Sparse Mean-Reverting Portfolios via Penalized Likelihood Optimization

Jize Zhang [a] Tim Leung [b] Aleksandr Aravkin [c]

[a] *Department of Applied Mathematics, University of Washington, USA (e-mail: jizez@uw.edu)*

[b] *Department of Applied Mathematics, University of Washington, USA (e-mail: timleung@uw.edu)*

[c] *Department of Applied Mathematics, University of Washington, USA (e-mail: saravkin@uw.edu)*

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 and low variance though 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. The problem requires custom analysis because the objective function does not have a Lipschitz-continuous gradient. Through our experiments using simulated and empirical price data, the proposed algorithm significantly outperforms standard approaches that do not exploit problem structure.

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 correlated or cointegrated assets, 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 exhibit 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 [1]. A number of studies on trading mean-reverting prices [2,3] and the empirical performance of pairs trading [4].

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 Ornstein-Uhlenbeck (OU) process [5] through penalized maximum likelihood estimation. A major feature of our joint optimization approach is that we *simultaneously* solve for the optimal portfolio and the corresponding parameters for maximum likelihood. This unified approach is different from prior work for OU portfolio selection, which break the problem up into stage-wise computations. For example, [6] first finds an OU representation for the time series of multiple assets, and then solves a second optimization problem to find the portfolio based on that representation. Conversely, [1] fits 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 authors' short proceedings paper [7]. In particular, the current paper

- Develops an efficient projection onto the intersection of ℓ_0 level sets and the nonconvex set $\|x\|_1 = 1$, with a correctness proof (Lemma 1).
- Establishes differentiability properties of the key value function used in the approach (Section 3.1).
- Proves convergence of the proposed algorithm (Theorem 2), even though the objective fails to have a Lipshitz-continuous gradient (Section 3.3).
- Develops numerical examples with empirical prices (Section 4.2) and compares results with those in [1].

[1] Corresponding author: Aleksandr Aravkin. This research was partially supported by the Washington Research Foundation Data Science Professorship.

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, as well as terms that select fewer assets from a larger candidate set. In Section 3 we develop an algorithm for the nonsmooth, nonconvex objective based on partial minimization and projection, and show that it performs much better than a standard algorithm that doesn't exploit problem structure. 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 lower variance, 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

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

where $A(c) = S_{1:T} - cS_{0:T-1}$, w is the portfolio to be selected, and a, c, θ, σ are likelihood parameters. The objective function is nonconvex, since $A(c)$ multiplies w, and also includes a nonconvex constraint $\|w\|_1 = 1$. The derivations are presented below.

An OU process is defined by the SDE

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

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

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

with $y = \theta(1-\exp(-\Delta t\mu))\mathbf{1}$, and $A \in \mathbb{R}^{n\times 2}$ defined as

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

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

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

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

$$a = \Delta t\sigma^2, \quad c = 1 - \Delta t\mu. \quad (5)$$

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 [7].

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. 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

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

A higher γ drives c to be lower, implying a higher μ as a result. In addition to increasing the speed of mean reversion μ, we can also constrain the resulting volatility of the portfolio to be larger than a given threshold [8].

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:

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

Our main strategy is to to use these value functions to recast (6) as the optimization problem

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

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 $\mathcal{W} = \{w : \|w\|_1 = 1, \|w\|_0 \leq \eta\}$ 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 $\|w\|_2 \geq \epsilon$, we assume that

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

where $x = Sw$ and $B = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{T}$. 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 $w \in \{w : \|w\|_2 \geq \epsilon\}$. Problem (6) is equivalent to*

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

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

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

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

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

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 [11, Theorem 2], $f_3(w)$ is twice differentiable if $F_y(\bar{w},\bar{y}) = 0$ and $F_{yy}(\bar{w},\bar{y})$ is invertible. 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 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 $\|b_0\|^2\|b_1\|^2 - (b_0^T b_1)^2 \neq 0$, in order for $\bar{a}$ to be a real number, γ has to be small enough so that

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

The infimum can be attained because $\|b_0\|^2$ is bounded below by the assumption on input data and $\|b_0\|^2\|b_1\|^2 - (b_0^T b_1)^2 \leq \|b_0\|^2\|b_1\|^2$ 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, [11] are satisfied, implying that f_3 is twice differentiable on $\{w : \|w\|_2 \geq \epsilon\}$. 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*

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

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*

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

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 [7] for a detailed derivation.

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

The set of interest,

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

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

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

where Δ_1 is the 1-simplex, $\odot$ element-wise multiplication, and the second equality is obtained by a change of variable $u = \text{sign}(w) \odot z$.

Next, to find $\text{proj}_{\mathcal{W}}(|w|)$, we propose the following procedure, whose correctness is proved in Lemma 1.

- Order $|w|$ such that $|w_1| \geq |w_2| \geq \cdots \geq |w_m|$.
- Let

$$w^+_{1:\eta} \leftarrow \text{argmin}_{u_{1:\eta} \in \Delta_1} \||w_{1:\eta}| - u_{1:\eta}\|^2$$

$$w^+_{\eta+1:m} = 0.$$

Lemma 1 *Suppose $v \in \mathbb{R}^m$. Let K be any size k subset of $I = \{1, ..., m\}$ 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 $v_1 \geq v_2 \geq ... \geq v_m$. The claim is that the optimal $K_{opt} = \{1, 2, .., k\}$, 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

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

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

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

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 [9].

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 [10]. However our problem does not satisfy the assumptions of [10] because the gradient of loss function $f_3(w)$ is not globally Lipschitz. As shown in 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. Figure 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.

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

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

1: $\mathcal{W} = \{w : \|w\|_1 = 1, \|w\|_0 \leq \eta\}$
2: **while** not converged **do**
3: $\quad w^k \leftarrow \text{Proj}_{\mathcal{W}}(w^{k-1} - \delta_i \nabla_w f_3(w^{k-1}; \gamma, \eta))$
4: $\quad \text{loss}_i \leftarrow f_3(w^k; \gamma, \eta)$
5: Recover a, c, θ from w.
6:

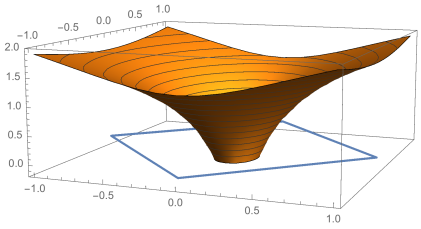


Fig. 1. 3D plot of the objective function in (8) for $w \in \mathbb{R}^2$, with constraint set $\|w\|_1 = 1, \|w\|_0 \leq 2$. Our goal is to find the minimum value of f_3 (yellow 3D plot) restricted to $\mathcal{W}$ (edges of the blue diamond).

Theorem 2 *Consider the optimization problem*

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

where f is the objective function f_3 in (8) and $\mathcal{W} = \{w : \|w\|_1 = 1, \|w\|_0 \leq \eta\}$ 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 $t \geq \underline{t}$, then

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

as $k \to \infty$.

Proof: This theorem is proved with the following lemmas:

- 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 $\epsilon \le \sqrt{2}/2$.
- 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.

We now present detailed statements and proofs for the lemmas used to establish the theorem.

Lemma 2 *Let $C = \mathbb{R}^n - \mathbb{B}_\epsilon(0)$ where $\mathbb{B}_\epsilon(0) = \{w : \|w\|_2 < \epsilon\}$ and $\epsilon \le \sqrt{2}/2$. 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*

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

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

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

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

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

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,

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

By the definition of projection,

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

Adding them together yields

$$f(w^+) + \frac{1}{2t}\|w - w^+\|^2 \le f(w) + \frac{15L(\epsilon)}{2}\|w^+ - w\|^2,$$

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

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 $\underline{t} \le t_k < \frac{1}{K}$, 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 $\underline{t} \le t_k < \frac{1}{K}$, 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

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

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

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

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 descent 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 $k \to \infty$, $A^k \to 0$.

4 Numerical Results

4.1 Selection for Multiple Time Series

Algorithmic Comparison using Simulated Data. In our first experiment, we show (1) that we can identify mean-reverting time series using simulated data and (2) that Algorithm 1 is faster than a standard approach that does not use partial minimization. We simulate five time series; four from an OU process with different μ and σ as specified in Table 1, and one is non-OU time series with $\sigma = .1$ (the fifth time series). All have $T = 500$ and $\Delta t = 0.01$. We use the first 70% of data for training and 30% for testing.

#	1	2	3	4
μ	1	4	1	4
σ	1	1	0.5	0.5
θ	0	1	1	0

Table 1
Model parameters used for the simulated OU time series

Figure 2 shows the comparison between Algorithm 1 and regular projected gradient descent on all unknowns (without partial minimization) starting from iteration 1.

Table 2 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 estimated parameter values are $\mu = 2.41, \sigma^2 = 0.09, \theta = 0.07$ with weight vector $w = [0.12, -0.11, 0.33, 0.31, -0.12]$. The model puts 64% of the weights into the pair of OU time series with $\sigma = 0.5$. When η is decreased to 4, the model drops the non-OU time series. As we further decrease η, it drops an OU time series with larger σ value. In other words, it favors OU time series with a lower σ value but remains relatively indifferent to μ values. Figure 3 plots those time series and the portfolio selected by the model.

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

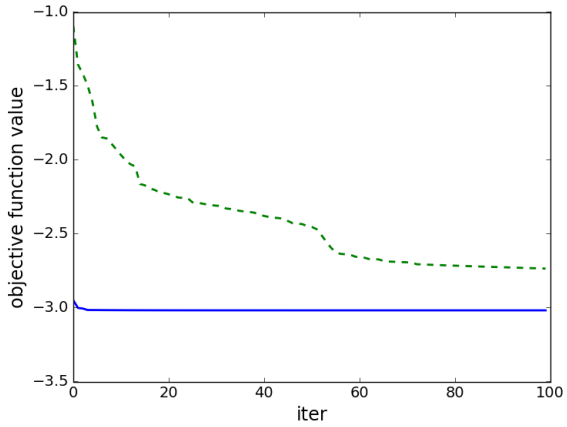


Fig. 2. Comparison of Algorithm 1 (solid) with standard projected gradient (dashed) using objective function values.

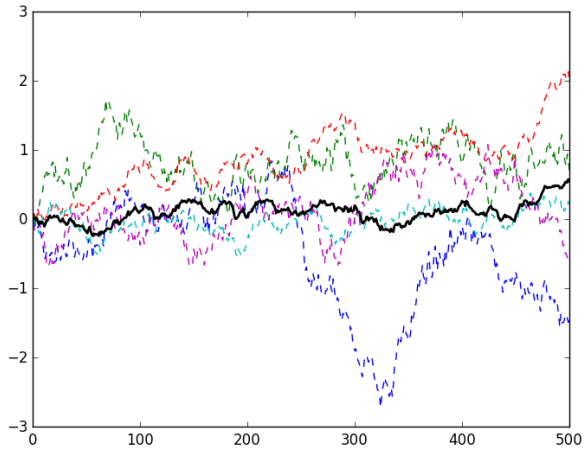


Fig. 3. Plot of simulated asset time series (dashed) and time series for final selected portfolio (solid).

η	μ	σ^2	θ	w	nll
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)

Table 2
Estimated parameters, weights, and nll. We set $\gamma = 0$ for top three rows, $\gamma = 0.5$ for bottom three rows.

Remark 5 *As noted in Remark 3, γ will drive $\bar{c}$ to be negative. If γ is large, the model many not find a feasible time series combination. 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, see Table 3. 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.

Groups	Assets (Tickers)
Precious Metals	GLD GDX, GDXJ, SLV, GG, ABX
Large Equities	GOOG, JNJ, NKE, MCD,
	SBUX, SPY, VIG, VO
Oil Companies	BP, COP, CVX, OIL, USO, VLO, XOM

Table 3
Asset Groups for Empirical Experiments

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 4. 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.

We would like to draw a connection to the pairs trading problem by looking at the top pair from each group observed from Table 4 with $\eta = 2$. In Figure 4, we plot the selected pairs which display similarity and synchrony in trend, which is usually desirable in pairs trading.

We also conducted experiments varying γ to promote larger μ. As summarized in Table 5, 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.

Remark 6 *Because of nonconvexity in objective function and constraints, in practice we may need to try different initializations to avoid local minima.*

4.2 Comparison with Pairs Trading

We compared our approach with that in chapter 2 of [1] on pairs trading. In [1], two assets are selected first, from which a portfolio is constructed as

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

where S_1 and S_2 are asset price time series. This "β-method" requires longing the first asset and shorting 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 maximizes 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 6, we can simply take the results from our model with $\eta = 2$ from Table 4.

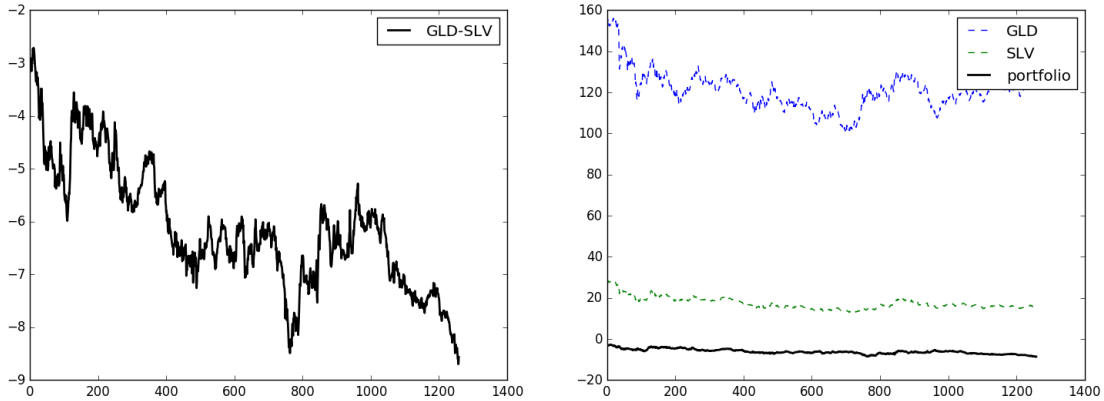
Assets	2	4	6	indiv. nll (train,test)
GLD	-0.17	-0.08	-0.07	0.77,0.44
GDX		-0.21	-0.29	0.05,-0.30
GDXJ			0.03	0.70, 0.38
SLV	0.83	0.44	0.30	-0.69, -1.0
GG			0.10	-0.04, -0.44
ABX		0.27	0.21	-0.24 ,-0.54
Port.	-1.48,-1.72	-1.95,2.12	-2.18,-2.35	
Assets	2	4	6	ind. nll (train,test)
GOOG				2.66, 3.06
JNJ		-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	
Assets	2	4	6	ind. nll (train,test)
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 4
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.

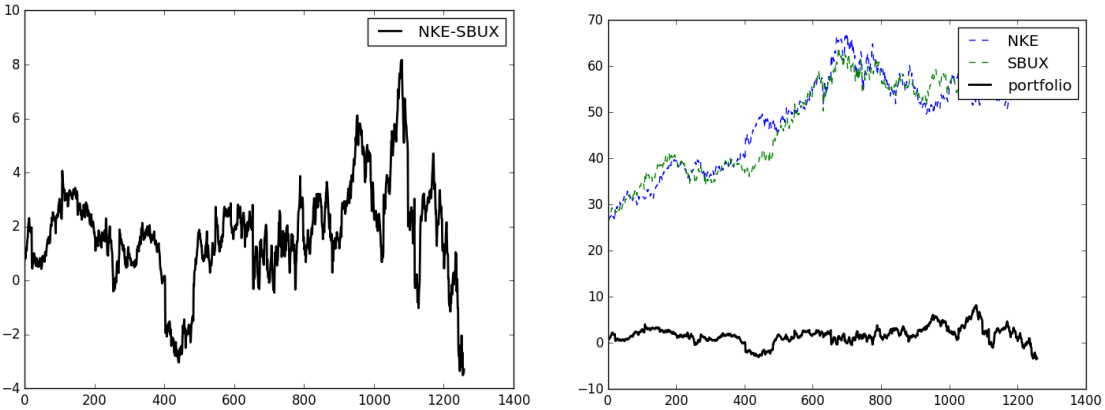
5 Concluding Remarks

We have solved a joint optimization problem for simultaneous portfolio selection and OU-fitting. We also in-

(a) Precious Metals



(b) Large-Cap Equities



(c) Oil Companies

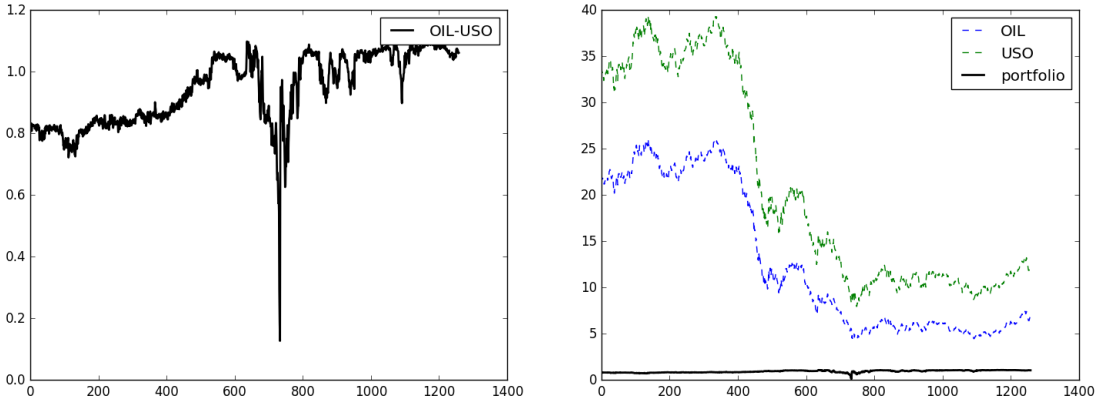


Fig. 4. Top pairs and corresponding portfolios from empirical price data.

Group	1				
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
Group	2				
γ	η	μ	σ^2	θ	nll (train, test)
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
Group	3				
γ	η	μ	σ^2	θ	nll (train, test)
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

Table 5
Model estimations with different γ and η for asset groups

(GLD, SLV)	β	portfolio weights
GLD - βSLV	3.68	[0.21, -0.79]
-βGLD + SLV	0.19	[-0.17,0.83]
our model	–	[-0.17,0.83]
(NKE, SBUX)		
NKE - βSBUX	0.61	[0.63, -0.37]
- βNKE + SBUX	0.52	[-0.33, 0.67]
our model	–	[-0.49, 0.51]
(OIL, USO)		
OIL - βUSO	0.67	[0.59, -0.41]
- βOIL + USO	1.42	[-0.59, 0.41]
our model	–	[-0.59, 0.41]

Table 6
Summary of portfolio weights from our model and method in (13) applied to different pairs.

corporated 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 [1], develops a convergence analysis for the algorithm, and provides a comparison analysis.

References

[1] T. Leung and X. Li, *Optimal Mean Reversion Trading: Mathematical Analysis and Practical Applications*, ser. Modern Trends in Financial Engineering. World Scientific, Singapore, 2016.

[2] ——, "Optimal mean reversion trading with transaction costs and stop-loss exit," *International Journal of Theoretical & Applied Finance*, vol. 18, no. 3, p. 15500, 2015.

[3] Y. Kitapbayev and T. Leung, "Optimal mean-reverting spread trading: nonlinear integral equation approach," *Annals of Finance*, vol. 13, no. 2, pp. 181–203, 2017.

[4] E. Gatev, W. Goetzmann, and K. Rouwenhorst, "Pairs trading: Performance of a relative-value arbitrage rule," *Review of Financial Studies*, vol. 19, no. 3, pp. 797–827, 2006.

[5] L. S. Ornstein and G. E. Uhlenbeck, "On the theory of the Brownian motion," *Physical Review*, vol. 36, pp. 823–841, 1930.

[6] A. d'Aspremont, "Identifying small mean-reverting portfolios," *Quant. Finance*, vol. 11, no. 3, pp. 351–364, 2011.

[7] J. Zhang, T. Leung, and A. Aravkin, "Mean reverting portfolios via penalized maximum likelihood estimation and optimization," in *Proceedings of the IEEE Conference on Decision and Control (CDC)*, 2018.

[8] M. Cuturi and A. d'Aspremont, "Mean reversion with a variance threshold," in *Proceedings of the international Conference on Machine learning*, vol. 28, no. 3. ACM, 2013, pp. 271–279.

[9] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l 1-ball for learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 272–279.

[10] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, 2013.

[11] B. M. Bell and J. V. Burke, "Algorithmic differentiation of implicit functions and optimal values," in *Advances in Automatic Differentiation*. Springer, 2008, pp. 67–77.