weight vector to the formulation in (D.2), and that takes the form of:

$$\rho_\lambda(\boldsymbol{w}) = \lambda \times \sum_{i=1}^{k} w_i^2 \tag{D.4}$$

As opposed to the LASSO, the RIDGE is not singular at the origin and thus does not promote sparse solutions. Still, imposing the ℓ_2 - Norm on the portfolio problem is equal to adding an identity matrix, weighted by the regularization parameter λ to the inverse of the variance-covariance matrix, i.e. $(\boldsymbol{\Sigma}^{-1} + \lambda\boldsymbol{I})$, where $\boldsymbol{I}$ is the $k \times k$ identity matrix. This leads to more numerical stability and makes the RIDGE penalty especially appealing in environments that suffer from multicollinearity (Zou and Hastie, 2005).

Equal Risk Contribution Portfolio. Finally, we consider the Equal Risk Contribution (ERC) portfolio, which aims to equalize the marginal risk contributions of the assets to the overall portfolio risk. That is, given that portfolio variance can be decomposed as:

$$\sigma_p^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} w_i w_j \sigma_{ij} = \sum_{i=1}^{k} w_i \sum_{j=1}^{k} w_j \sigma_{ij} \tag{D.5}$$

the marginal contribution to the portfolio risk for asset i is given as:

$$c_i^{var} = w_i \sum_{j=1}^{k} w_j \sigma_{ij} = w_i \ (\boldsymbol{\Sigma}\boldsymbol{w})_i \quad \text{with} \quad \sum_{i=1}^{k} c_i^{var} = \sigma_p^2 \tag{D.6}$$

where $(\boldsymbol{\Sigma}\boldsymbol{w})_i$ denotes the i^{th} row of the product of Σ and $\boldsymbol{w}$ (Roncalli, 2013). As the marginal risk is dependent on the portfolio weight magnitude, the ERC portfolio has no analytically solution and must be obtained numerically,by solving:

$$\min_{\boldsymbol{w}\in\mathbb{R}^N} \sum_{i=1}^{k}\left(\frac{w_i \ (\boldsymbol{\Sigma}\boldsymbol{w})_i}{\sigma_p^2} - \frac{1}{k}\right)^2 \quad s.t. \quad \sum_{i=1}^{k} w_i = 1, \quad 0 \le w_i \le 1 \quad \forall\ i \in \{1, 2, ..., k\} \tag{D.7}$$

The ERC favors assets with lower volatility, lower correlation with other assets, or both, and is less sensitive to small changes in the covariance matrix as compared to the GMV portfolio (Kremer et al., 2018). Furthermore, (Maillard et al., 2010) show that the volatility of the ERC is between that of the EW and the GMV, and that it coincides with the latter, when both, correlations and SRs, are assumed to be equal (Maillard et al., 2010).

Acknowledgement

Małgorzata Bogdan acknowledges the grant of the Polish National Center of Science Nr 2016/23/B/ST1/00454, and together with Sandra Paterlini further ac-knowledge ICT COST Action IC1408 from CRoNoS. Sangkyun Lee acknowledges the support of the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP; Ministry of Science, ICT & Future Planning) (No. 2017R1C1B5018367).

References

Bellec, P., Lecué, G., Tysbakov, A., 2016a. Bounds on the prediction error of penalized least squares estimators with convex penalty. arXiv:1609.06675, to appear in Modern Problems of Stochastic Analysis and Statistics, Festschrift in honor of Valentin Konakov.

Bellec, P., Lecué, G., Tysbakov, A., 2016b. Slope meets lasso: improved oracle bounds and optimality. arXiv:1605.08651, 1–29.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., Candes, E., 2015. Slope - adaptive variable selection via convex optimization. Annals of Applied Statistics 9 (3), 1103–1140.

Bogdan, M., van den Berg, E., Su, W., E.J., C., 2013. Statistical estimation and testing via the ordered ℓ_1 norm. arXiv:1310.1969, 1–46.

Bondell, H., Reich, B., March 2008. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. Biometrics 64 (1), 115–123.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning 3 (1), 1–122.

Boyle, P., Garlappi, L., Uppal, R., Wang, T., 2012. Keynes meets Markowitz: The trade-off between familarity and diversification. Management Science 58 (2), 253–272.

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

Carrasco, M., Noumon, N., 2012. Optimal portfolio selection using regularization. Working Paper University of Montreal, 1–52.

Cazalet, Z., Grison, P., Roncalli, T., 2014. The smart beta indexing puzzle. The Journal of Index Investing 5 (1), 97–119.

Chopra, V., Ziemba, W., 1993. The effect of errors in means, variances, and covariances on optimal portfolio choice. Journal of Portfolio Management 19 (2), 6–11.

Choueifaty, Y., Coignard, Y., 2008. Toward maximum diversification. Journal of Portfolio Management 34 (4), 40–51.

DeMiguel, V., Garlappi, L., Nogales, F., Uppal, R., 2009a. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norm. Management Science 55 (5), 798–812.

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

DeMiguel, V., Nogales, F., 2009. Portfolio selection with robust estimation. Operations Research 55 (5), 798–812.

Fan, J., Fan, Y., Lv, J., 2008. High didimension covariance matrix estimation using a factor model. Journal of Econometrics 147 (1), 186–197.

Fan, J., Zhang, J., You, K., 2012. Vast portfolio selection with gross-exposure constraint. Journal of the American Statistical Association 107 (498), 592–606.

Fastrich, B., Paterlini, S., Winker, P., 2014. Cardinality versus q-norm constraints for index tracking. Quantitative Finance 14 (11), 2019–2032.

Fastrich, B., Paterlini, S., Winker, P., 2015. Constructing optimal sparse portfolios using regularization methods. Computational Management Science 12 (3), 417–434.

Figueiredo, M., Nowak, R., 2014. Sparse estimation with strongly correlated variables using ordered weighted ℓ_1 regularization. Working Paper, 1–15.
URL `arXiv:1409.4005`

Gabaix, X., 2014. A sparsity-based model of bounded rationality. The Quarterly Journal of Economics 129, 1661–1710.

Gasso, G., Rakotomamonjy, A., Canu, S., 2010. Recovering sparse signals with a certain family of non-convex penalties and dc programming. IEEE Transactions on Signal Processing 57 (12), 4686–4698.

Giuzio, M., Paterlini, S., 2016. Un-diversifying during crises: Is it a good idea? FRB of Cleveland Working Paper (16-28), 1–38.

Hestenes, M. R., 1969. Multiplier and gradient methods. Journal of Optimization Theory and Applications 4 (5), 303–320.

Jagannathan, R., Ma, T., August 2003. Risk reduction in large portfolios: Why imposing the wrong constraints helps. The Journal of Finance 58 (4), 1651–1683.

Jorion, P., 1986. Bayes-stein estimator for portfolio analysis. The Journal of Financial and Quantitative Analysis 21 (3), 279–292.

Kolm, P., Tütüncü, R., Fabozzi, F. J., 2014. 60 years of portfolio optimization: Practical challenges and current trends. European Journal of Operational Research 234 (2), 356–371.

Kremer, P., Talmaciu, A., Paterlini, S., 7 2018. Risk minimization in multi-factor portfolios: What is the best strategy? Annals of Operations Research 266 (1-2), 255–291.

Ledoit, O., Wolf, M., 2008. Robust performance hypothesis testing with the sharpe ratio. Journal of Empirical Finance 15 (5), 850–859.

Ledoit, O., Wolf, M., 2011. Robust performance hypothesis testing with the variance. Wilmott Technical Paper 2011 (55), 86–89.

Ledoit, O., Wolff, M., 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance 10 (5), 603–621.

Ledoit, O., Wolff, M., 2004a. Honey, i shrunk the covariance matrix. Journal of Portfolio Management 30 (4), 110–119.

Ledoit, O., Wolff, M., 2004b. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88 (2), 365–411.

Li, J., 2015. Sparse and stable portfolio selection with parameter uncertainty. Journal of Business and Economic Statistics 33 (3), 381–392.

Maillard, S., Roncalli, T., Teiletche, J., 2010. The properties of equally weighted risk contribution portoflios. The Journal of Portfolio Management 36 (4), 60–70.

Markowitz, H., March 1952. Portfolio selection. The Journal of Finance 7 (1), 77–91.

Merton, R. C., 1980. On estimating the expected return on the market: An exploratory investigation. Journal of Financial Economics 8 (4), 323–361.

Michaud, R., 1989. The Markowitz optimization enigma: is 'optimized' optimal? Financial Analyst Journal 45 (1), 31–42.

Parikh, N., Boyd, S., 2014. Proximal algorithms. Foundations and Trends in Optimization 1 (3), 127–239.

Powell, M. J. D., 1969. A method for nonlinear constraints in minimization problems. In: Fletcher, R. (Ed.), Optimization. Academic Press, New York, pp. 283–298.

Roncalli, T., 2013. Introduction to Risk Parity and Budgeting. Chapman & Hall/CRC Financial Mathematics Series.

Shefrin, H., Statman, M., 2000. Behavioral portfolio theory. Journal of Financial and Quantitative Analysis 35 (2), 127–151.

Stevens, J., 1992. Applied multivariate Statistics for Social Sciences, 2nd Edition. New-Jersey:Lawrance Erlbaum Associates Publishers, pp. 260 - 269.

Su, W., Candès, E., 2016. Slope is adaptive to unknown sparsity and asymptotically minimax. Annals of Statistics 44 (3), 1038–1068.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society 58 (1), 267–288, series B.

Xing, X., Hub, J., Yang, Y., 2014. Robust minimum variance portfolio with ℓ_∞ constraints. Journal of Banking and Finance 46, 107–117.

Yen, Y.-M., 2015. Sparse weighted norm minimum variance portfolio. Review of Finance 20 (3), 1259–1287.

Zeng, X., Figueiredo, M., 2014. Decreasing weighted sorted l_1 regularization. IEEE Signal Processing Letters 21 (10), 1240–1244.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society B 67 (2), 301–320.