INSEAD
The Business School for the World®

Faculty & Research Working Paper

Regularized Robust Portfolio Estimation

Theodoros EVGENIOU
Massimiliano PONTIL
Diomidis D. SPINELLIS
Rafal SWIDERSKI
Nick NASSUPHIS
2013/79/DS

Regularized Robust Portfolio Estimation

Theodoros Evgeniou*

Massimiliano Pontil**

Diomedis D. Spinellis***

Rafal Swiderski****

Nick Nassuphis*****

* Associate Professor of Decision Sciences and Technology Management, Academic Director, INSEAD eLab at INSEAD, Boulevard de Constance 77305 Fontainebleau, France..
Email: theodoros.evgeniou@insead.edu

** Professor of Computational Statistics and Machine Learning at University College London Gower Street, London WC1E 6BT, United Kingdom. Email: m.pontil@cs.ucl.ac.uk

*** Professor at the Athens University of Economics and Business 28is Oktovriou 76, Athens, Greece. Email: dds@aueb.gr

**** 31, St. Martin's Lane WC2N 4ER London, United Kingdom. Email: rswid@web.de

***** 31, St. Martin's Lane WC2N 4ER London, United Kingdom.
Email: nicknassuphis@gmail.com

Regularized Robust Portfolio Estimation

Theodoros Evgeniou
Decision Sciences and Technology Management
INSEAD
Fontainebleau, France
theodoros.evgeniou@insead.edu

Diomidis Spinellis
Dept. of Management Science and Technology
Athens University of Economics and Business
Athens, Greece
dds@aueb.gr

Nick Nassuphis
31 St. Martin's Lane
WC2N 4ER
London, UK
nicknassuphis@gmail.com

Massimiliano Pontil
Department of Computer Science
University College London
Gower Street, London WC1E 6BT
England, UK
m.pontil@cs.ucl.ac.uk

Rafal Swiderski
31 St. Martin's Lane
WC2N 4ER
London, UK
rswid@web.de

Abstract

We study the problem of learning a combination (e.g. a portfolio) of time series that has large autocorrelation. This is a challenging task as it involves the lag-1 autocovariance matrix of the series, which is difficult to estimate in practice. To address this issue we develop regularized versions of the autocorrelation function based on a robust optimization formulation of the problem. We highlight different forms of regularizers and present a method to solve the underlying optimization problem. We also discuss an extension of our approach to find maximally cross-correlated combinations of time series, which provides a novel class of regularization techniques for canonical correlation analysis. Experiments in the context of financial time series, where the estimation of the lag-1 autocovariance matrix is notoriously difficult, indicate the potential of the proposed approach.

1 Introduction

Given a vector-valued time series, we study the problem of learning the weights of a linear combination of the series' components (e.g. a portfolio), which has large autocorrelation, and discuss the extension to the problem of learning two combinations, which have large cross-correlation. Both problems have been studied from different perspectives in various areas, ranging from computational neuroscience [19], to computer vision [15, 9], to information retrieval [10], among others. In this paper, we address these problems from the point of view of robust optimization (see e.g. [4, 6] and references therein) and regularization, and highlight their application to the context of financial time series analysis, see e.g. [17] and references therein.

The autocorrelation (or cross-correlation) function is a quantity difficult to measure, as it depends on the lag-1 autocovariance matrix of the time series, which is typically unstable. To mitigate this

problem, we propose a robust optimization approach that leads to regularized versions of the autocorrelation (or cross-correlation) function. We describe various forms of regularizers derived from different constraints on the uncertainty region of the lag-1 autocovariance matrix, which in particular induce ℓ_2 or ℓ_1 regularized portfolio estimation methods. We present an optimization algorithm to solve the ℓ_1 regularization problem, which is inspired by recent work on sparse principal component analysis (PCA) [11, 12], also linking this work to the broader literature on sparsity regularization. We then apply the proposed methods to estimate high lag-1 autocorrelation portfolios for financial time series. On the way, we link specific instances of our method to portfolio creation strategies previously considered in the Finance literature. Finally, we extend the proposed approach to the setting of canonical correlation analysis (CCA), an older statistical technique [1, 8], which has seen revived interest in machine learning and statistics, see e.g. [10, 20] and references therein.

In summary, the key contributions of this paper are both methodological, namely developing novel regularization methodologies and optimization based estimation algorithms, as well as theoretical, namely establishing a link between robust optimization and regularization in the context of portfolio estimation. Although the methods are more broadly applicable, we study them in the context of portfolio creation for financial time series. This type of data is among the most challenging ones to develop predictive models for. Another important contribution of this paper is a demonstration of the proposed approach's application potential on financial time series. Although developing portfolio estimation methods for such series that can be used in practice (e.g. for trading) is beyond the scope of this paper, we discuss potential future research that can lead to such methods building on the approach we develop. Over the past few years there has been a rising interest within the financial industry to employ machine learning techniques; this work also builds in that direction.

The paper is organized as follows. In Section 2, we introduce the portfolio learning problem. In Section 3, we address the issue of implementing the learning method numerically. In Section 4, we extend our regularization approach to the setting of canonical correlation analysis. Finally, in Section 5 we report experiments with the proposed methods in the context of financial time series.

2 Finding Robust Autocorrelation Portfolios

We start with the relation between a robust optimization formulation of the maximally autocorrelated portfolio estimation problem and regularization. As the experimental focus of the paper is on financial time series, we first introduce some notation from that context.

2.1 Financial Time Series: Notation and Definitions

Let $r_1, \ldots, r_T \in \mathbb{R}^n$ be the realization of a vector-valued time series over T consecutive time frames (e.g. days). In the experiments, r_t represents the vector of log-returns on day t of n assets (e.g. stocks). A common goal in practice is to learn a weight vector $x \in \mathbb{R}^n$ that maximizes some investment performance, such as the cumulative return or Sharpe ratio. The former quantity is defined as the sum of the daily returns, that is $\sum_{t=2}^{T} f(x^\top r_{t-1})\, x^\top r_t$, where the function f can, for example, be $\mathrm{sign}(\cdot)$ or $-\mathrm{sign}(\cdot)$, depending on whether the portfolio follows a momentum or mean reversion strategy; we refer to [5] for background. The Sharpe ratio is defined as the ratio between the average daily returns and the standard deviation of the daily returns. Both quantities are difficult to optimize in x (e.g. they are not differentiable), therefore in this paper we use as a surrogate function the lag-1 autocorrelation of the portfolio. Intuitively, a positive (resp. negative) autocorrelated portfolio will

favour a momentum (resp. mean reversion) strategy: we buy (resp. sell) the portfolio on day t if it had a positive (resp. negative) return on the previous day.

2.2 Regularization Problem

The vector x gives rise to the scalar time series $p_t := x^\top r_t, t = 1, \ldots, T$, called the portfolio series. Our goal is to find a portfolio that has maximal lag-1 autocorrelation, which is defined as the correlation between p_{t-1} and p_t. If the series is stationary[1] this quantity simplifies to

$$\nu(x) = \frac{x^\top \Theta x}{x^\top \Gamma x} \tag{1}$$

where Θ and Γ are the lag-1 autocovariance and the covariance of the time series, respectively. The latter matrix is assumed to be invertible. In order to emphasize the fact that the autocorrelation ν depends on matrix Θ, we will sometimes use the notation $\nu(\cdot|\Theta)$. We then solve the problem

$$\max_{x \in \mathbb{R}^n} \nu(x). \tag{2}$$

This is a generalized eigenvalue problem [7], whose solution is given by $x = \Gamma^{-\frac{1}{2}} u$, where u is the leading eigenvector of the matrix $\Gamma^{-\frac{1}{2}} \Theta \Gamma^{-\frac{1}{2}}$.

In practice, matrices Θ and Γ are estimated from historical data, the most common estimates being $\hat{\Theta} = \frac{1}{T} \sum_{t=2}^{T} r_t r_{t-1}^\top$ and $\hat{\Gamma} = \frac{1}{T} \sum_{t=1}^{T} r_t r_t^\top$ (see e.g. [17]), where for simplicity we assumed that the series has mean zero. Often, these estimates are inaccurate and it can be therefore useful to introduce robust versions of problem (2), in which we suppose that the matrix Θ and/or Γ is known to belong to some uncertainty set. For simplicity here we only address the robustness of Θ, which is typically the main problem of concern since the covariance is in practice often more stable than the lag-1 autocovariance [5, 17]. In particular, we prescribe an *uncertainty set* $\mathcal{A}$ and consider the problem

$$\max_{x \in \mathbb{R}^n} \min_{\Theta \in \mathcal{A}} \nu(x|\Theta). \tag{3}$$

That is, we maximize the worst autocorrelation obtained when varying matrix Θ in the set $\mathcal{A}$.

A natural choice for the uncertainty set is a ball centered at the empirical estimate $\hat{\Theta}$, namely we choose $\mathcal{A} = \{\Theta : \|\Theta - \hat{\Theta}\|_p \leq \epsilon\}$, where $\|\cdot\|_p$ is the ℓ_p norm of the matrix elements.[2] In this case, it is not difficult to see that $\min\{x^\top \Theta x : \Theta \in \mathcal{A}\} = x^\top \hat{\Theta} x - \epsilon \|x\|_q^2$, where q is the conjugate exponent of p, as determined by the equation $1/p + 1/q = 1$; a detailed derivation is presented in Lemma 2 in the appendix. Using this observation, problem (3) becomes

$$\max_{x \in \mathbb{R}^n} \left\{ \frac{x^\top \hat{\Theta} x - \epsilon \|x\|_q^2}{x^\top \Gamma x} \right\}. \tag{4}$$

Parameter ϵ is related to the size (confidence level) of the uncertainty set $\mathcal{A}$, and, as we adopt the regularization framework below, for simplicity we call it hereafter the regularization parameter. In this paper we consider the cases $q = 2$ and $q = 1$. In the first case, problem (4) is still a generalized

[1] Relaxing this assumption within our framework is left for future research.

[2] Other matrix norms could be used, but we leave this for future research.

eigenvalue one of the form (1) with matrix Θ replaced by $\hat{\Theta} - \epsilon I$. In the second case, the problem becomes a nonlinear one, for which we present an optimization method in the next section.

We note that, in the case of financial data, when $\epsilon \to \infty$ the solutions of problem (4) are related to portfolios studied before in the Finance literature. Specifically, for $q = 2$ we recover the leading eigenvector of Γ, which for financial time series is close to the "market portfolio" [2]. If $q = 1$, and the series components all have the same unit variance, then the solution of problem (4) is (up to a nonzero multiplicative constant) given by $x = e_j$, where j is the most positively autocorrelated series component [14], that is $\Theta_{jj} = \max_{i=1}^{n} \Theta_{ii}$. To see this, note that problem (4) is equivalent to $\max\{x^\top \Theta x - \epsilon \|x\|_1^2 : x^\top \Gamma x = 1\}$. If ϵ is large enough there is an advantage in choosing x such that $\|x\|_1$ is as small as possible provided that $x^\top \Gamma x = 1$. Since we assumed that the series all have the same unit variance, we have that

$$1 = x^\top \Gamma x = \sum_{i=1}^{n} x_i^2 + \sum_{i \neq j} \Gamma_{ij} x_i x_j = \|x\|_1^2 + \sum_{i \neq j} |x_i x_j| \left(\Gamma_{ij} \text{sign}\,(x_i x_j) - 1\right).$$

Since $|\Gamma_{ij}| < 1$ if $i \neq j$ (otherwise Γ would not be strictly positive definite) we conclude that

$$\|x\|_1^2 = 1 + \sum_{i \neq j} |x_i x_j| \left(1 - \Gamma_{ij} \text{sign}\,(x_i x_j)\right) \geq 1$$

and $\|x\|_1 = 1$ is and only if $x \in \{e_1, \ldots, e_n\}$.

The above robust analysis can be applied in a similar way to the problem of finding the most negative autocorrelated portfolio, namely $\min\{\nu(x|\Theta) : x \in \mathbb{R}^n\}$. This merely requires replacing Θ by $-\Theta$ in problem (4). Moreover, the above analysis can be extended to take into account uncertainty in both matrices Θ and Γ, with not much additional difficulty. Specifically, if we choose the set $\mathcal{A} = \{\Theta : \|\hat{\Theta} - \Theta\|_p \leq \epsilon\} \times \{\Gamma : \|\hat{\Gamma} - \Gamma\|_p \leq \lambda\}$ then we obtain a robust optimization problem similar to (4) but with the denominator replaced by $x^\top \Gamma x + \lambda \|x\|_q^2$. As we have already noted, in financial time series Θ is much more unstable that Γ and, so, we do not explore robustness with respect to both Θ and Γ further in this paper. We also refer to [18] for related ideas on the ℓ_2 case.

We end this section by noting a connection between our approach and a method of unsupervised learning. Using the identity $2(x^\top r_{t-1})(x^\top r_t) = (x^\top r_{t-1})^2 + (x^\top r_t)^2 - (x^\top (r_t - r_{t-1}))^2$, we can rewrite twice the numerator in (1) as $x^\top \Gamma_{t-1} x + x^\top \Gamma_t x - x^\top V x$, where $V = \mathbb{E}\delta_t \delta_t^\top$ is the covariance of the "velocity" process $\delta_t = r_t - r_{t-1}$. If the process is stationary then $\nu(x) := 1 - \frac{1}{2}\psi(x)$, where

$$\psi(x) := \frac{x^\top V x}{x^\top \Gamma x}.$$

Thus, maximizing ν is the same as minimizing ψ. The latter optimization problem is very similar to the method of slow feature analysis [19] (see also [15]), an unsupervised learning technique which was originally designed to extract invariant representations from time varying visual signals.

3 Optimization Method

In this section, we address the issue of implementing the learning method (4) in the case $q = 1$.[3] We begin by rewriting problem (4) as $\min\{\eta(x) : x \in \mathbb{R}^d\}$, where

$$\eta(x) = \frac{x^\top N x + \epsilon \|x\|_1^2 - x^\top P x}{x^\top \Gamma x}$$

[3] Similar observations apply to the general case $q \in (1, \infty]$.

Algorithm 1 ℓ_1-Regularized Autocorrelation

Choose a starting point $x^0 \in \mathbb{R}^n$ and tolerance parameter tol.
for $k = 1, 2, \ldots$ **do**
  Let $x^k = \operatorname{argmin}\{\phi(x|x^{k-1}) : x \in \mathbb{R}^n\}$
  If $|\eta(x^k) - \eta(x^{k-1})| \leq tol$ terminate.
end for

and P, N are symmetric positive definite matrices such that $P - N = S := (\Theta + \Theta^\top)/2$. Matrices P and N can be obtained via the eigenvalue decomposition of S as $P = (S)_+$ and $N = P - S$, where $(\cdot)_+$ is a spectral function that acts on the eigenvalues $\lambda \in \mathbb{R}$ as $(\lambda)_+ = \max(\lambda, 0)$.

Fix a point x^0 and let $\phi(\cdot|x^0)$ be the function, defined, for every $x \in \mathbb{R}^n$, as

$$\phi(x|x^0) = x^\top N x + \epsilon\|x\|_1^2 - (x^0)^\top P x^0 - 2(x - x^0)^\top P x^0 - \eta(x^0)\gamma(x|x^0) \tag{5}$$

where

$$\gamma(x|x^0) = \begin{cases} (x^0)^\top \Gamma x^0 + 2(x - x^0)\Gamma x^0 & \text{if } \eta(x^0) > 0, \\ x^\top \Gamma x & \text{otherwise.} \end{cases}$$

We then consider the convex optimization problem

$$\min_x \phi(x|x^0). \tag{6}$$

The following lemma provides a rational behind this problem.

Lemma 1. *If* $\phi(x|x^0) < 0$ *then* $\eta(x) < \eta(x^0)$.

Proof. The result follows from the inequality $x^\top N x + \epsilon\|x\|_1^2 - x^\top P x - \eta(x^0) x^\top \Gamma x \leq \phi(x|x^0)$. □

This observation leads to the descent Algorithm 1, which is inspired by a method outlined in [11] for sparse PCA. Algorithm 1 iteratively solves a sequence of problems of the form (6). Lemma 1 guarantees that the algorithm produces a sequence of points $\{x^k : k \in \mathbb{N}\}$ such that the corresponding sequence of function values $\{\eta(x^k) : k \in \mathbb{N}\}$ is strictly monotonically decreasing or the algorithm terminates. In practice, we terminate the algorithm when $|\eta(x^k) - \eta(x^{k-1})|$ is less than some tolerance parameter, e.g. 10^{-4}.

It remains to show how to solve problem (6). This problem is of the form $\min \|Ax - b\|_2^2 + \epsilon\|x\|_1^2$, for an appropriate choice of the $n \times n$ matrix A and vector $b \in \mathbb{R}^n$. Hence, it is equivalent to the Lasso method and can be solved up to numerical precision by proximal gradient methods, see e.g. [3, 16]. In our numerical experiments below we have found that we do not need to solve (6) exactly, it is enough to find a point x which strictly decreases the objective. The simplest choice for the update rule is to set

$$x^k = \operatorname{prox}_{r\|\cdot\|_1^2}\left[\left(I + \frac{1}{\alpha^{k-1}}\big(S + \eta(x^{k-1})\Gamma\big)\right)x^{k-1}\right] \tag{7}$$

where $\alpha^{k-1} = |||N|||$ if $\eta(x^{k-1}) > 0$ and $\alpha^{k-1} = |||N - \eta(x^{k-1})\Gamma|||$ otherwise, with $|||\cdot|||$ the spectral norm. This corresponds to a single step of the proximal gradient method [16]. The function $\operatorname{prox}_{r\|\cdot\|_1^2}$

Algorithm 2 Robust CCA

Choose a starting point $z^0 \in \mathbb{R}^n$ and tolerance parameter tol.
for $k = 1, 2, \ldots$ **do**
  Let $(\hat{x}, \hat{y}) = \hat{z}$ where $\hat{z} = \operatorname{argmin}\{\phi(z|z^{k-1}) : z \in \mathbb{R}^n\}$
  Set $z^k = (x^k, y^k)$ with $x^k = \hat{x}/\sqrt{\hat{x}^\top \Gamma \hat{x}}$, $y^k = \hat{y}/\sqrt{\hat{y}^\top \Sigma \hat{y}}$.
  If $|\eta(z^k) - \eta(z^{k-1})| \leq tol$ terminate.
end for

is the proximity operator of the function $r\|\cdot\|_1^2$, where $r := \frac{\epsilon}{2\alpha^{k-1}}$ and it is defined, for every $z \in \mathbb{R}^n$, as

$$\operatorname{prox}_{r\|\cdot\|_1^2}(z) = \operatorname{argmin}\left\{\frac{1}{2}\|x - z\|_2^2 + r\|x\|_1^2 : x \in \mathbb{R}^n\right\}. \tag{8}$$

Appendix B presents a method to solve this problem.

4 Robust Canonical Correlation Analysis

In this section, we present an extension of the proposed approach to the problem of maximizing the correlation between the one-dimensional projections of the time series r_t and a prescribed m-dimensional signal series, denoted by s_t. For every pair of vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, we define the cross-correlation function

$$\rho(x, y) = \frac{x^\top \Theta y}{\sqrt{x^\top \Gamma x}\sqrt{y^\top \Sigma y}} \tag{9}$$

where Γ and Σ are estimates of the covariance of r_t and s_t respectively (both assumed to be invertible) and the $n \times m$ matrix Θ is an estimate of the cross-covariance of r_t and s_t. The simplest example of this setting is when $s_t = r_{t-1}$, but in general s_t may include for example technical indicators/more lags or any other time series (other than stocks in our case). Notice also that if $s_t = r_{t-1}$ and $x = y$ we recover problem (1).

Our goal is to solve the problem

$$\max_{x,y} \rho(x, y) = \max_{x,y} \frac{2x^\top \Theta y}{x^\top \Gamma x + y^\top \Sigma y} \tag{10}$$

where the equality follows by the arithmetic-geometric mean inequality and the fact that the solutions are invariant by rescaling, see e.g. [1, 10]. The right problem is a generalized eigenvalue problem of the form $\max\{z^\top A z : z \in \mathbb{R}^{n+m},\ z^\top B z = 1\}$, where

$$z = \begin{pmatrix} x \\ y \end{pmatrix}, \quad A = \begin{bmatrix} 0 & \Theta \\ \Theta^\top & 0 \end{bmatrix}, \quad B = \begin{bmatrix} \Gamma & 0 \\ 0 & \Sigma \end{bmatrix}.$$

We can now have an extension of problem (3) to the case of CCA where again we consider robustness w.r.t. perturbations of matrix Θ, that is we consider the problem

$$\max_{x,y} \min_{\Theta \in \mathcal{A}} \rho(x, y|\Theta) \tag{11}$$

for $\mathcal{A} = \{\Theta : \|\Theta - \hat{\Theta}\|_p \leq \epsilon\}$. Using Lemma 2 in the appendix to compute the inner minimum in (11), we obtain the problem

$$\max_{x,y} \frac{x^\top \hat{\Theta} y - \epsilon \|x\|_q \|y\|_q}{\sqrt{x^\top \Gamma x}\sqrt{y^\top \Sigma y}}. \tag{12}$$

Two choices of interest are $q = 2$ and $q = 1$. In the latter case we obtain again an interpretation of the problem for large values of ϵ, which is equivalent to selecting the pair of most positively correlated series, studied in [14]. This analysis requires that Γ and Σ have all their diagonal elements equal to one. Indeed, if ϵ is sufficiently large the second term in the numerator of (12) dominates, hence we want this term to be as small as possible. Thus, it must be the case that $x \in \{e_1, \ldots, e_n\}$ and $y \in \{e_1, \ldots, e_m\}$. We conclude that the solution of (12) is given by $x = e_j$, $y = e_k$ such that $\hat{\Theta}_{jk} = \max\{\hat{\Theta}_{i\ell} : i = 1, \ldots, n,\ \ell = 1, \ldots, m\}$. A similar reasoning applies for the most negatively autocorrelated series pair. This is obtained by replacing Θ by $-\Theta$ in the above analysis.

In Algorithm 2 we sketch a method to find a solution of problem (12). The algorithm, which is similar in spirit to Algorithm 1, starts from a vector z^0 and iteratively decreases the objective by solving the convex optimization problem

$$\phi\left(z|z^k\right) = z^\top \left(N + |\eta\left(z^k\right)|B\right) z - \left(z - z^k\right)^\top \left(2Pz^k + \epsilon u + 1_{\{\eta(z^k)>0\}} v\right)$$

where $P = (A)_+$, $N = P - A$, $u \in \partial \|z^k\|_{2,q}^2$ and $v \in \partial\left(\|\Gamma^{\frac{1}{2}} x^k\|_2 + \|\Sigma^{\frac{1}{2}} y^k\|_2\right)^2$. Similarly to Lemma 1 one can show that function $\phi(\cdot|z^k)$ has the property that if $\phi(z|z^k) < 0$ then $\eta(z) < \eta(z^k)$. Furthermore this function can be optimized by means of proximal gradient methods, see e.g. [3, 16]. A detailed description of the algorithm in presented in Appendix C.

5 Experiments

In this section, we present experiments with the proposed methods on a synthetic and a real dataset. As the main goal of this paper is to study the use of robust optimization through the development of regularization methods for time series prediction, a key aim of the experiments is to explore whether robustness — and the corresponding regularization — has the potential to improve test performance when the data is highly noisy, such as in the case of financial data. Under these circumstances, one would expect that as we "add robustness" by increasing the regularization parameter ϵ, the test performance of the methods improves and, potentially, drops after some point.

5.1 Synthetic Data

We used synthetic data to test the ℓ_1 method above. We generated 30 synthetic time series of length 200. Among these, three were governed by a stationary AR(1) process [17], whereas the remaining 27 series were white noise (normally distributed with zero mean and unit variance). The AR(1) time series were centered and normalized in order to have the same mean and variance as the remaining 27 time series. We trained Algorithm 1 using this dataset for 100 values of the regularization parameter ϵ. The left plot in Figure 1 shows the weights of the portfolio found by Algorithm 1 as a function of ϵ. It can be seen that for the portfolios of sparsity less than or equal to 3, the AR(1) time series were heavily favoured by our method. While the noise time series had non-zero weights for small values of ϵ, these weights were usually smaller than the weights of the AR(1) time series for all values of ϵ, and, as ϵ increased, they tended to be suppressed to zero earlier than the weights of the AR(1) time

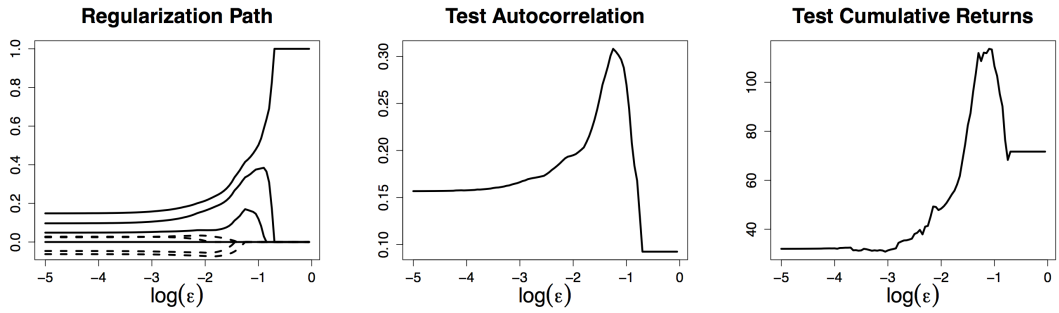

Figure 1: Synthetic Data: Solution path (Left), Test Autocorrelation (Center) and Test Cumulative Return (Right), as a function of the regularization parameter ϵ, for the ℓ_1-Method.

series. The next two plots in Figure 1 show the performance of the portfolios on 800 consecutive test data points, again as a function of ϵ. Both test performances are maximized at approximately the same value of ϵ in this case.

5.2 S&P 500 Stock Data

Next, we tested the methods using a dataset that is known to be notoriously challenging: we focused on the problem of constructing portfolios x of n stocks using daily adjusted close prices of stocks in the S&P 500 index.[4] We used data for the past 10 years, from January 1, 2003 until April 12, 2013 (the date of the final data construction). This corresponds to a total of 2586 daily (close to close) returns. We tested the methods by constructing portfolios using only stocks from specific sectors, for a number of different sectors. This was done both in order to perform multiple experiments and because companies in the same sector are known to "co-move" [2], making the proposed methods more applicable. We consider four large sectors defined based on a standard industry classification:[5] Energy, Financial, Healthcare, and Technology (other sectors led to similar conclusions). These consist of respectively $n = 30, 30, 24$ and 35 stocks. For each sector we use the first 1000 days for training, the next 250 days for validation, and the remaining 1336 days for testing.[6] The values of the regularization parameter ϵ considered were $\{0, 0.1^k, 10^{10}\}$ for k between -6 to 6 at increments of 0.005 (hence a total of 2403 values) for the ℓ_2 method, and for k between 0.05 and 5 at increments of 0.05 (hence a total of 100 values) for the ℓ_1 method, which led to "complete U-curves" below for both methods. We used fewer values for the ℓ_1 method as it is a computationally more costly one. We report the following performance metrics:

- *Cumulative return*: the sum of the 1336 daily returns of the constructed portfolios. It corresponds to the cumulative returns one would get if one had invested using the method over the period's last 1336 days, investing "one unit" every day.

[4]The data (downloaded from Yahoo!) and the R code used to run the experiments are available from the authors upon request.

[5]Based on the industry classification at *http://www.nasdaq.com/screening/industries.aspx*, sorted by market capitalization, and using only those companies with market capitalization larger than $10 billion in April 2013.

[6]Although the conclusions are similar for other data splits, the problem (discussed below) that the selected parameter ϵ changes across windows - for some of which the performances of the different portfolios tested are similar - makes the analysis of the "average across windows" effects of ϵ on performance (figures below) noisy. To better present this effect we only report the results for one data split.

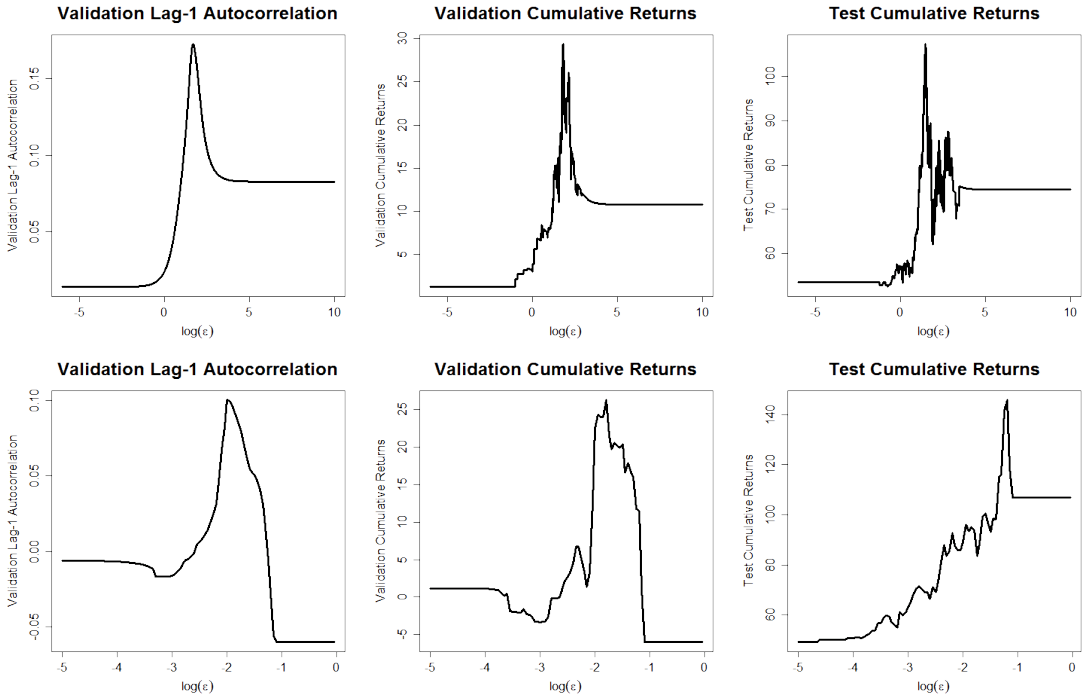

Figure 2: S&P 500 Stock Data: Validation ν, Validation Cumulative Return and Test Cumulative Return over 1336 days for the ℓ_2-Method (top) and ℓ_1-Method (bottom) as functions of the regularization parameter ϵ.

- *Yearly Sharpe ratio:* the ratio of the average daily return over the standard deviation of the returns in the 1336 days. We scale it by a factor of $\sqrt{250}$ to get an annual Sharpe ratio as typically reported in practice, e.g. [17]. Note that, as we have 1336 test data, any Sharpe ratio larger than $\frac{1.96}{\sqrt{1336/250}} = 0.85$ is statistically significant.

We report the performances of the two methods (ℓ_1 and ℓ_2) using both the regularization parameter ϵ selected using the cumulative returns in the validation data, and the optimal regularization parameter based on the test data (hence with hindsight). The latter indicates potential space for improvement, as discussed below. We also report three benchmarks: a) the performance of the average of the stocks (the "Market") during the same test window, typically used as a benchmark in practice; b) the performance of the portfolio x with equal weights (e.g. $1/n$) for all stocks, corresponding to a "market autocorrelation portfolio" ("Market AC" in Table 1); c) the non-regularization based solution (i.e. $\epsilon = 0$) ("Max AC" in Table 1). For comparison in all cases we normalize the solution x so that it has ℓ_1 norm 1 (hence in all cases "we invest 1 unit every day").

Figure 2 shows the validation (minus) autocorrelaton, the validation cumulative return and the test cumulative return as a function of the regularization parameter ϵ, both for the case of the ℓ_2 (top row) and ℓ_1 (bottom row) methods. The plots shown are for mean reversion, hence most negatively autocorrelated portfolio, for the healthcare sector; similar conclusions can be drawn from plots for the other sectors and for momentum/most positively autocorrelated portfolio. The figure illustrates the main experimental finding: using the proposed robust optimization approach improves performance and, more interestingly, the observed inverted U-curve indicates that the proposed methodologies capture "structure" even in the highly unpredictable S&P 500 daily stock returns time series.

Method	Healthcare	Financials	Energy	Technology
Market	37.5% (0.47)	2.0% (0.01)	7.8% (0.04)	28.1% (0.19)
Market AC	57.2% (0.47)	353.3% (1.60)	66.25% (0.31)	22.44% (0.15)
Max AC	53.6% (1.52)	115.7% (1.80)	-3.6% (-0.12)	5.1% (0.16)
Selected L2	85.8% (0.77)	300.8 % (1.36)	34.4% (0.17)	**52.7%** (0.34)
Best L2	107.3% (2.13)	346.3% (1.57)	67.7% (0.30)	88.56% (0.71)
Selected L1	**94.0%** (1.02)	**439.6%** (1.43)	**161.1%** (1.03)	-24.7% (-0.47)
Best L1	145.7% (1.04)	446.9% (1.48)	162.9% (0.88)	281.3% (0.83)

Table 1: Comparison of methods for the S&P 500 Stock Data: for each of the 4 sectors we note with "Market" the average of the stocks in that sector, with "Market AC" the selected between momentum and mean reversion of the "Market", with "Max AC" the non-regularized maximum autocorrelation solution ($\epsilon = 0$), with "Selected" the regularization methods with the selected regularization parameter ϵ using the validation data, and with "Best" the best ϵ using the test data (hence with hindsight). Cumulative returns during 1336 test days are reported, with annual Sharpe ratio (values above 0.85 are statistically significant – see text) in parentheses. Best performance for each sector – without considering the "Best" cases – is indicated in bold.

Using the validation data, we choose between momentum (maximum positive autocorrelation) and mean reversion (minimum negative autocorrelation), and select the regularization parameter for the proposed methods. We report the performances in Table 1. For each case we report the cumulative return and the Sharpe ratio (in parentheses) in the test data. From the values in Table 1 we can make the following observations.

1. For all sectors, the proposed approach leads to portfolios that outperform the sector's market. Given that in practice outperforming the market, particularly for stocks in the S&P 500 index, is considered challenging, the results indicate the potential of the proposed approach.

2. For all sectors, regularization improves performance relative to the case $\epsilon = 0$.

3. For all sectors the best ϵ with hindsight is, as expected, (much) better than the performance of the selected ϵ. This indicates that there can be further improvements if a better method to select ϵ for this data is developed. Note that Figure 2 also illustrates this challenge of selecting ϵ for the specific financial data: the best performing regularization parameter ϵ for the validation data may differ from that in the test data. Although the "inverse u-curves" are observed for different time windows indicating that the proposed methods capture structure in this data, selecting the regularization parameter in a "rolling window" setup can be a challenge in practice as the "u-curve" may shift across time windows (e.g. this structure may be non-stationary).

4. The largest performance improvement is for the financial sector. This indicates that it may be the case that the proposed methods work better for certain groups of time series/stocks. Future work can improve our understanding of the characteristics of such groups of time series and/or lead to other methods for different types of groups of time series.

6 Conclusion and Future Research

We proposed an approach to estimate large autocorrelation portfolios using regularization methods derived from a robust optimization formulation. We developed two regularization methods as special cases of the proposed general approach. For one of these methods we developed an iterative optimization learning algorithm which estimates sparse portfolios. We then tested the methods using notoriously noisy financial time series data. The experiments indicate the potential of the proposed approach to uncover structure in time series of daily S&P 500 stock returns. The results also indicate that the proposed method can lead to portfolios that outperform "the market" for this data. Finally, we discussed an extension and a novel algorithm for the more general case of CCA.

A number of future research directions can further improve the proposed approach. One of the key questions is the selection of the regularization parameter for (non-stationary – among others) time series such as the stock data we explored. Another question may be to build on the proposed methods in order to better identify subsets of time series for which the approach performs best. Yet another direction for research is to further develop the proposed CCA approach and test it using potentially diverse "predictors" e.g. for the financial time series explored. Finally, potentially novel regularization methodologies for time series analysis can be developed based on the robust optimization approach used in this paper.

Acknowledgements

We would like to thank for suggestions Julien Bonhé, Christoph Burgard, Matthias Hein, Raphael Hauser, Gunnar Klinkhammer, Nick Leyhane, Andreas Maurer, Charles Micchelli, Luke Pebody, Ioana Popescu, Bin Ren, Lucie Tepla, Dimitri Vayanos and Ding-Xuan Zhou. A special thanks to Andreas Argyriou for many useful discussions. This work was in part supported by EPSRC grant EP/H027203/1.

References

[1] T.M. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, second edition, 1984.

[2] M. Avellaneda and J.H. Lee. Statistical arbitrage in the U.S. equities market. *Quantitative Finance*, 10:761-782, 2010.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal of Imaging Sciences*, 2(1):183–202, 2009.

[4] A. Ben-Tal, L. El Ghaoui, A. Nemirovski. *Robust Optimization*, Princeton University Press, 2009.

[5] J.Y. Campbell, A.W.C. Lo, A.C. MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, 1997.

[6] G. Cornuejols and R. Tütüncü. *Optimization Methods in Finance.* Cambridge University Press, 2011.

[7] G.H. Golub and C.F. Van Loan. *Matrix Computations.* John Hopkins University Press, 1996.

[8] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[9] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(6):1005–1018, 2007.

[10] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[11] M. Hein and T. Buehler. An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA. *Advances in Neural Information Processing Systems 23*, pages 847–855, 2010.

[12] M. Hein and S. Setzer, Beyond spectral clustering - tight relaxations of balanced graph cuts. *Advances in Neural Information Processing Systems 24*, pages 2366–2374, 2011.

[13] J-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms, Part I.* Springer, 1996.

[14] A.W. Lo and A.C. MacKinlay. When are contrarian profits due to stock market overreaction? *Review of Financial studies*, 1990.

[15] A. Maurer. Unsupervised slow subspace-learning from stationary processes. *Theoretical Computer Science*, 405(3):237–255, 2008.

[16] Y. Nesterov. Gradient methods for minimizing composite objective functions. *ECORE Discussion Paper*, 2007/96, 2007.

[17] R.S. Tsay. *Analysis of Financial Time Series*. John Wiley & Sons, 2002.

[18] P. Xanthopoulos, M.R. Guarracino, P.M. Pardalos. Robust generalized eigenvalue classifier with ellipsoidal uncertainty. *Annals of Operations Research*, to appear.

[19] L. Wiskott and T.J. Sejnowski. Slow feature analysis: unsupervised learning of invariances. *Neural Computation*, 14:715–770, 2002.

[20] D.M. Witten, R. Tibshirani, T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

Appendix

In this appendix, we collect some auxiliary results and present the main steps for the solution of the robust CCA problem.

A Derivation of the Robust Optimization Problem

The following lemma is the key step in the derivation of problems (4) and (12).

Lemma 2. *Let $\hat{\Theta} \in \mathbb{R}^{n\times m}$, let $\|\cdot\|_p$ be the elementwise ℓ_p-norm of a matrix or vector, where $p \in [1,\infty]$, and let q satisfy $1/p + 1/q = 1$. Then it holds that*

$$\min\{x^\top \Theta y : \Theta \in \mathbb{R}^{n\times m},\ \|\Theta - \hat{\Theta}\|_p \leq \epsilon\} = x^\top \hat{\Theta} y - \epsilon \|x\|_q \|y\|_q.$$

Proof. We write $\Theta = \hat{\Theta} + \Delta$ for some $\|\Delta\|_p \leq \epsilon$. Using Hölder's inequality we obtain that

$$x^\top \Theta y = x^\top \hat{\Theta} y + x^\top \Delta y \geq x^\top \hat{\Theta} y - \|\Delta\|_p \|xy^\top\|_q.$$

We shall show that this inequality is tight for a particular choice of Δ. To this end, let $\delta \in \mathbb{R}^n$ such that $\|\delta\|_p = 1$ and $\delta^\top x = \|x\|_q$, and let $\gamma \in \mathbb{R}^m$ such $\|\gamma\|_p = 1$ and $\gamma^\top y = \|y\|_q$. The result then follows by choosing $\Delta = \epsilon \delta \gamma^\top$. □

B Computation of the Proximity Operator

We describe a method to compute the proximity operator of the function $\|\cdot\|_1^2$, see equation (8). To this end, we recall the following identity[7]

$$\|x\|_1^2 = \inf\left\{ \sum_{i=1}^n \frac{x_i^2}{\lambda_i} : \lambda > 0, \sum_{i=1}^n \lambda_i = 1 \right\}$$

where $\lambda \in \mathbb{R}^n$ denotes the vector $(\lambda_1, \ldots, \lambda_n)$ and $\lambda > 0$ means that all components of λ must be greater than zero. Replacing the above expression in the right hand side of (8), fixing λ and minimizing over x we obtain the solution

$$x_i(\lambda) = \frac{\lambda_i z_i}{2r + \lambda_i}. \tag{13}$$

Using this equation, we obtain the problem

$$\min\left\{ \sum_{i=1}^n \frac{r z_i^2}{2r + \lambda_i} : \lambda \geq 0, \sum_{i=1}^n \lambda_i = 1 \right\}.$$

One verifies that the minimizing λ is given by

$$\lambda_i = (\rho |z_i| - 2r)_+$$

where the positive parameter ρ is found by binary search in order to ensure that $\|\lambda\|_1 = 1$. Finally, we replace the obtained value of λ in the right hand side of (13) to obtain the solution of (8).

[7] This is a direct consequence of the arithmetic-geometric mean inequality.

C Robust CCA Algorithm

We describe the main steps behind Algorithm 2. If f is a convex function we denote by $\mathrm{Lin}_{f,u}(z|z^0) = f(z^0) + u^\top(z - z^0)$ a linear approximation of f at z^0, for some $u \in \partial f(z^0)$. We sometimes omit u and write $\mathrm{Lin}_f(z|z^0)$ to denote a generic linear approximation. This approximation can be visualized as a linear lower bound for f which touches f at z^0.

We first rewrite problem (12) as that of minimizing the quantity

$$\eta(x, y) = \frac{-x^\top \Theta y + \epsilon \|x\|_q \|y\|_q}{\sqrt{x^\top \Gamma x}\sqrt{y^\top \Sigma y}} \tag{14}$$

over $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Recall the notation

$$z = \begin{pmatrix} x \\ y \end{pmatrix}, \quad A = \begin{bmatrix} 0 & \Theta \\ \Theta^\top & 0 \end{bmatrix}, \quad B = \begin{bmatrix} \Gamma & 0 \\ 0 & \Sigma \end{bmatrix}$$

and note that $-2x^\top \Theta y = z^\top A z$. Using the equality $2\|x\|_q\|y\|_q = (\|x\|_q + \|y\|_q)^2 - \|x\|_q^2 - \|y\|_q^2$ and the formula $A = P - N$, for $P = (A)_+$ and $N = P - A$, twice the numerator in (14) has the DC decomposition (difference of convex functions)

$$R_1(z) - R_2(z) = \left(z^\top N z + \epsilon \|z\|_{1,q}^2\right) - \left(z^\top P z + \epsilon \|z\|_{2,q}^2\right)$$

where we defined the mixed norms $\|z\|_{1,q} = \|x\|_q + \|y\|_q$ and $\|z\|_{2,q} = \sqrt{\|x\|_q^2 + \|y\|_q^2}$. Using this formula, problem (12) can be rewritten as

$$\min_z \left\{ \frac{z^\top N z + \epsilon \|z\|_{1,q}^2 - z^\top P z - \epsilon \|z\|_{2,q}^2}{2\sqrt{x^\top \Gamma x}\sqrt{y^\top \Sigma y}} \right\}. \tag{15}$$

For every vector $z^0 \in \mathbb{R}^{n+m}$ we shall construct a function $\phi(\cdot|z^0)$ which has the property that if $\phi(z|z^0) < 0$ then $\eta(z) < \eta(z^0)$. We distinguish between two cases.

Case 1: $\eta(z^0) \leq 0$. We replace the denominator in (15) by the quadratic form $z^\top B z$, which provides a simplification of the problem. Indeed if the objective is negative by the arithmetic-geometric mean inequality we obtain that

$$\eta(z) \leq h(z) := \frac{R(z)}{z^\top B z}$$

and equality holds if and only if $x^\top \Gamma x = y^\top \Sigma y$.

Fix $z^0 \in \mathbb{R}^{n+m}$ and let $\phi(\cdot|z^0)$ be the convex function defined, for every $z \in \mathbb{R}^{n+m}$, as

$$\phi(z|z^0) = R_1(z) - \mathrm{Lin}_{R_2}(z, z^0) - \eta(z^0) z^\top B z. \tag{16}$$

Note that $\phi(z^0|z^0) = 0$. If z is a point such that $\phi(z|z^0) < 0$, we conclude that $h(z) < h(z^0)$. Indeed

$$0 > \phi(z|z^0) \geq R_1(z) - R_2(z) - h(z^0) z^\top B z.$$

Furthermore, if we rescale z so that $x^\top \Gamma x = y^\top \Sigma y$, then $\eta(z) = h(z)$.

Case 2: $\eta(z^0) > 0$. In this case we need to work with the original denominator, $\sqrt{x^\top \Gamma x}\sqrt{y^\top \Sigma y}$. We rewrite twice this quantity as the DC decomposition

$$S_1(z) - S_2(z) = \left(\|\Gamma^{\frac{1}{2}} x\|_2 + \|\Sigma^{\frac{1}{2}} y\|_2\right)^2 - x^\top \Gamma x - y^\top \Sigma y.$$

Now, we choose

$$\phi(z|z^0) = R_1(z) - \mathrm{Lin}_{R_2}(z, z^0) - \eta(z^0)\left(\mathrm{Lin}_{S_1}(z|z^0) - S_2(z)\right). \tag{17}$$

This case is slightly more difficult to handle since we also need to compute a subgradient of the function S_1. Combining [13, Thms 4.2.1 and 4.3.1], we obtain that

$$\partial\left(\|\Gamma^{\frac{1}{2}}x\|_2 + \|\Sigma^{\frac{1}{2}}y\|_2\right)^2 = 2(\|\Gamma^{\frac{1}{2}}x\|_2 + \|\Sigma^{\frac{1}{2}}y\|_2)\Big\{(\Gamma^{\frac{1}{2}}\alpha, \Sigma^{\frac{1}{2}}\beta) : \alpha \in \partial\|\Gamma^{\frac{1}{2}}x\|_2, \beta \in \partial\|\Sigma^{\frac{1}{2}}y\|_2\Big\}.$$

We are now ready to summarize the formula for $\phi(z|z^0)$. Combining equations (16) and (17) we have

$$\phi(z|z^0) = z^\top(N + |\eta(z^0)|B)z - (z - z^0)^\top(2Pz^0 + \epsilon u + 1_{\{\eta(z^0)>0\}}v)$$

where $u \in \partial\|z^0\|_{2,q}^2$ and $v \in \partial S_1(z^0)$.

Algorithm 2 solves a sequence of convex optimization problems of the form

$$\min\{\phi(z|z^k) : z \in \mathbb{R}^{n+m}\}$$

In practice it is sufficient to solve this problem approximately, finding a point z such that $\phi(z|z^k) < 0$. Similarly to Lemma 1 it is easily seen that if $\phi(z|z^k) < 0$ then $\eta(z) < \eta(z^0)$. The simplest updating rule is provided by one step of the proximal gradient method [3, 16]

$$z^{k+1} = \mathrm{prox}_{r\|\cdot\|_{1,q}^2}\left[z^k - \frac{1}{|||N + |\eta^k|B|||}\Big((A + |\eta^k|B)z^k - \frac{1}{2}\left(\epsilon u + (\eta^k)_+ v\right)\Big)\right] \tag{18}$$

where, recall, $|||\cdot|||$ denotes the spectral norm of a matrix, $r = \frac{\epsilon}{2|||N+|\eta(z^0)|B|||}$, and we have defined $\eta^k = \eta(z^k)$.

Next, we discuss how to compute a subgradient of $\|z\|_{2,q}^2$ and the proximity operator when $q \in \{1, 2\}$. For this purpose, we recall that if $\|\cdot\|$ is a norm, then by [13, §4.3] the subdifferential of $\|\cdot\|^2$ at z is equal to $2\|z\|$ times the subdifferential of $\|\cdot\|$ at z, that is $\partial\|z\|^2 = 2\|z\|\partial\|z\|$. In particular, we obtain that

$$\partial\|z\|_{2,1}^2 = \{2(\|x\|_1\alpha, \|y\|_1\beta) : \alpha \in \partial\|x\|_1,\ \beta \in \partial\|y\|_1\}.$$

On the other hand, if $q = 2$ then $\|z\|_{2,2}^2$ is just the square ℓ_2 norm of z and its gradient is equal to $2z$.

It remains to obtain the formula for the proximity operator in (18). The case $q = 1$ is conceptually identical to the derivation in Appendix B, with the understanding that n is replaced by $n + m$ and x by the vector z. The case $q = 2$ is derived along the same lines and we only sketch the main points here. We obtain, for every $z = (x, y) \in \mathbb{R}^{n+m}$, that

$$\mathrm{prox}_{r\|z\|_{1,2}^2}(z) = \left(\frac{\lambda x}{2r + \lambda}, \frac{\tau y}{2r + \tau}\right)$$

where $\lambda = (\rho\|x\|_2 - 2r)_+$, $\tau = (\rho\|y\|_2 - 2r)_+$ and the positive parameter ρ is determined by binary search in order to ensure that $\lambda + \tau = 1$.

Europe Campus
Boulevard de Constance
77305 Fontainebleau Cedex, France
Tel: +33 (0)1 60 72 40 00
Fax: +33 (0)1 60 74 55 00/01

Asia Campus
1 Ayer Rajah Avenue, Singapore 138676
Tel: +65 67 99 53 88
Fax: +65 67 99 53 99

Abu Dhabi Campus
Muroor Road - Street No 4
P.O. Box 48049
Abu Dhabi, United Arab Emirates
Tel: +971 2 651 5200
Fax: +971 2 443 9461

www.insead.edu

INSEAD

The Business School
for the World®

Printed by INSEAD