Automatica 111 (2020) 108651

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

Sparse mean-reverting portfolios via penalized likelihood optimization☆

Jize Zhang, Tim Leung, Aleksandr Aravkin *

Department of Applied Mathematics, University of Washington, USA

ARTICLE INFO

Article history:
Received 2 November 2018
Received in revised form 23 July 2019
Accepted 25 September 2019
Available online 22 October 2019

ABSTRACT

An optimization approach is proposed to construct sparse portfolios with mean-reverting price behaviors. Our objectives are threefold: (i) design a multi-asset long-short portfolio that best fits an Ornstein–Uhlenbeck process in terms of maximum likelihood, (ii) select portfolios with desirable characteristics of high mean reversion through penalization, and (iii) select a parsimonious portfolio using ℓ_0-regularization, i.e. find a small subset of a larger universe of assets that can be used for long and short positions. We present the full problem formulation, and develop a provably convergent algorithm for the nonsmooth, nonconvex objective based on partial minimization and projection. We demonstrate model functionalities on simulated and empirical price data, and include comparison with a pairs trading algorithm.

1. Introduction

Mean reversion trading is a major class of trading strategies used by professional traders and fund managers. The strategy typically involves a portfolio of positions in two or more highly cointegrated assets (with a strong financial or economic relationship that prevents them from diverging), such as stocks and exchange-traded funds (ETFs), or derivatives, such as futures, across many asset classes. The challenge is to systematically construct a portfolio whose value over time exhibits mean-reverting behaviors. Once such a portfolio is identified, then the pattern can be exploited by traders and the estimated parameters can inform the optimal trading strategies, such as those developed in Leung and Li (2016). There are also a number of studies on trading mean-reverting prices (Kitapbayev & Leung, 2017; Leung & Li, 2015) and the empirical performance of pairs trading (Gatev, Goetzmann, & Rouwenhorst, 2006).

Previous works on mean-reverting portfolio design have used different empirical proxies for mean reversion and can usually be converted into semi-definite programming problems (see e.g. d'Aspremont, 2011; Zhao & Palomar, 2018; Zhao, Zhou, & Palomar, 2019). In this paper, we instead consider using an Ornstein–Uhlenbeck (OU) process (Ornstein & Uhlenbeck, 1930) as a measure for mean reversion. Given an arbitrary set of assets with their price histories, our main goal is to design a mean-reverting portfolio whose evolution over time can be characterized by an OU process (Ornstein & Uhlenbeck, 1930) through penalized maximum likelihood estimation (MLE). A major feature of our joint optimization approach is that we *simultaneously* solve for the optimal portfolio and the corresponding parameters for maximum likelihood using *gradient-based method*. This unified approach is different from prior work since (a) it does not rely on SDP, and (b) we do not break the problem up into stage-wise computations. For example, d'Aspremont (2011) first determines optimal weights using mean-reversion proxies other than OU, and then fits the resulting portfolio to an OU process. Since our formulation is based on MLE of OU and does not involve other proxies, it is more natural in our case to perform simultaneous optimization than a two-stage procedure. Conversely, Leung and Li (2016) fit an OU process to each of a range of candidate (pair) portfolios, and takes the candidate with the highest OU likelihood. Our unified approach looks for the best OU-representable portfolio from a set of candidates, making the quality of the OU fit part of the optimization problem.

This paper is a revised and expanded version of the short proceedings paper (Zhang, Leung, & Aravkin, 2018). In particular, the current paper (1) develops a new efficient projection onto the intersection of ℓ_0 level sets and the nonconvex set $\|x\|_1 = 1$ (Lemma 1), (2) establishes differentiability of value function used in the approach (Section 3.1), (3) proves convergence of the proposed algorithm (Theorem 2), and (4) presents numerical

☆ This research was partially supported by the Washington Research Foundation Data Science Professorship, USA. The material in this paper was presented at the 57th IEEE Conference on Decision and Control, December 17–19, 2018, Miami Beach, Florida, USA. This paper was recommended for publication in revised form by Associate Editor Kok Lay Teo under the direction of Editor Ian R. Petersen.

* Corresponding author.
E-mail addresses: jizez@uw.edu (J. Zhang), timleung@uw.edu (T. Leung), saravkin@uw.edu (A. Aravkin).

https://doi.org/10.1016/j.automatica.2019.108651
0005-1098/

examples with empirical prices (Remark 6) compared with the approach of Leung and Li (2016).

The paper proceeds as follows. In Section 2 we derive the optimization problem associated with the maximum likelihood estimation (MLE). We then modify the MLE formulation to include terms that promote portfolio sparsity and high mean reversion; in Section 3 we develop an algorithm for the nonsmooth, nonconvex objective based on partial minimization and projection; in Section 4 we provide numerical illustrations using both simulated and real data. We end with a discussion in Section 5.

2. Problem formulation

We first present the maximum likelihood formulation for simultaneously selecting a portfolio from a set of assets, and representing that selection using an Ornstein–Uhlenbeck (OU) process. We also make several theoretical observations about the well-posedness of the estimation problem. We then extend the maximum likelihood formulation to allow selection of higher mean reversion and parsimony in the portfolio.

2.1. OU MLE via optimization

We are given historical data for m assets, with $S^{(T+1)\times m}$ the matrix for assets values over time. Our first goal is to find w, the linear combination of assets that comprise our portfolio, such that the corresponding portfolio price process $x_t := S_t w$ best follows an OU process. We first show that solving for the portfolio with the optimal OU likelihood leads to the optimization problem

$$\min_{a,c,\theta,\|w\|_1=1} \frac{1}{2}\ln(a) + \frac{1}{2Ta}\|\mathcal{A}(c)w - \theta(1-c)\mathbf{1}\|^2, \tag{1}$$

where $\mathcal{A}(c) = S_{1:T} - cS_{0:T-1}$, w is the portfolio to be selected, and a, c, θ are likelihood parameters. The objective function is nonconvex, since $\mathcal{A}(c)$ multiplies w, and also includes a nonconvex constraint $\|w\|_1 = 1$. The 1-norm constraint limits both long and short positions. We are primarily interested in the relative not the absolute magnitude of w_i's. The portfolio weights w_i's and thus value of the constraint (i.e. 1 on the right-hand side) can be scaled, and our method can still be applied (see Remark 5). The derivations of problem (1) are presented below.

An OU process is defined by the SDE

$$dx_t = \mu(\theta - x_t)dt + \sigma dB_t, \tag{2}$$

where B_t is a standard Brownian motion under the physical probability measure. The likelihood of an OU process observed over a sequence $\{x_t\}_{t=1}^T$ is given by

$$\prod_{t=1}^{T} f(x_t|x_{t-1}) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \times \exp\left(-\frac{(x_t - x_{t-1}\exp(-\Delta t\mu) - \theta(1-\exp(-\Delta t\mu)))^2}{2\tilde{\sigma}^2}\right)$$

where $\tilde{\sigma}^2 = \sigma^2\frac{1-\exp(-\Delta t\mu)^2}{2\mu}$. Minimizing the negative log-likelihood results in the optimization problem

$$\min_{\mu,\sigma^2,\theta,w} \frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln(\tilde{\sigma}^2(\mu,\sigma^2)) + \frac{\|A(\mu)w - y(\theta,\mu)\|^2}{2T\tilde{\sigma}^2(\mu,\sigma^2)}, \tag{3}$$

with $y = \theta(1-\exp(-\Delta t\mu))\mathbf{1}$, and $A(\mu) \in \mathbb{R}^{T\times m}$ defined as

$$A(\mu) := S_{1:T} - \exp(-\Delta t\mu)S_{0:T-1},$$

where the subscripts denote ranges for t.

Remark 1. The objective function in (3) is unbounded. Set $w = 0, \theta = 0$; the objective function is then given by

$$\frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln(\sigma^2) + \frac{1}{2}\ln\left(\frac{1-\exp(-2\mu\Delta t)}{2\mu}\right),$$

which goes to $-\infty$ as $\sigma^2 \to 0$.

To solve the issue exposed in Remark 1, we add a 1-norm equality constraint on w, setting $\|w\|_1 = 1$. This constraint is also convenient from a modeling perspective, as it eliminates the need to select which assets in the portfolio are to be long or short *a priori*.

To obtain formulation (1), we denote

$$a = \tilde{\sigma}^2 = \frac{\sigma^2(1-\exp(-2\Delta t\mu))}{2\mu}, \; c = \exp(-\Delta t\mu). \tag{4}$$

Applying the linear approximation $e^x \approx 1 + x$ to (4), we obtain simplified expressions for a and c:

$$a = \Delta t\sigma^2, \quad c = 1 - \Delta t\mu. \tag{5}$$

We can recover μ and σ^2 once we know a and c. For a detailed relationship between the OU model and discrete-time approximation in (5), see Zhang et al. (2018).

Remark 2. The term $\frac{1}{2}\ln(2\pi)$ is dropped from the objective as it is simply a constant. In the subsequent sections when we mention negative log likelihood it refers to value without this constant term.

2.2. Promoting sparsity and mean reversion

Given a set of candidate assets, we want to select a small parsimonious subset to build a portfolio. To add this feature to the model, we want to impose a sparsity penalty on w. While the 1-norm is frequently used, in our case we have already imposed the 1-norm equality constraint $\|w\|_1 = 1$. To obtain sparse solutions under this constraint, we add a multiple of the *nonconvex* constraint $\|w\|_0 \le \eta$ to the maximum likelihood (1). This constraint limits the maximum number of assets to be η.

In addition to sparsifying the solution, we may also want to promote other features of the portfolio. The penalized likelihood framework is flexible enough to allow these enhancements. An important feature is encapsulated by the mean-reverting coefficient μ; a higher μ may be desirable for trading, where positions are opened when deviations are observed, and closed when the portfolio returns to the mean. We can obtain a higher μ by promoting a lower c, e.g. with a linear penalty on $c = 1 - \Delta t\mu$ with a constant penalization coefficient γ. The augmented likelihood function is

$$\min_{a,c,\theta,\|w\|_1=1,\|w\|_0\le\eta} \frac{\ln(a)}{2} + \frac{\|A(c)w - \theta(1-c)\mathbf{1}\|^2}{2Ta} + \gamma c. \tag{6}$$

A higher γ drives c to be lower, and hence drives μ higher.

3. Value function optimization

We develop an algorithm to solve the *nonsmooth, nonconvex* problem (6) by exploiting its rich structure. We define the following nested value functions:

$$\begin{aligned} f(w,a,c,\theta) &= \frac{\ln(a)}{2} + \gamma c + \frac{\|A(c)w - \theta(1-c)\mathbf{1}\|^2}{2Ta} \\ f_1(w,a,c) &= \min_{\theta} f(w,a,c,\theta) \\ f_2(w,a) &= \min_{c} f_1(w,a,c) = \min_{c,\theta} f(w,a,c,\theta) \\ f_3(w) &= \min_{a} f_2(w,a) = \min_{a,c,\theta} f(w,a,c,\theta). \end{aligned} \tag{7}$$

In other words we project out variables a, c, θ. This technique is known as variable projection, or partial minimization. Our main strategy is to use these value functions to recast (6) as the optimization problem

$$\min_{\|w\|_1=1, \|w\|_0 \le \eta} f_3(w), \tag{8}$$

and solve it using projected gradient descent as detailed in Algorithm 1.

To prove Algorithm 1 converges for (8) requires several steps. First, we establish the differentiability of f_3 and Lipschitz continuity of its gradient on region bounded away from the origin in Theorem 1. Second, we develop a projection map onto the set $\mathcal{W} = \{w : \|w\|_1 = 1, \|w\|_0 \le \eta\}$ in Lemma 1 and prove its correctness. Finally we develop the convergence analysis in Theorem 2.

3.1. Differentiability of $f_3(w)$ and Lipschitz continuity of $\nabla f_3(w)$

We first make an assumption on the input data S: for any $\|w\|_2 \ge \epsilon$, we assume that

$$\|Bx(w)_{0:T-1}\|_2 \ge \delta > 0 \tag{9}$$

where $x = Sw$ and $B = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{T} \in \mathbb{R}^{T\times T}$. If $\|Bx(w)_{0:T-1}\|_2 = 0$ for some w, that implies

$$\exists w,\ x(w)_{0:T-1} = \frac{\mathbf{1}^T x(w)_{0:T-1}}{T}\mathbf{1}, \tag{10}$$

but this is a linear system with m (the number of assets) unknowns and T equations, where T usually is much larger than m. Intuitively, (10) says that the portfolio value $x(w)$ must be constant over time and exactly equal to its mean, which is very unlikely with stock market data. Hence assumption (9) is reasonable.

We now state the theorem.

Theorem 1. *Consider $w \in \{w : \|w\|_2 \ge \epsilon\}$. Problem (6) is equivalent to*

$$\min_{\|w\|_1=1, \|w\|_0 \le \eta} f_3(w)$$

where $f_3(w)$ is a differentiable function for small enough γ and ∇f_3 is Lipschitz continuous.

Proof. We start by deriving an explicit expression for the f_1 value function. Taking $\partial_\theta f = 0$, we get

$$0 = \frac{\partial f}{\partial \theta} = (1-c)\mathbf{1}^T(\theta(1-c)\mathbf{1} - A(c)w)$$
$$\Rightarrow \theta^*(c, w) = \frac{\mathbf{1}^T(x(w)_{1:T} - cx(w)_{0:T-1})}{T(1-c)}.$$

Plugging $\theta^*(c, w)$ into f, we get an explicit form of f_1:

$$f_1(w, a, c) = \frac{1}{2}\ln(a) + \gamma c + \frac{\|B(x(w)_{1:T} - cx(w)_{0:T-1})\|^2}{2Ta} \tag{11}$$

with $B = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{T}$ a projection matrix onto the space of vectors in $\mathbb{R}^T$ with mean 0. To simplify the following analysis, we define

$$b_1(w) := Bx(w)_{1:T}, \quad b_0(w) := Bx(w)_{0:T-1}.$$

We now apply a differential variant of the implicit function theorem to f_1. Let $F(w, y)$ be f_1 in (11) where $y = [a, c]$, so that $f_3(w) = \min_y F(w, y)$. From Bell and Burke (2008, Theorem 2), if there exist $\bar{w}, \bar{y}$ such that $F_y(\bar{w}, \bar{y}) = 0$ and $F_{yy}(\bar{w}, \bar{y})$ is positive definite, then in the neighborhood of $(\bar{w}, \bar{y})$ where $f_3(w)$ is defined, it is twice differentiable. In our case,

$$F_y(w, y) = \begin{bmatrix} \frac{1}{2a} - \frac{\|b_1 - cb_0(w)\|^2}{2Ta^2} \\ \gamma - \frac{1}{Ta} b_0(w)^T(b_1(w) - cb_0(w)) \end{bmatrix}$$

$$F_{yy}(w, y) = \begin{bmatrix} -\frac{1}{2a^2} + \frac{\|b_1(w)-cb_0(w)\|^2}{Ta^3} & \frac{b_0(w)^T(b_1(w)-cb0(w))}{Ta^2} \\ \frac{b_0(w)^T(b_1(w)-cb_0(w))}{Ta^2} & \frac{1}{Ta} b_0(w)^T b_0(w) \end{bmatrix}.$$

When $F_y(\bar{w}, \bar{y}) = 0$, we have

$$\bar{a}(\bar{w}) = \frac{1}{T}\|b_1(\bar{w}) - \bar{c}b_0(\bar{w})\|^2,$$
$$\gamma(\bar{w}) = \frac{1}{T\bar{a}} b_0(\bar{w})^T(b_1(\bar{w}) - \bar{c}b_0(\bar{w})),$$

and F_{yy} simplifies to

$$F_{yy}(\bar{w}, \bar{y}) = \begin{bmatrix} \frac{1}{2\bar{a}(\bar{w})^2} & \frac{\gamma}{\bar{a}(\bar{w})} \\ \frac{\gamma(\bar{w})}{\bar{a}(\bar{w})} & \frac{1}{T\bar{a}(\bar{w})} b_0(\bar{w})^T b_0(\bar{w}) \end{bmatrix}.$$

Given assumption in (9), when $\gamma = 0$, we immediately have that $F_{yy}(\bar{w}, \bar{y})$ is diagonal with positive entries. If $\gamma > 0$, we write $\bar{a}$ in terms of $\bar{w}$ by solving $F_y(\bar{w}, \bar{y}) = 0$ and

$$\bar{a} = \frac{\|b_0\|^2}{2T\gamma^2} - \frac{\sqrt{\|b_0\|^4 - 4\gamma^2(\|b_0\|^2\|b_1\|^2 - (b_0^T b_1)^2)}}{2T\gamma^2}$$

where b_0, b_1 are evaluated at $\bar{w}$.

When $\|b_0\|^2\|b_1\|^2 - (b_0^T b_1)^2 \ne 0$, in order for $\bar{a}$ to be a real number, γ has to be small enough so that

$$\|b_0\|^4 - 4\gamma^2(\|b_0\|^2\|b_1\|^2 - (b_0^T b_1)^2) \ge 0 \ \forall \|w\|_2 \ge \epsilon$$
$$\Rightarrow 0 \le \gamma \le \inf_{\|w\|_2 \ge \epsilon} \frac{1}{2}\sqrt{\frac{\|b_0\|^4}{\|b_0\|^2\|b_1\|^2 - (b_0^T b_1)^2}}.$$

The infimum can be attained because $\|b_0\|^2$ is bounded below by the assumption on input data and $\|b_0\|^2\|b_1\|^2 - (b_0^T b_1)^2 \le \|b_0\|^2\|b_1\|^2$ is bounded above.

Thus the determinant of $F_{yy}(\bar{w}, \bar{y})$ is

$$\det(F_{yy}(\bar{w}, \bar{y})) = \frac{\|b_0\|^2 - 2T\bar{a}\gamma^2}{2T\bar{a}^3} > 0$$

using the expression for $\bar{a}$. Since $F_{yy}(\bar{w}, \bar{y})$ is a 2×2 matrix with a positive first minorant and positive determinant, it must be positive definite. Hence the conditions in Theorem 2, Bell and Burke (2008) are satisfied, implying that f_3 is twice differentiable on $\{w : \|w\|_2 \ge \epsilon\}$. Moreover, the eigenvalues of F_{yy} depend continuously on w, which is restricted to a compact set $\mathcal{W}$. Hence the operator norm of F_{yy} has an upper bound for all $w \in \mathcal{W}$, and this value is also a Lipschitz constant for $\nabla f(w)$. □

Remark 3. The expression for $\bar{c}$ is

$$\bar{c} = \frac{b_0^T b_1 - T\bar{a}\gamma}{\|b_0\|^2}.$$

There is no guarantee that $\bar{c}$ is positive. Indeed $\bar{c}$ can potentially be negative, in which case no corresponding positive μ exists. This means that the given data and γ do not permit the construction of a mean-reverting time series. The γ term in numerator drives $\bar{c}$ towards negative values, which means that the higher mean-reverting level we request, the less likely such a process can be constructed.

Remark 4. When $\gamma > 0$, $f_3(w)$ is given by

$$f_3(w) = \frac{1}{2}\ln(\bar{a}) + \frac{\|b_1\|^2}{2T\bar{a}} - \frac{(b_0^T b_1)^2}{2T\bar{a}\|b_0\|^2} - \frac{T\bar{a}\gamma^2}{2\|b_0\|^2} + \frac{\gamma b_0^T b_1}{\|b_0\|^2}.$$

When $\gamma = 0$, $f_3(w)$ simplifies to

$$f_3(w) = \frac{1}{2}\ln(\bar{a}) + 1/2.$$

In both expressions, $\bar{a}, b_0, b_1$ are evaluated at w as in the proof of Theorem 1. See Zhang et al. (2018) for a detailed derivation.

Remark 5. If we scale w to Kw, then

$$\begin{aligned} f_3(Kw) =& \frac{1}{2}\ln(K^2\bar{a}) + \frac{K^2\|b_1\|^2}{2TK^2\bar{a}} - \frac{K^4(b_0^T b_1)^2}{2TK^4\bar{a}\|b_0\|^2} \\ & - \frac{TK^2\bar{a}\gamma^2}{2K^2\|b_0\|^2} + \frac{K^2\gamma b_0^T b_1}{K^2\|b_0\|^2} \\ =& \ln(K) + f_3(w). \end{aligned}$$

Let $v = Kw$. Then

$$\min_{\|v\|_1=K, \|v\|_0\le\eta} f_3(v) \Leftrightarrow \min_{\|w\|_1=1, \|w\|_0\le\eta} f_3(w).$$

3.2. Projection map onto $\mathcal{W}$

The set of interest,

$$\mathcal{W} = \{w : \|w\|_1 = 1, \|w\|_0 \le \eta\} \tag{12}$$

is highly nonconvex, but admits an efficient projection. First, we reduce the problem to projection of a non-negative vector, and recovering the true projection by element-wise multiplication:

$$\begin{aligned} \text{proj}_{\mathcal{W}}(x) &\leftarrow \text{argmin}_{\|z\|_1=1, \|z\|_0\le\eta} \|x - z\|^2 \\ &= \text{sign}(x) \odot \text{argmin}_{\|u\|_1=1, \|u\|_0\le\eta} \||x| - u\|^2 \\ &= \text{sign}(x) \odot \text{argmin}_{u^T 1=1, u\ge 0, \|u\|_0\le\eta} \||x| - u\|^2 \\ &= \text{sign}(x) \odot \text{proj}_{\Delta_1\cap\|\cdot\|_0\le\eta}(|x|), \end{aligned}$$

i.e. the projection of $|x|$ onto the intersection of the 1-simplex and the set of vectors with at most η nonzero entries. Above $\odot$ is element-wise multiplication, and the second equality is obtained by a change of variable $u = \text{sign}(x) \odot z$.

Next, to find $\text{proj}_{\Delta_1\cap\|\cdot\|_0\le\eta}(|x|)$, we propose the following procedure, whose correctness is proved in Lemma 1.

- Order $|x|$ such that $|x_1| \ge |x_2| \ge \cdots \ge |x_m|$.
- $x^+_{1:\eta} \leftarrow \text{argmin}_{u_{1:\eta}\in\Delta_1} \||x_{1:\eta}| - u_{1:\eta}\|^2, \quad x^+_{\eta+1:m} = 0.$

Lemma 1. *Suppose $v \in \mathbb{R}^m$. Let K be any size k subset of $I = \{1, \ldots, m\}$ and $\mathcal{K}$ the union of all such Ks. $I - K$ denotes the complement of K in I. The problem is to find*

$$\min_{u_K\in\Delta_1, u_{I-K}=0, K\in\mathcal{K}} \frac{1}{2}\|v - u\|^2.$$

Let us reorder v such that $v_1 \ge v_2 \ge \cdots \ge v_m$. The claim is that the optimal $K_{opt} = \{1, 2, \ldots, k\}$, i.e. the indices corresponding to the k largest components in v.

Proof. Equivalently, the problem can be stated as

$$\begin{aligned} & \min_{u_K\in\Delta_1, K\in\mathcal{K}} \frac{1}{2}\sum_{j\in K}(v_j - u_j)^2 + \frac{1}{2}\sum_{j\in I-K} v_j^2 \\ =& \min_{u_K\in\Delta_1, K\in\mathcal{K}} \frac{1}{2}\|v_K - u_K\|^2 + \frac{1}{2}\|v_{I-K}\|^2 \\ \Leftrightarrow& \min_{u_K\in\Delta_1, K\in\mathcal{K}} \frac{1}{2}\|v_K - u_K\|^2 - \frac{1}{2}\|v_K\|^2 + \frac{1}{2}\|v\|^2. \end{aligned}$$

Note that the last term $\frac{1}{2}\|v\|^2$ does not depend on u_K, so we can focus on the first two terms, i.e.

$$\min_{u_K\in\Delta_1, K\in\mathcal{K}} \frac{1}{2}\|v_K - u_K\|^2 - \frac{1}{2}\|v_K\|^2.$$

Suppose there is some K' that is different from K_{opt} and denote the corresponding v as $v_{K'}$. Define $f(y)$ and $g(t)$ by

$$\begin{aligned} f(y) &= -\frac{1}{2}\|y\|^2 + \min_{z\in\Delta_1}\frac{1}{2}\|y - z\|^2, \\ g(t) &= f((1-t)v_{K_{opt}} + t v_{K'}). \end{aligned}$$

Then we have

$$\begin{aligned} & f(v_{K'}) - f(v_{K_{opt}}) = g(1) - g(0) = \int_0^1 g'(t)dt, \\ & g'(t) = \nabla f((1-t)v_{K_{opt}} + t v_{K'})^T(-v_{K_{opt}} + v_{K'}), \\ & \nabla f(y) = -y + y - z^* = -z^* \in -\Delta_1, \end{aligned}$$

where z^* is the projection of y onto the simplex Δ_1. $\nabla f(y)$ is nonpositive in all components and strictly negative in some components. Therefore, $\nabla f((1-t)v_{K_{opt}} + t v_{K'}) \le 0$. Further $-v_{K_{opt}} + v_{K'} \le 0$ because $v_{K_{opt}}$ contains the k-largest components of v. As a result,

$$g'(t) \ge 0 \Rightarrow \int_0^1 g'(t)dt \ge 0 \Rightarrow f(v_{K'}) \ge f(v_{K_{opt}}).$$

This shows that K_{opt} must be the optimal choice. Once we have determined K, we can apply simplex projection onto v_K with existing techniques (Duchi, Shalev-Shwartz, Singer, & Chandra, 2008).

Algorithm 1 Projected Gradient Descent for $f_3(w; \gamma, \eta)$ (7).

Input: $w \in \mathbb{R}^m, S, f_3, \gamma, \eta$

1: $\mathcal{W} = \{w : \|w\|_1 = 1, \|w\|_0 \le \eta\}$
2: **while** not converged **do**
3: $\quad w^k \leftarrow \text{Proj}_{\mathcal{W}}(w^{k-1} - \delta_i \nabla_w f_3(w^{k-1}; \gamma, \eta))$
Recover a, c, θ from w.
4: (δ_i denotes stepsize via line search.)

3.3. Convergence analysis

Algorithm 1 is projected gradient descent for the value function f_3 over the nonconvex set $\mathcal{W}$, which converges for a large class of nonconvex functions (Attouch, Bolte, & Svaiter, 2013). However our problem does not satisfy the assumptions of Attouch et al. (2013) because the gradient of loss function $f_3(w)$ is not globally Lipschitz. As shown in the previous section, when w is bounded away from the origin, the gradient is Lipschitz; when w approaches the origin, however, the function value goes to ∞ and the gradient is not Lipschitz. Fig. 1 shows a schematic plot of the loss function f_3. Global Lipschitz of gradient is used to establish sufficient decrease in the loss, a key component of any convergence theory. We derive Lemma 2 to establish sufficient decrease of f_3, taking advantage of the fact that $\mathcal{W}$ is bounded away from the origin. We also include additional lemmas to provide a full picture of the analysis. The main result is presented in Theorem 2.

Theorem 2. *Consider the optimization problem*

$$\min_{w\in\mathcal{W}} f(w),$$

where f is the objective function f_3 in (8) and $\mathcal{W} = \{w : \|w\|_1 = 1, \|w\|_0 \le \eta\}$ is the nonconvex constraint set in (8). In particular, f is nonconvex and is not smooth and has singularities near the origin.

Let $\{w^k\}$ be the sequence generated by the line search $w^+ \leftarrow \Pi_{\mathcal{W}}(w - t\nabla f(w))$ with $t \ge \underline{t}$, then

$$\nabla f(w^k) + \partial\delta_{\mathcal{W}}(w^k) \to 0$$

as $k \to \infty$. Here $\Pi_{\mathcal{W}}$ denotes the projection onto $\mathcal{W}$, $\underline{t}$ a lower bound on t, $\delta_{\mathcal{W}}(z) = \begin{cases} 0 & z \in \mathcal{W} \\ \infty & o.w. \end{cases}$, and $\partial\delta$ denotes the limiting subdifferential, which is the appropriate generalization of derivative for this situation; see e.g. Rockafellar and Wets (2009).

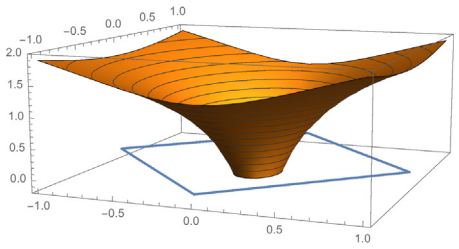

Fig. 1. 3D plot of the objective function in (8) for $w \in \mathbb{R}^2$, with constraint set $\|w\|_1 = 1, \|w\|_0 \leq 2$. Our goal is to find the minimum value of f_3 (yellow 3D plot) restricted to $\mathcal{W}$ (edges of the blue diamond).

Proof. This theorem is proved using the following lemmas (detailed below):

- Lemma 2 relates decrease in function values $f(w) - f(w^+)$ to consecutive differences $\|w^+ - w\|^2$, using Lipschitz continuity of ∇f on the set

  $C = \mathbb{R}^n - \mathbb{B}_\epsilon(0) \supset \mathcal{W}$,

  where $\mathbb{B}_\epsilon(0) = \{w : \|w\|_2 < \epsilon\}$ and $\epsilon \leq \sqrt{2}/2$.
- Lemma 3 uses Lemma 2 to show that $\|w^{k+1} - w^k\| \downarrow 0$.
- Lemma 4 shows that elements in the subdifferential $\nabla f + \delta_{\mathcal{W}}$ converge to 0 using Lemma 3.

Lemma 2. *Let $C = \mathbb{R}^n - \mathbb{B}_\epsilon(0)$ where $\mathbb{B}_\epsilon(0) = \{w : \|w\|_2 < \epsilon\}$ and $\epsilon \leq \sqrt{2}/2$. In other words, $\mathbb{B}_\epsilon(0)$ is inside the 1-norm sphere $\{w : \|w\|_1 = 1\}$. Let $L(\epsilon)$ be the upper bound such that*

$$\|\nabla f(w) - \nabla f(w')\| \leq L(\epsilon)\|w - w'\| \ \forall w, w' \in C.$$

Suppose $w \in \mathcal{W}$, and let $w^+ \leftarrow \Pi_{\mathcal{W}}(w - t\nabla f(w))$. Then we have

$$f(w^+) \leq f(w) - \frac{1/t - 15L(\epsilon)}{2}\|w^+ - w\|^2.$$

Proof. If the line segment from w to w^+ does not go through $\mathbb{B}_\epsilon$, then by $L(\epsilon)$-Lipschitz,

$$f(w^+) \leq f(w) + \langle w^+ - w, \nabla f(w)\rangle + \frac{L(\epsilon)}{2}\|w^+ - w\|^2.$$

Otherwise, let w_1, w_4 denote the intersection of the line segment with the closed ball $\bar{\mathbb{B}}_\epsilon$ and $w_0 = w, w_5 = w^+$. We can find a 2D circle centered at the origin with diameter 2ϵ that passes through w_1, w_4. Then we can find a tight box with length 2ϵ that contains the circle. Let w_2, w_3 be two vertices on the box, through which we can define a path from w_1 to w_4 along the box. This path does not go through $\mathbb{B}_\epsilon$.

By $L(\epsilon)$-Lipschitz of f on C,

$$\begin{aligned}
&f(w_{i+1}) \leq f(w_i) + \langle w_{i+1} - w_i, \nabla f(w_i)\rangle \\
&\quad + \frac{L(\epsilon)}{2}\|w_{i+1} - w_i\|^2 \\
\Rightarrow &f(w^+) \leq \sum_{i=0}^{4}\langle w_{i+1} - w_i, \nabla f(w_i)\rangle + \frac{L(\epsilon)}{2}\|w_{i+1} - w_i\|^2 \\
&\quad + f(w) \\
\Rightarrow &f(w^+) \leq f(w) + \langle w^+ - w, \nabla f(w)\rangle \\
&\quad + \sum_{i=0}^{4}\langle w_{i+1} - w_i, \nabla f(w_i) - \nabla f(w)\rangle + \frac{L(\epsilon)}{2}\|w_{i+1} - w_i\|^2 \\
\Rightarrow &f(w^+) \leq f(w) + \langle w^+ - w, \nabla f(w)\rangle \\
&\quad + \sum_{i=0}^{4} L(\epsilon)\|w_{i+1} - w_i\|\|w_i - w_0\| + \frac{L(\epsilon)}{2}\|w_{i+1} - w_i\|^2 \\
\Rightarrow &f(w^+) \leq f(w) + \langle w^+ - w, \nabla f(w)\rangle + \frac{15L(\epsilon)}{2}\|w^+ - w\|^2.
\end{aligned}$$

By the definition of projection,

$$\begin{aligned}
w^+ =&\operatorname{argmin}_{y\in\mathcal{W}}\frac{1}{2}\|w - t\nabla f(w) - y\|^2 \\
\Rightarrow &\frac{1}{2}\|w - w^+ - t\nabla f(w)\|^2 \leq \frac{1}{2}\|t\nabla f(w)\|^2 \\
\Rightarrow &\frac{1}{2t}\|w - w^+\|^2 + \langle w^+ - w, \nabla f(w)\rangle \leq 0.
\end{aligned}$$

Adding them together yields

$$\begin{aligned}
&f(w^+) + \frac{1}{2t}\|w - w^+\|^2 \leq f(w) + \frac{15L(\epsilon)}{2}\|w^+ - w\|^2, \\
&f(w^+) \leq f(w) - \frac{1/t - 15L(\epsilon)}{2}\|w^+ - w\|^2.
\end{aligned}$$

Lemma 3. *Let $\{w^k\}$ be a sequence generated by $w^+ \leftarrow \Pi_{\mathcal{W}}(w - t\nabla f(w))$ with initial guess $w^0 \in C$, and let $K = 15L(\epsilon)$. If we choose t_k at each step such that $\underline{t} \leq t_k < \frac{1}{K}$, then*

$$\sum_{k=1}^{\infty}\|w^{k+1} - w^k\|^2 < \infty \Rightarrow \lim_{k\to\infty}\|w^{k+1} - w^k\| = 0.$$

Proof. Since $\underline{t} \leq t_k < \frac{1}{K}$, the expression $\frac{2}{1/t_k - K}$ is bigger than 0, and is upper bounded by some $M > 0$ for all k. By Lemma 2

$$\begin{aligned}
\|w^{k+1} - w^k\|^2 &\leq \frac{2}{1/t_k - K}[f(w^k) - f(w^{k+1})] \\
&\leq M[f(w^k) - f(w^{k+1})].
\end{aligned}$$

Summing up k from 0 to $N - 1$ gives

$$\begin{aligned}
&\sum_k \|w^{k+1} - w^k\|^2 \leq M\sum_k f(w^k) - f(w^{k+1}) \\
&= M[f(w^0) - f(w^N)] \leq M[f(w^0) - f(w^*)].
\end{aligned}$$

Taking $N \to \infty$ yields the desired result.

Lemma 4. *Let $\{w^k\}$ be a sequence generated by $w^+ \leftarrow \Pi_{\mathcal{W}}(w - t\nabla f(w))$. Define*

$$A^k = \frac{1}{t_{k-1}}(w^{k-1} - w^k) + \nabla f(w^k) - \nabla f(w^{k-1}).$$

Then $A^k \in \nabla f(w^k) + \partial\delta_{\mathcal{W}}(w^k)$ and $A^k \to 0$ as $k \to \infty$.

Proof. By the definition of projected gradient step,

$$\begin{aligned}
&0 \in \nabla f(w^{k-1}) + \frac{1}{t_{k-1}}(w^k - w^{k-1}) + \partial\delta_{\mathcal{W}}(w^k) \\
\Rightarrow &\frac{1}{t_{k-1}}(w^{k-1} - w^k) \in \nabla f(w^{k-1}) + \partial\delta_{\mathcal{W}}(w^k).
\end{aligned}$$

Hence,

$$\begin{aligned}
A^k \in&\nabla f(w^{k-1}) + \partial\delta_{\mathcal{W}}(w^k) + \nabla f(w^k) - \nabla f(w^{k-1}) \\
=&\partial\delta_{\mathcal{W}}(w^k) + \nabla f(w^k).
\end{aligned}$$

In turn, we have

$$\begin{aligned}
\|A^k\| \leq&\frac{1}{t_{k-1}}\|w^{k-1} - w^k\| + L(\epsilon)\|w^k - w^{k-1}\| \\
&\leq \left(\frac{1}{\underline{t}} + L(\epsilon)\right)\|w^k - w^{k-1}\|.
\end{aligned}$$

By Lemma 3, as $k \to \infty$, $A^k \to 0$.

Table 1
Estimated parameters, weights, and nll. We set $\gamma = 0$ for top three rows, $\gamma = 0.5$ for bottom three rows.

η	μ	σ^2	θ	w	nll (train, test)
5	2.4	0.09	0.07	[.12, −.11, .33, .31, −.12]	−(3.03, 3.04)
4	2.9	0.10	0.32	[.13, .12, .38, .36, 0]	−(2.96, 2.94)
3	2.6	0.11	0.23	[0.16, 0, 0.43, 0.42, 0]	−(2.90, 2.90)
5	5.0	0.09	0.08	[.11, −.11, −.33, .32, −.12]	−(3.02, 2.99)
4	5.8	0.10	0.18	[.14, 0, .35, .34, .16]	−(2.93, 2.84)
3	4.7	0.11	0.27	[0, 0, .40, .40, −.19]	−(2.88, 2.83)

4. Numerical results

Algorithm 1 is much faster than the standard projected gradient descent on all unknowns (Zhang et al., 2018, Section IV B). We give additional examples to show how the approach identifies mean-reverting time series using simulated data. We simulate five time series; four from an OU process with (μ, σ, θ) given by $(1, 1, 0), (4, 1, 1), (1, 0.5, 1), (4, 0.5, 0)$, and one is non-OU time series with $\sigma = .1$ (the fifth time series). All have $T = 500$ and $\Delta t = 0.01$. We divide the data into training set (70% of data) and test sets (30% of data).

Table 1 compares the estimated OU parameters and weight vectors as we tune γ and η. Top three rows correspond to $\gamma = 0$, and bottom three rows $\gamma = 0.5$. When $\gamma = 0, \eta = 5$, the model puts 64% of the weights into the pair of OU time series with $\sigma = 0.5$. It is evident from the results that the model favors OU time series with a lower σ value but is less sensitive to μ values.

With larger η we reach lower negative log likelihood (nll) since that means more freedom in choosing assets.

Remark 6. As noted in Remark 3, γ will drive $\bar{c}$ to be negative. If γ is large, the model may not find a feasible time series combination. In addition, γ controls the balance between negative log likelihood and mean-reversion promoting term (i.e. γc). If γ is too large, the model may choose a portfolio that has high negative log likelihood, i.e. low likelihood. Also, since $\mu = -\frac{1}{\Delta t}\log(c)$ and $c \in (0, 1)$, a small increase in c will be amplified in μ. Hence a small γ usually suffices and is preferred. In practice it is a good idea to start with $\gamma = 0$. The tuning of η is straightforward. One can set it to be the desired number of assets for the portfolio.

Real data. We performed experiments with empirical price data from three groups of selected assets: precious metals, large equities and oil companies. Data were taken from Yahoo Finance, and give closing stock prices for each asset over the past five years. The first 70% of data (over time) is used for training, and the rest for testing.

For each group, we progressively augmented the set of candidate assets in pairs, and applied our approach. The model determined asset weights, along with negative log-likelihoods of portfolios and of individual assets are given in Table 2. The portfolios' negative log likelihoods are generally smaller than negative log likelihoods of individual assets in that portfolio and decrease as we include more assets, which means we can obtain more OU-representable portfolios as the candidate sets expand. The negative log likelihood on the test set can sometimes be significantly larger than that on the training set. In Group 2, individual assets such as GOOG, JNJ and MCD have significantly larger negative log likelihood on test than on training. The discrepancy in likelihood indicates that those assets have very different patterns before and after the split of training/test. As a result, the constructed portfolios also tend to have larger negative log likelihood on test set. This also suggests that one should check individual asset patterns before generalizing fitted model to another time period.

Table 2
Negative log-likelihood (nll) of assets groups for $\eta \in \{2, 4, 6\}$ (no. of assets in portfolio) and $\gamma = 0$. The bottom row shows the (training, testing) nll of our optimal portfolios.

Assets	2	4	6	indiv. nll (train, test)
GLD	−0.17	−0.08	−0.07	0.77, 0.44
GDX		−0.21	−0.29	0.05, −0.30
GDXJ			0.03	0.70, 0.38
SLV	0.83	0.44	0.30	−0.69, −1.0
GG			0.10	−0.04, −0.44
ABX		0.27	0.21	−0.24 ,−0.54
Port.	−1.48, −1.72	−1.95, 2.12	−2.18, −2.35	
GOOG				2.66, 3.06
JNJ		−0.12	−0.10	0.40, 0.86
NKE	−0.49	−0.36	−0.27	0.09, 0.43
MCD		−0.11	−0.07	0.49, 1.09
SBUX	0.51	0.41	0.36	0.02, 0.05
SPY			0.12	0.95, 1.00
VIG				−0.07, 0.01
VO			−0.08	0.53, 0.45
Port.	−0.70, −0.08	−0.77, −0.14	−1.07, −0.52	
BP			−0.01	−0.09, −0.33
COP		−0.01	−0.01	0.46, 0.25
CVX		−0.02	−0.01	0.79, 0.73
OIL	−0.59	−0.57	−0.57	−0.84, −1.25
USO	0.41	0.41	0.41	−0.45, −0.86
VLO			0.002	0.58, 0.43
XOM				0.48, 0.26
Port.	−2.89, −3.29	−2.94, −3.35	−2.96, −3.37	

Table 3
Model estimations with different γ and η for precious metals, large equities, and oil companies.

γ	η	μ	σ^2	θ	nll (train, test)
0	2	2.69	4.77	−6.42	−1.48, −1.72
0.5	2	4.51	4.78	−6.14	−1.48, −1.72
0	4	2.28	1.87	−2.90	−1.95, −2.13
0.5	4	7.06	2.35	−2.65	−1.84, −2.07
0	6	1.20	1.17	−3.30	−2.18, −2.35
0.5	6	12.70	1.11	−0.98	−2.21, −2.39
0	2	5.74	22.85	−0.57	−0.70, −0.08
0.5	2	11.49	23.11	−0.56	−0.69 −0.04
0	4	1.90	19.76	−22.84	−0.77, −0.14
0.5	4	4.12	19.63	−23.61	−0.77, 0.01
0	6	3.54	10.87	1.00	−1.07, −0.52
0.5	6	6.35	10.64	−1.78	−1.08, −0.46
0	2	11.80	0.28	0.89	−2.89, −3.29
0.5	2	34.43	0.29	1.00	−2.87, −3.26
0	4	16.84	0.26	0.47	−2.94, −3.35
0.5	4	37.80	0.27	0.44	−2.92, −3.31
0	6	17.39	0.25	0.49	−2.95, −3.37
0.5	6	42.63	0.26	0.63	−2.93, −3.31

We also conducted experiments varying γ to promote larger μ. As summarized in Table 3, when $\gamma > 0$, we see increasing μ across asset groups. As $c = \exp(-\Delta t\mu) \approx 1 - \Delta t\mu$, the change in c due to γ will be magnified in μ, hence we may see fairly drastic increase in μ.

Comparison with pairs trading We compared our approach with that in chapter 2 of Leung and Li (2016) on pairs trading. In Leung and Li (2016), two assets are selected first, from which a portfolio is constructed as

$$X = S_1 - \beta S_2 \tag{13}$$

Table 4
Summary of portfolio weights from our model and method of (13) applied to different pairs.

	β	Portfolio weights
GLD $-\beta$SLV	3.68	[0.21, −0.79]
$-\beta$GLD + SLV	0.19	[−0.17, 0.83]
Our model	–	[−0.17, 0.83]
NKE $-\beta$SBUX	0.61	[0.63, −0.37]
$-\beta$NKE + SBUX	0.52	[−0.33, 0.67]
Our model	–	[−0.49, 0.51]
OIL $-\beta$USO	0.67	[0.59, −0.41]
$-\beta$OIL + USO	1.42	[−0.59, 0.41]
Our model	–	[−0.59, 0.41]

where S_1 and S_2 are asset price time series. This "β-method" requires making long the first asset and short the other. With the weight of the first asset fixed to be 1, this method first determines, for each fixed β, the model parameters that maximize the OU likelihood of the corresponding portfolio X. Then, in a separate step, it searches over a range of β for the MLE. For this approach to identify the optimal pairs, one needs to further find two optimal β's by switching positions of two assets in (13). In contrast, our model solves for the optimal portfolio in a single step. For the examples in Table 4, we can simply take the results from our model with $\eta = 2$ from Table 2.

5. Concluding remarks

We have solved a joint optimization problem for simultaneous portfolio selection and OU-fitting. We also incorporated desirable portfolio features, including higher mean-reversion and sparser portfolios, both important for practical trading purposes. We developed a fast algorithm for the nonsmooth nonconvex optimization problem, and presented our solutions using both simulated and real data, resulting in useful portfolios from several asset classes. Our model extends the pairs trading model in Leung and Li (2016), develops a convergence analysis for the algorithm, and provides a comparison analysis.

References

Attouch, H., Bolte, J., & Svaiter, B. (2013). Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, *137*(1–2), 91–129.

Bell, B. M., & Burke, J. V. (2008). Algorithmic differentiation of implicit functions and optimal values. In *Advances in automatic differentiation* (pp. 67–77). Springer.

d'Aspremont, A. (2011). Identifying small mean-reverting portfolios. *Quantitative Finance*, *11*(3), 351–364.

Duchi, J., Shalev-Shwartz, S., Singer, Y., & Chandra, T. (2008). Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on machine learning* (pp. 272–279). ACM.

Gatev, E., Goetzmann, W., & Rouwenhorst, K. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, *19*, (3), 797–827.

Kitapbayev, Y., & Leung, T. (2017). Optimal mean-reverting spread trading:Nonlinear integral equation approach. *Annals of Finance*, *13*(2), 181–203.

Leung, T., & Li, X. (2015). Optimal mean reversion trading with transaction costs and stop-loss exit. *International Journal of Theoretical & Applied Finance*, *18*(3), 15500.

Leung, T., & Li, X. (2016). *Modern trends in financial engineering, Optimal mean reversion trading: mathematical analysis and practical applications*. Singapore: World Scientific.

Ornstein, L. S., & Uhlenbeck, G. E. (1930). On the theory of the Brownian motion. *Physical Review*, *36*, 823–841.

Rockafellar, R. T., & Wets, R. J.-B. (2009). *Variational analysis, vol. 317*. Springer Science & Business Media.

Zhang, J., Leung, T., & Aravkin, A. (2018). Mean reverting portfolios via penalized maximum likelihood estimation and optimization. In *Proceedings of the IEEE conference on decision and control*.

Zhao, Z., & Palomar, D. P. (2018). Mean-reverting portfolio with budget constraint. *IEEE Transactions on Signal Processing*, *66*(9), 2342–2357.

Zhao, Z., Zhou, R., & Palomar, D. P. (2019). Optimal mean-reverting portfolio with leverage constraint for statistical arbitrage in finance. *IEEE Transactions on Signal Processing*.

Jize Zhang is a Ph.D. student in Applied Mathematics at the University of Washington. She obtained her B.S. from Smith College in 2014. She works on optimization research with applications to multiple fields, including finance, engineering, and global health.

Tim Leung is an Associate Professor in the Department of Applied Mathematics and the Director of the Computational Finance & Risk Management (CFRM) program at University of Washington in Seattle. Previously, he had been an Assistant Professor in the Department of Applied Mathematics & Statistics at Johns Hopkins University (2008–2011) and in the Department of Industrial Engineering & Operations Research at Columbia University (2011–2016). He obtained his B.S. from Cornell University and Ph.D. from Princeton University where he was supported by the Charlotte Procter Honorific Fellowship. His research areas are Quantitative Finance and Stochastic Optimal Control.

Aleksandr Aravkin received B.S. degrees in Mathematics and Computer Science from the University of Washington in 2004. He then received an M.S. in Statistics and a Ph.D. in Mathematics from the University of Washington in 2010. He was a joint postdoctoral fellow in Earth and Ocean Sciences and Computer Science at the University of British Columbia from 2010 to 2012, and a research staff member at the IBM T.J. Watson Research Center from 2012 to 2015. During this time he also worked at Columbia as an Adjunct Professor in Computer Science and IEOR. In 2015, Dr. Aravkin joined the faculty at UW Applied Mathematics, where he works on theoretical and practical problems connected to data science, including convex and variational analysis, statistical modeling, and algorithm design.