Mean-reverting portfolio optimization: higher-order parametric models with regularization

Brian Zhu

6 December 2021

Abstract

In this paper, we frame the problem of finding a mean-reverting portfolio from a basket of assets as maximizing the likelihood that the portfolio follows a stationary time series model. To this end, we develop algorithms to optimize asset weights for autoregressive models of arbitrary order for the portfolio value time series with unregularized, L1-regularized, and L2-regularized objective functions. Using real-world asset price data to fit portfolios, we evaluate the effects of model order and regularization with information criteria for model selection and proxies for mean reversion. The results of our numerical experiments suggest that most metrics considered show improvements in out-of-sample testing for models with moderate complexity and moderate regularization, and that L1 regularization promotes sparse mean-reverting portfolios.

1 Introduction

In finance, mean reversion—the long-term oscillation of an asset's price or value about a mean level—is a highly desirable property as it allows for statistical arbitrage, a family of trading strategies based on changing positions in the asset depending on its price or value relative to the mean level. The idea of finding stationary linear combinations of time series, or cointegration, was pioneered by the work of Engle & Granger (1987). Since then, there has emerged a vast body of literature on cointegration in econometrics and finance, part of which has been concerned with finding asset weights such that the resulting portfolio value optimizes some criterion related to mean reversion. Constraints and penalties have been incorporated to the optimization problem formulations as well. Cuturi & d'Aspremont (2013) identify three proxies for mean reversion: the predictability statistic of Box & Tiao (1977), the portmanteau statistic of Ljung & Box (1978), and the crossing statistic of Kedem & Yakowitz (1994). The work of Zhang et al. (2018) explores maximizing the likelihood that a portfolio follows the Ornstein-Uhlenbeck (OU) process of Ornstein & Uhlenbeck (1930).

Several statistical arbitrage trading strategies are described in Jurek & Yang (2007) and Leung & Li (2016). To potentially improve returns from trading on these strategies, Cuturi & d'Aspremont (2013) consider restricting the variance of the portfolio to be greater than a given threshold, and Zhang et al. (2018) add a penalty term that promotes higher values of the mean reversion parameter for the OU process, motivated by generating higher magnitudes of oscillations and more frequent oscillations in the portfolio time series, respectively. Both works also consider sparsity constraints for the portfolio weights.

Sparsity is another desirable property for a portfolio. Practically, the issues of transaction costs and execution error are less severe for sparser portfolios. Statistically, sparsity can be viewed as a form of regularization on the portfolio weights, which may mitigate issues of overfitting. Cuturi & d'Aspremont (2013) and Zhang et al. (2018) add L0 constraints in the mean-reverting setting to promote sparsity, while Brodie et al. (2009) adds an L1 constraint in the minimum-variance setting to promote sparsity.

Our paper observes that the AR(1) model is equivalent to a discretization of the OU process, and considers the problem of maximizing the likelihood that a portfolio follows an AR(p) model with homoskedastic normal errors where $p \in \mathbb{N}$ is arbitrary. Higher-order models may capture any long-term memory or autoregression in the asset prices, and may better characterize the asset price data depending on patterns in the data. As fitting higher-order models results in increased model complexity, we also consider regularizing the objective functions to counteract overfitting.

Our main contributions in this paper are algorithms to solve the maximum likelihood optimization problem for the aforementioned AR(p) models with no regularization, L1 regularization, and L2 regularization. The algorithms presented take advantage of the structure of the problem, specifically that the full optimization problems over $(N+p+2)$ parameters (where N is the number of assets in the basket, p is the AR order, and the remaining two parameters are the intercept and error variance for the AR model) can be reduced to just being over p parameters, highlighting the usefulness of our algorithms when N is large. Using real-world data sets, we use our algorithms to fit portfolios, varying the model order and regularization parameters. The fitted portfolios are evaluated with the Akaike and Bayes information criteria and the portmanteau and crossing statistics.

This paper is organized as follows. Section 2 describes the models and formulates the optimization problems. Section 3 motivates and presents the algorithms. Section 4 describes the setup and results of the numerical experiments, and Section 5 concludes.

2 Models and Problem Formulations

Let there be N assets. Denote the price of asset i at time t by $s_t = (s_{t,1}, \ldots, s_{t,N})^\top$. A portfolio of the assets is specified by a vector of weights $w = (w_1, \ldots, w_N)^\top$ normalized such that $\sum_{i=1}^N w_i = 1$, where w_i represents the proportion of an investor's wealth invested in asset i and negative weights represent short positions. The value of the portfolio at time t is then $v_t := w^\top s_t$.

The AR(p) model for a process $\{v_t\}$ is

$$v_t = \phi_0 + \phi_1 v_{t-1} + \cdots + \phi_p v_{t-p} + \sigma_\varepsilon \varepsilon_t$$

where we assume $\varepsilon_t \sim \mathcal{N}(0,1)$ i.i.d. Then

$$v_t - \phi_0 - \phi_1 v_{t-1} - \cdots - \phi_p v_{t-p} \sim \mathcal{N}(0, \sigma_\varepsilon^2) \text{ i.i.d.}$$

Replacing v_t with $w^\top s_t$ gives

$$w^\top s_t - \phi_0 - \phi_1 w^\top s_{t-1} - \cdots - \phi_p w^\top s_{t-p} \sim \mathcal{N}(0, \sigma_\varepsilon^2) \text{ i.i.d.}$$

Then one formulation of the likelihood function for w, ϕ_0, $\phi = (\phi_1, \ldots, \phi_p)$, and σ_ε given observed data $S = (s_1, \ldots, s_T)$ is

$$L(w, \phi_0, \phi, \sigma_\varepsilon \mid S) = \prod_{t=p+1}^{T} \frac{1}{(2\pi\sigma_\varepsilon^2)^{1/2}} \cdot \exp\left(-\frac{(w^\top s_t - \phi_0 - \phi_1 w^\top s_{t-1} - \cdots - \phi_p w^\top s_{t-p})^2}{2\sigma_\varepsilon^2}\right)$$

This is typically called the *conditional likelihood function* in the time series analysis literature, since the first p observations are utilized to condition the remaining observations and do not have their densities appear in the likelihood function. The conditional log-likelihood function is thus

$$\ell(w, \phi_0, \phi, \sigma_\varepsilon \mid S) = \sum_{t=p+1}^{T} \left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma_\varepsilon^2 - \frac{(w^\top s_t - \phi_0 - \phi_1 w^\top s_{t-1} - \cdots - \phi_p w^\top s_{t-p})^2}{2\sigma_\varepsilon^2}\right)$$

Note that maximizing the log-likelihood is equivalent to minimizing the negative log-likelihood. We flip the sign, drop constants, and scale by $(T-p)$ to arrive at an objective function:

$$f(w, \phi_0, \phi, \sigma_\varepsilon) = \frac{1}{2}\log \sigma_\varepsilon^2 + \frac{1}{2\sigma_\varepsilon^2(T-p)} \sum_{t=p+1}^{T} (w^\top s_t - \phi_0 - \phi_1 w^\top s_{t-1} - \cdots - \phi_p w^\top s_{t-p})^2$$

Let $S_{-k} \in \mathbb{R}^{(T-p)\times N}$ for $k \in \mathbb{Z}_{\geq 0}$ where $s_{p+1-k}, \ldots, s_{t-k}$ form the rows of S_{-k}. Then note that

$$\sum_{t=p+1}^{T} (w^\top s_t - \phi_0 - \phi_1 w^\top s_{t-1} - \cdots - \phi_p w^\top s_{t-p})^2 = \|S_{-0}w - \phi_0 1_{T-p} - \phi_1 S_{-1}w - \cdots - S_{-p}w\|_2^2$$

where 1_d denotes the all-ones vector in $\mathbb{R}^d$. As we are to eventually reduce the optimization problem from being over $(N+p+2)$ parameters to being over p parameters, we fix $\phi = (\phi_1, \ldots, \phi_p)$, define $D(\phi) := S_{-0} - \phi_1 S_{-1} - \cdots - \phi_p S_{-p}$, and note that

$$\|S_{-0}w - \phi_0 1_{T-p} - \phi_1 S_{-1}w - \cdots - S_{-p}w\|_2^2 = \|D(\phi)w - \phi_0 1_{T-p}\|_2^2$$

The objective function given ϕ then becomes

$$g_0(w, \phi_0, \sigma_\varepsilon \mid \phi) = \frac{1}{2}\log \sigma_\varepsilon^2 + \frac{1}{2\sigma_\varepsilon^2(T-p)}\|D(\phi)w - \phi_0 1_{T-p}\|_2^2$$

Regularizing w with tuning parameter γ, we first let $\tilde{S}_{-k}$ be S_{-k} standardized by asset and define $\tilde{D}(\phi) := \tilde{S}_{-0} - \phi_1 \tilde{S}_{-1} - \cdots - \phi_p \tilde{S}_{-p}$. The L1- and L2-regularized objective functions are then

$$g_1(w, \phi_0, \sigma_\varepsilon \mid \phi, \gamma) = \frac{1}{2}\log \sigma_\varepsilon^2 + \frac{1}{2\sigma_\varepsilon^2(T-p)}\left(\left\|\tilde{D}(\phi)w - \phi_0 1_{T-p}\right\|_2^2 + \gamma\|w\|_1^1\right)$$

$$g_2(w, \phi_0, \sigma_\varepsilon \mid \phi, \gamma) = \frac{1}{2}\log \sigma_\varepsilon^2 + \frac{1}{2\sigma_\varepsilon^2(T-p)}\left(\left\|\tilde{D}(\phi)w - \phi_0 1_{T-p}\right\|_2^2 + \gamma\|w\|_2^2\right)$$

The resulting inner optimization problems are

$$\underset{w,\phi_0,\sigma_\varepsilon}{\operatorname{argmin}} g_0(w, \phi_0, \sigma_\varepsilon \mid \phi) \text{ s.t. } 1_N^\top w = 1 \tag{2.1}$$

$$\underset{w,\phi_0,\sigma_\varepsilon}{\operatorname{argmin}} g_1(w, \phi_0, \sigma_\varepsilon \mid \phi, \gamma) \text{ s.t. } 1_N^\top w = 1 \tag{2.1-L1}$$

$$\underset{w,\phi_0,\sigma_\varepsilon}{\operatorname{argmin}} g_2(w, \phi_0, \sigma_\varepsilon \mid \phi, \gamma) \text{ s.t. } 1_N^\top w = 1 \tag{2.1-L2}$$

Then the outer objective functions are

$$\begin{aligned}
f_0(\phi) &= g_0(w^\star, \phi_0^\star, \sigma_\varepsilon^\star \mid \phi) \text{ where } (w^\star, \phi_0^\star, \sigma_\varepsilon^\star) \text{ solves (2.1) given } \phi \\
f_1(\phi \mid \gamma) &= g_1(w^\star, \phi_0^\star, \sigma_\varepsilon^\star \mid \phi) \text{ where } (w^\star, \phi_0^\star, \sigma_\varepsilon^\star) \text{ solves (2.1-L1) given } (\phi, \gamma) \\
f_2(\phi \mid \gamma) &= g_2(w^\star, \phi_0^\star, \sigma_\varepsilon^\star \mid \phi) \text{ where } (w^\star, \phi_0^\star, \sigma_\varepsilon^\star) \text{ solves (2.1-L2) given } (\phi, \gamma)
\end{aligned}$$

so the outer optimization problems are

$$\underset{\phi}{\operatorname{argmin}} f_0(\phi) \tag{2.2}$$

$$\underset{\phi}{\operatorname{argmin}} f_1(\phi \mid \gamma) \tag{2.2-L1}$$

$$\underset{\phi}{\operatorname{argmin}} f_2(\phi \mid \gamma) \tag{2.2-L2}$$

3 Algorithms

For each of the unregularized, L1-regularized, and L2-regularized settings, we derive a closed-form solution or provide an efficient method to compute an apprxoimate solution for the global minima of the inner optimization problems. The outer optimization problems are nonconvex in general and can be solved with an algorithm of choice, typically L-BFGS-B or simulated annealing.

3.1 No Regularization

We first observe that (2.1), reproduced below for convenience, has a closed-form solution.

$$\underset{w,\phi_0,\sigma_\varepsilon}{\operatorname{argmin}} \left(\frac{1}{2} \log \sigma_\varepsilon^2 + \frac{1}{2\sigma_\varepsilon^2 (T-p)} \| D(\phi) w - \phi_0 1_{T-p} \|_2^2 \right) \text{ s.t. } 1_N^\top w = 1$$

Lemma 3.1. Assuming $D(\phi)^\top D(\phi)$ is invertible, define quantities and vectors

$$\begin{aligned}
\alpha &= 1_N^\top (D(\phi)^\top D(\phi))^{-1} D(\phi)^\top 1_{T-p} \\
\beta &= 1_N^\top (D(\phi)^\top D(\phi))^{-1} 1_N \\
x &= (D(\phi)^\top D(\phi))^{-1} D(\phi)^\top 1_{T-p} \\
y &= (D(\phi)^\top D(\phi))^{-1} 1_N \\
z &= D(\phi)^\top 1_{T-p}
\end{aligned}$$

Then the global minimum for (2.1) is attained by

$$\phi_0^\star = \frac{\langle y, z\rangle}{\alpha\langle y, z\rangle + \beta(T - p - \langle x, z\rangle)}$$
$$w^\star = \phi_0^\star x + \left(\frac{1 - \phi_0^\star \alpha}{\beta}\right) y$$
$$\sigma_\varepsilon^\star = \frac{\|D(\phi)w^\star - \phi_0^\star 1_{T-p}\|_2}{(T-p)^{1/2}}$$

Proof. Note that w and ϕ_0 do not appear in (2.1) outside of the squared norm expression, so

$$(w^\star, \phi_0^\star) = \underset{w,\phi_0}{\operatorname{argmin}} \|D(\phi)w - \phi_0 1_{T-p}\|_2^2 \text{ s.t. } 1_N^\top w = 1$$

We form the Lagrangian with multiplier 2λ for the constraint:

$$\begin{aligned}\mathcal{L}(w, \phi_0, \lambda) &= \|D(\phi)w - \phi_0 1_{T-p}\|_2^2 - 2\lambda(1_N^\top w - 1) \\ &= w^\top D(\phi)^\top D(\phi)w - 2\phi_0 w^\top D(\phi)^\top 1_{T-p} + \phi_0^2(T-p) - 2\lambda(1_N^\top w - 1)\end{aligned}$$

The gradients of $\mathcal{L}(w, \phi_0, \lambda)$ with respect to each argument are

$$\begin{aligned}\nabla_w \mathcal{L}(w, \phi_0, \lambda) &= 2D(\phi)^\top D(\phi)w - 2\phi_0 D(\phi)^\top 1_{T-p} - 2\lambda 1_N \\ \nabla_{\phi_0} \mathcal{L}(w, \phi_0, \lambda) &= -2w^\top D(\phi)^\top 1_{T-p} + 2\phi_0(T-p) \\ \nabla_\lambda \mathcal{L}(w, \phi_0, \lambda) &= 1_N^\top w - 1\end{aligned}$$

and the Hessians are $2D(\phi)^\top D(\phi)$, $2(T-p)$, and 0, respectively, so $\mathcal{L}(w, \phi_0, \lambda)$ is convex in each argument. Then it suffices to set the gradients to zero to minimize $\mathcal{L}$, giving optimality conditions

$$w^\star = (D(\phi)^\top D(\phi))^{-1}(\phi_0^\star D(\phi)^\top 1_{T-p} + \lambda^\star 1_N) \tag{3.1.1}$$
$$\phi_0^\star = \frac{(D(\phi)w^\star)^\top 1_{T-p}}{T-p} \tag{3.1.2}$$
$$1 = 1_N^\top w^\star \tag{3.1.3}$$

Substituting (3.1.1) into (3.1.3), we have

$$\begin{aligned}1 &= 1_N^\top (D(\phi)^\top D(\phi))^{-1}(\phi_0^\star D(\phi)^\top 1_{T-p} + \lambda^\star 1_N) \\ &= \phi_0^\star 1_N^\top (D(\phi)^\top D(\phi))^{-1} D(\phi)^\top 1_{T-p} + \lambda^\star 1_N^\top (D(\phi)^\top D(\phi))^{-1} 1_N \\ \lambda^\star &= \frac{1 - \phi_0^\star 1_N^\top (D(\phi)^\top D(\phi))^{-1} D^\top 1_{T-p}}{1_N^\top (D(\phi)^\top D(\phi))^{-1} 1_N}\end{aligned} \tag{3.1.4}$$

Substituting (3.1.4) into (3.1.1), we have

$$w^\star = (D(\phi)^\top D(\phi))^{-1}\left(\phi_0^\star D(\phi)^\top 1_{T-p} + \left(\frac{1 - \phi_0^\star 1_N^\top (D(\phi)^\top D(\phi))^{-1} D(\phi)^\top 1_{T-p}}{1_N^\top (D(\phi)^\top D(\phi))^{-1} 1_N}\right) 1_N\right)$$

$$= \phi_0^\star (D(\phi)^\top D(\phi))^{-1} D^\top 1_{T-p} + \left(\frac{1 - \phi_0^\star 1_N^\top (D(\phi)^\top D(\phi))^{-1} D(\phi)^\top 1_T}{1_N^\top (D(\phi)^\top D(\phi))^{-1} 1_N} \right) (D(\phi)^\top D(\phi))^{-1} 1_N$$

With α, β, x, y, and z defined in the statement of the lemma, (3.1.1) and (3.1.2) become

$$w^\star = \phi_0^\star x + \left(\frac{1 - \alpha \phi_0^\star}{\beta} \right) y \tag{3.1.5}$$

$$\phi_0^\star = \langle w^\star, z \rangle / (T-p) \tag{3.1.6}$$

Substituting (3.1.6) into (3.1.5), we have

$$\begin{aligned} w^\star = \frac{\langle w^\star, z \rangle}{T-p} x + \left(\frac{1 - \frac{\alpha}{T-p} \langle w^\star, z \rangle}{\beta} \right) y &= \frac{\langle w^\star, z \rangle}{T-p} x - \left(\frac{1}{\beta} - \frac{\alpha \langle w^\star, z \rangle}{\beta (T-p)} \right) y \\ &= \langle w^\star, z \rangle \left(\frac{x}{T-p} - \frac{\alpha y}{\beta (T-p)} \right) + \frac{y}{\beta} \end{aligned}$$

Taking the inner product with z on both sides gives

$$\begin{aligned} \langle w^\star, z \rangle &= \langle w^\star, z \rangle \left\langle \frac{x}{T-p} - \frac{\alpha y}{\beta (T-p)}, z \right\rangle + \frac{\langle y, z \rangle}{\beta} \\ &= \frac{\langle y, z \rangle}{\beta \left(1 - \left\langle \frac{x}{T-p} - \frac{\alpha y}{\beta (T-p)}, z \right\rangle \right)} \end{aligned} \tag{3.1.7}$$

Substituting (3.1.7) back into (3.1.5), we have

$$\phi_0^\star = \frac{\langle w^\star, z \rangle}{T-p} = \frac{\langle y, z \rangle}{\beta (T-p) \left(1 - \left\langle \frac{x}{T-p} - \frac{\alpha y}{\beta (T-p)}, z \right\rangle \right)} = \frac{\langle y, z \rangle}{\alpha \langle y, z \rangle + \beta (T - p - \langle x, z \rangle)}$$

as desired. The expression for $w^\star$ follows directly from (3.1.5). Now let $\xi = \log \sigma_\varepsilon^2$. Then

$$\begin{aligned} &\operatorname*{argmin}_{w, \phi_0, \sigma_\varepsilon} \left(\frac{1}{2} \log \sigma_\varepsilon^2 + \frac{1}{2 \sigma_\varepsilon^2 (T-p)} \| D(\phi) w - \phi_0 1_{T-p} \|_2^2 \right) \text{ s.t. } 1_N^\top w = 1 \\ &= \operatorname*{argmin}_{\xi} \left(\frac{\xi}{2} + \frac{\exp(-\xi)}{2(T-p)} \| D(\phi) w^\star - \phi_0^\star 1_{T-p} \|_2^2 \right) \end{aligned} \tag{3.1.8}$$

Note that ξ and $\exp(-\xi)$ are convex, so to minimize the objective function in (3.1.8), it suffices to set its gradient with respect to ξ to zero, giving the optimality condition

$$\begin{aligned} \frac{1}{2} - \frac{\| D(\phi) w^\star - \phi_0^\star 1_{T-p} \|_2^2}{2(T-p)} \exp(-\xi) &= 0 \\ \sigma_\varepsilon^\star = (\exp \xi)^{1/2} &= \frac{\| D(\phi) w^\star - \phi_0^\star 1_{T-p} \|_2}{(T-p)^{1/2}} \end{aligned}$$

as desired. □

By Lemma 3.1, we have an efficient oracle for $f_0(\phi)$, so we can approximately optimize f_0 over ϕ with an algorithm of choice. As the resulting AR(p) model should be stationary, it may be helpful to impose bounds of $[-1, 1]$ on each component ϕ_i to narrow the search region for ϕ, although these bounds alone do not guarantee stationarity. In our implementation for the numerical experiments in Section 4, we impose these bounds and use the L-BFGS-B algorithm.

3.2 L1 Regularization

Taking the same approach as Section 3.1, we would like to first solve the inner optimization problem. We reproduce (2.1-L1) below.

$$\operatorname*{argmin}_{w,\phi_0,\sigma_\varepsilon} \left(\frac{1}{2} \log \sigma_\varepsilon^2 + \frac{1}{2\sigma_\varepsilon^2 (T-p)} \left(\left\| \tilde{D}(\phi) w - \phi_0 1_{T-p} \right\|_2^2 + \gamma \|w\|_1^1 \right) \right) \text{ s.t. } 1_N^\top w = 1$$

Similarly to (2.1), we can first compute the optimal w and ϕ_0 by considering

$$\operatorname*{argmin}_{w,\phi_0} \left(\left\| \tilde{D}(\phi) w - \phi_0 1_{T-p} \right\|_2^2 + \gamma \|w\|_1^1 \right) \text{ s.t. } 1_N^\top w = 1 \tag{3.2.1}$$

One can verify that the objective function in (3.2.1) is convex in w and ϕ_0. However, (3.2.1) does not have a closed-form solution in general, so we compute an approximate solution with a projected subgradient method.

The gradient of the objective function in (3.2.1) with respect to ϕ_0 is

$$\nabla_{\phi_0} \mathcal{L} = -2 w^\top \tilde{D}(\phi)^\top 1_{T-p} + 2\phi_0 (T-p)$$

giving the optimality condition

$$\phi_0^\star = \frac{(\tilde{D}(\phi) w^\star)^\top 1_{T-p}}{T-p}$$

Then we can write

$$\begin{aligned}
&\operatorname*{argmin}_{w,\phi_0} \left(\left\| \tilde{D}(\phi) w - \phi_0 1_{T-p} \right\|_2^2 + \gamma \|w\|_1^1 \right) \text{ s.t. } 1_N^\top w = 1 \\
&= \operatorname*{argmin}_{w} \left(\left\| \tilde{D}(\phi) w - \frac{(\tilde{D}(\phi) w)^\top 1_{T-p}}{T-p} 1_{T-p} \right\|_2^2 + \gamma \|w\|_1^1 \right) \text{ s.t. } 1_N^\top w = 1 \\
&= \operatorname*{argmin}_{w} \left(w^\top \tilde{D}(\phi)^\top \tilde{D}(\phi) w - \frac{(w^\top \tilde{D}(\phi)^\top 1_{T-p})^2}{T-p} + \gamma \|w\|_1^1 \right) \text{ s.t. } 1_N^\top w = 1
\end{aligned} \tag{3.2.2}$$

A subgradient oracle for the objective function in (3.2.2) is then

$$\textsc{Subgrad}(w; S, \phi, \gamma) = 2\tilde{D}(\phi)^\top \tilde{D}(\phi) w - \frac{2(w^\top \tilde{D}(\phi)^\top 1_{T-p})(\tilde{D}(\phi)^\top 1_{T-p})}{T-p} + \gamma \operatorname{sgn} w$$

Lemma 3.2. Let $b \in \mathbb{R}^N$ and $C = \{a \in \mathbb{R}^N : 1_N^\top a = 1\}$. Define $a_i^\star = b_i - (\sum_{i=1}^N b_i - 1)/N$ for $i = 1, \ldots, N$. Then $a^\star = (a_1^\star, \ldots, a_N^\star) = \mathrm{proj}_C(b)$.

Proof. Note that

$$\mathrm{proj}_C(b) = \operatorname*{argmin}_a \frac{1}{2}\|a - b\|_2^2 \text{ s.t. } 1_N^\top a = 1$$

We form the Lagrangian with Lagrange multiplier λ:

$$\mathcal{L}(a, \lambda) = \frac{1}{2}\|a - b\|_2^2 + \lambda(1_N^\top a - 1)$$

The gradients with respect to each argument are

$$\nabla_a \mathcal{L}(a, \lambda) = a - b + \lambda 1_N$$
$$\nabla_\lambda \mathcal{L}(a, \lambda) = 1_N^\top a - 1$$

One can verify that $\mathcal{L}(a, \lambda)$ is convex in each of its arguments, giving optimality conditions

$$a^\star = b - \lambda^\star 1_N \iff a_i^\star = b_i - \lambda^\star \tag{3.2.3}$$

$$1 = 1_N^\top a \iff 1 = \sum_{i=1}^N a_i^\star \tag{3.2.4}$$

Substituting (3.2.3) into (3.2.4), we have

$$1 = \sum_{i=1}^N b_i - N\lambda^\star$$

$$\lambda^\star = \frac{\sum_{i=1}^N b_i - 1}{N} \tag{3.2.5}$$

Substituting (3.2.5) into (3.2.3) gives the lemma. □

By Lemma 3.2, a projection oracle for the set $\{w \in \mathbb{R}^N : 1_N^\top w = 1\}$ is

$$\textsc{Project}(b) = b - \frac{\sum_{i=1}^n b_i - 1}{N} 1_N$$

We can now provide a projected subgradient method to compute approximately optimal w and ϕ_0, given an initial point w_0, step sizes $(\eta_1, \ldots, \eta_K)$, and hard-thresholding operator h_θ for $\theta > 0$.

Algorithm PSM

Input: w_0, $(\eta_1, \ldots, \eta_K)$, S, ϕ, γ

for $k = 1, \ldots, K$ **do**

  $w_k = \textsc{Project}(w_{k-1} - \eta_k \cdot \textsc{Subgrad}(w_{k-1}; S, \phi, \gamma))$

end for

Output: $h_\theta(w_K)/(\sum_{i=1}^N h_\theta(w_K)_i)$

Common step sizes include a constant step size ($\eta_k = \eta$ for all k), decreasing step sizes (e.g. $\eta_k = \eta_0/k$), or step sizes inspired by Barzilai & Borwein (1988), which we use in our implementation:

$$\eta_k = \frac{\left|(w_k - w_{k-1})^\top (\textsc{Subgrad}(w_k) - \textsc{Subgrad}(w_{k-1}))\right|}{\|\textsc{Subgrad}(w_k) - \textsc{Subgrad}(w_{k-1})\|_2^2}$$

The hard-thresholding operator h_θ sets all components less than $|\theta|$ to 0, and the hard-thresholded vector is re-normalized before being outputted. A small value of θ (e.g. 10^{-3} in our implementation) achieves the effect of sparsity promoted by L1 regularization that the iterative step alone might only approximate. Using an argument from the proof of Lemma 3.1, we can get a similar expression for the optimal σ_ε. Then an approximate solution to (2.1-L1) given ϕ and γ is

$$w^\star = \mathrm{PSM}(w_0; S, \phi, \gamma, K)$$

$$\phi_0^\star = \frac{\tilde{D}(\phi) w^\star)^\top 1_{T-p}}{T-p}$$

$$\sigma_\varepsilon^\star = \left(\frac{\left\| \tilde{D}(\phi) w - \phi_0 1_{T-p} \right\|_2^2 + \gamma \|w\|_1^1}{T-p} \right)^{1/2}$$

so we have an efficient oracle for $f_1(\phi \mid \gamma)$ and can then approximately optimize f_1 over ϕ with an algorithm of choice. In practice, the black-box optimization methods (found in the `scipy.optimize` package for example) are unstable here, often converging to poor local optima, so we present an algorithm to approximately optimize f_1 over ϕ using simulated annealing. We compare its performance against `scipy.optimize` methods in Section 4. The algorithm requires an initial point $\phi^{(0)}$, perturbation variance σ_δ, and decreasing acceptance parameters $(\alpha_1, \ldots, \alpha_M)$. This algorithm proposes new points $\phi^{(\text{prop})}$ via perturbation, and accepts them either immediately or with probability $\exp\big((f_1(\phi^{(\text{curr})}) - f_1(\phi^{(\text{prop})}))/\alpha_i\big)$ depending on if it yields a lower objective function value or not. The decreasing α_i help to prevent early iterates from getting stuck in poor local optima.

Algorithm SA

Input: $\phi^{(0)}$, σ_δ, $(\alpha_1, \ldots, \alpha_M)$, S, ϕ, γ

$\phi^{(\text{curr})} \leftarrow \phi^{(0)}$
for $i = 1, \ldots, M$ **do**
  $\phi^{(\text{prop})} \leftarrow \phi^{(\text{curr})} + \sigma_\delta \delta_i$ where $\delta_i \sim \mathcal{N}(0, I_p)$
  if $f_1(\phi^{(\text{prop})}) < f_1(\phi^{(\text{curr})})$ **then**
    $\phi^{(\text{curr})} \leftarrow \phi^{(\text{prop})}$
  else
    if $u < \exp\big((f_1(\phi^{(\text{curr})}) - f_1(\phi^{(\text{prop})}))/\alpha_i\big)$ where $u \sim \mathcal{U}(0,1)$ **then**
      $\phi^{(\text{curr})} \leftarrow \phi^{(\text{prop})}$
    end if
  end if
end for
Output: $\phi^{(\text{curr})}$

3.3 L2 Regularization

Like (2.1), (2.1-L2), reproduced below for convenience, has a closed-form solution as well.

$$\underset{w,\phi_0,\sigma_\varepsilon}{\operatorname{argmin}} \left(\frac{1}{2}\log\sigma_\varepsilon^2 + \frac{1}{2\sigma_\varepsilon^2(T-p)} \left(\left\| \tilde{D}(\phi)w - \phi_0 1_{T-p} \right\|_2^2 + \gamma\|w\|_2^2 \right) \right) \text{ s.t. } 1_N^\top w = 1$$

Lemma 3.3. Assuming $\tilde{D}(\phi)^\top \tilde{D}(\phi)$ is invertible, define quantities and vectors

$$\begin{aligned}
\alpha &= 1_N^\top (\tilde{D}(\phi)^\top \tilde{D}(\phi) + \gamma I_N)^{-1} \tilde{D}(\phi)^\top 1_{T-p} \\
\beta &= 1_N^\top (\tilde{D}(\phi)^\top \tilde{D}(\phi) + \gamma I_N)^{-1} 1_N \\
x &= (\tilde{D}(\phi)^\top \tilde{D}(\phi) + \gamma I_N)^{-1} \tilde{D}(\phi)^\top 1_{T-p} \\
y &= (\tilde{D}(\phi)^\top \tilde{D}(\phi) + \gamma I_N)^{-1} 1_N \\
z &= \tilde{D}(\phi)^\top 1_{T-p}
\end{aligned}$$

Then the global minimum for (2.1) is attained by

$$\begin{aligned}
\phi_0^\star &= \frac{\langle y, z\rangle}{\alpha\langle y, z\rangle + \beta(T - p - \langle x, z\rangle)} \\
w^\star &= \phi_0^\star x + \left(\frac{1 - \phi_0^\star \alpha}{\beta} \right) y \\
\sigma_\varepsilon^\star &= \left(\frac{\left\| \tilde{D}(\phi)w^\star - \phi_0^\star 1_{T-p} \right\|_2^2 + \gamma\|w\|_2^2}{(T-p)} \right)^{1/2}
\end{aligned}$$

Proof. Similar to the proof of Lemma 3.1. □

By Lemma 3.3, we have an efficient oracle for $f_2(\phi)$, so we can then approximately optimize f_2 over ϕ with an algorithm of choice. In our implementation for the numerical experiments in Section 4, we impose bounds of $[-1, 1]$ for each component ϕ_i and use the L-BFGS-B algorithm.

4 Numerical Experiments and Results

4.1 Experimental Setup

We consider two baskets of assets to build portfolios from: the top 20 currencies by trading volume[1] that have available data on Refinitiv and the top 20 oil/gas companies by market capitalization[2] that have available data on Yahoo Finance. Individual assets in these sectors tend to exhibit moderate mean-reverting behavior, so our algorithms, based on the likelihood of stationary processes, should be well-suited for building portfolios of assets from these sectors.

[1] Based on https://en.wikipedia.org/wiki/Template:Most_traded_currencies. The US dollar is used as the base.
[2] Based on https://companiesmarketcap.com/oil-gas/largest-oil-and-gas-companies-by-market-cap/.

For each asset, we use daily time series of its price or value for the past 10 years (around autumn 2011 to autumn 2021). We consider rolling windows of 210 observations for model fitting and 42 observations for out-of-sample testing of the model (i.e. $T_{\text{train}} = 210$ and $T_{\text{test}} = 42$), corresponding to 10 months for training and 2 months for testing. We then roll the window by 42 observations, which corresponds to updating the portfolio every 2 months based on the past 10 months of data, for around 45 total windows. In our experiments, we either vary the model order p or the regularization parameter γ, and collect the metrics described in the following sections.

4.2 Information Criteria

We consider two likelihood-based measures of model fit: the Akaike and Bayes information criterion (AIC and BIC, respectively). With $\ell(w^\star, \phi_0^\star, \phi^\star, \sigma_\varepsilon^\star \mid S)$ from Section 2, they are defined as

$$\begin{aligned}\text{AIC} &= -2\ell(w^\star, \phi_0^\star, \phi^\star, \sigma_\varepsilon^\star \mid S) + 2(p+2) \\ \text{BIC} &= -2\ell(w^\star, \phi_0^\star, \phi^\star, \sigma_\varepsilon^\star \mid S) + (p+2)\log(T-p)\end{aligned}$$

noting that $w^\star$ controls the series that the likelihood is taken over and thus should not be penalized in the information criteria. Since the regularization is over $w^\star$, the above expressions are suitable measures of model fit for the regularized cases as well.

4.3 Mean Reversion Proxies

We also consider two mean-reversion proxies. The out-of-sample *portmanteau statistic* for an AR(p) model of centered observations $(\dot{v}_1, \ldots, \dot{v}_{T_{\text{test}}})$, proposed by Ljung & Box (1978), is

$$\Pi_p = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{\hat{\mathbb{E}}[\dot{v}_t\dot{v}_{t-p}]}{\hat{\mathbb{E}}[\dot{v}_t\dot{v}_{t-0}]}\right)^2$$

where $\hat{\mathbb{E}}[\dot{v}_t\dot{v}_{t-p}]$ is defined to be

$$\hat{\mathbb{E}}[\dot{v}_t\dot{v}_{t-p}] = \frac{1}{T_{\text{test}}-p}\sum_{t=p+1}^{T_{\text{test}}}\dot{v}_t\dot{v}_{t-p}$$

A low portmanteau statistic suggests that the observations resembles white noise well. The sample *crossing statistic*, proposed by Kedem & Yakowitz (1994), of centered observations $(\dot{v}_1, \ldots, \dot{v}_{T_{\text{test}}})$ is

$$\chi = \frac{1}{T_{\text{test}}-1}\sum_{t=2}^{T_{\text{test}}}\mathbb{I}[\dot{v}_t\dot{v}_{t-1} < 0]$$

and measures the rate at which the observed series crosses its mean. As our means are fit parametrically, for both statistics, we center the out-of-sample portfolio series v_t with $\dot{v}_t = v_t - \mu^\star$ where $\mu^\star := \phi_0^\star/(1-\phi_1^\star-\cdots-\phi_p^\star)$, the mean of the fitted AR(p) model, assuming that it is stationary. As there may be windows where the statistics have extreme outlying values over all p or γ, we consider both the raw metrics and metrics standardized at each time window across the grids for p and γ.

4.4 Results

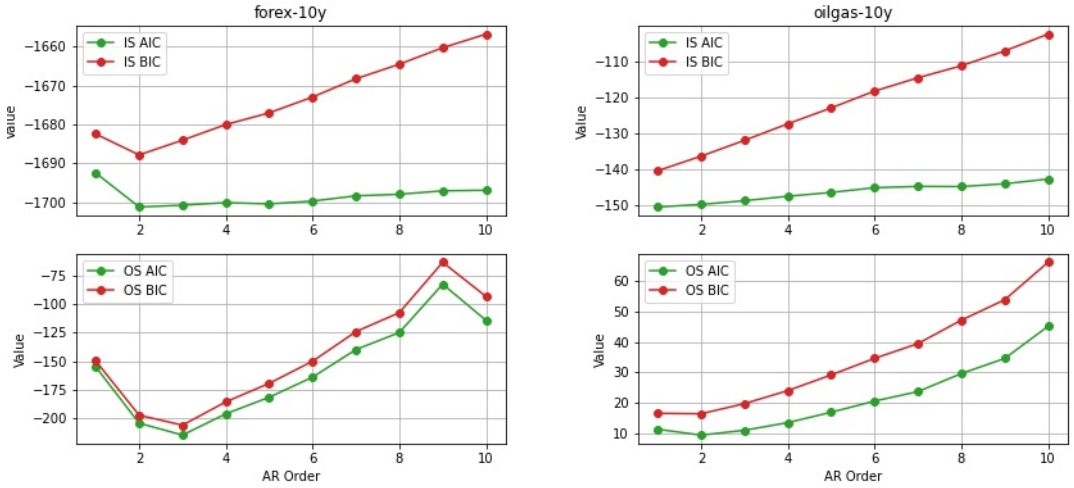

Figure 4.1. In-sample and out-of-sample AIC and BIC against model order on both data sets.

We first present the results for information criteria (IC). Figure 4.1 plots the IC for the training (in-sample) and testing (out-of-sample) periods against model order, with `forex` denoting the currency data set and `oilgas` denoting the oil/gas data set. While the in-sample IC tend to monotonically increase, indicating poorer fit, the out-of-sample IC tend to first decrease before increasing as model order increases, suggest that model orders of around 2-4 better fit the data based on the IC.

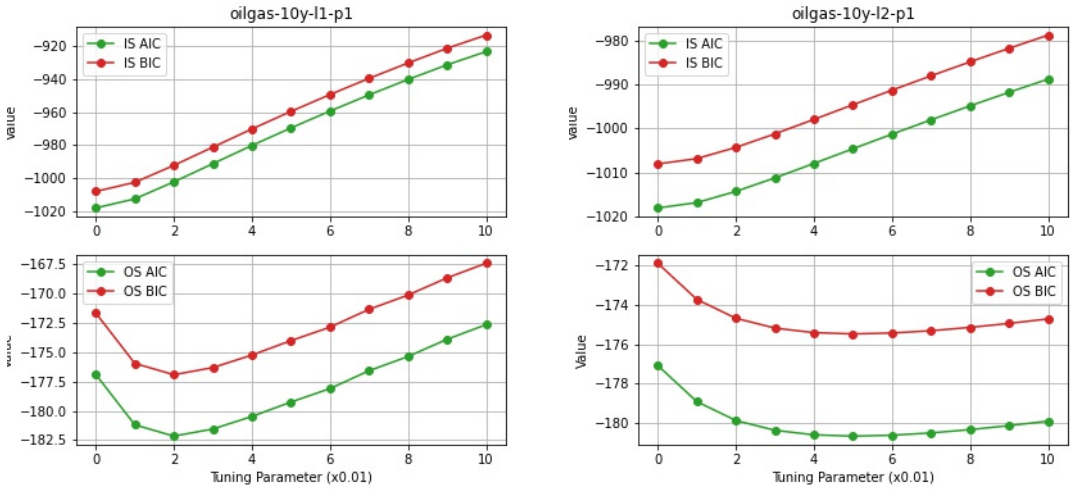

Figure 4.2. In-sample and out-of-sample AIC and BIC against L1 and L2 regularization parameters on the oil/gas data set with model order $p = 1$.

Figure 4.2 plots the in-sample and out-of-sample IC against L1 and L2 regularization parameters for the oil/gas data set with model order $p = 1$. This suggests that moderate L1 regularization and L2 regularization improves the out-of-sample IC, with moderate L1 regularization having slightly better performance than L2 regularization. The same trends generally hold for model order $p = 5$, as shown in Figure 4.3.

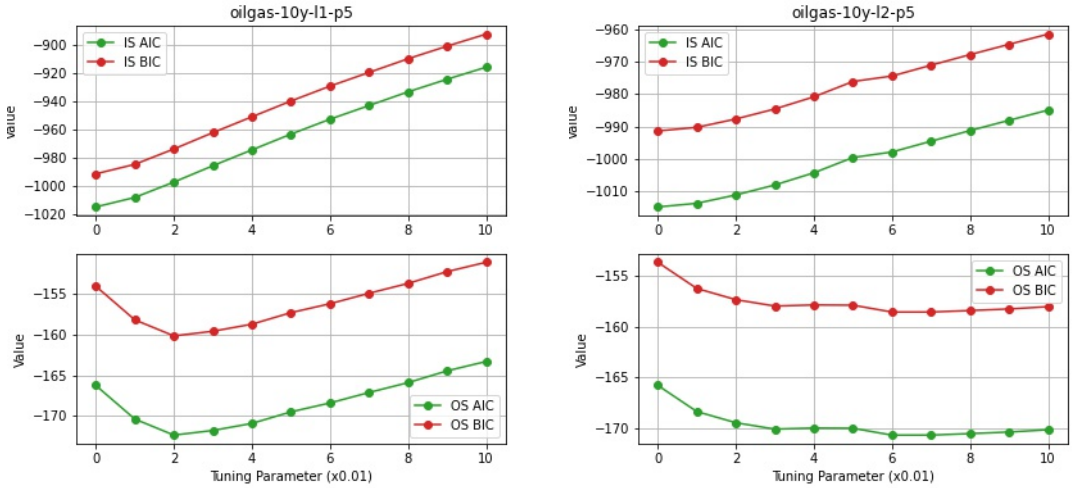

Figure 4.3. In-sample and out-of-sample AIC and BIC against L1 and L2 regularization parameters on the oil/gas data set with model order $p = 5$.

We then present the results for the mean reversion proxies. In Figure 4.4, it appears that the in-sample and out-of-sample standardized portmanteau statistic decrease in model order for both data sets, suggesting that the higher-order models tend to resemble white noise more than lower-order models. There are no meaningful trends in the standardized crossing statistic plots.

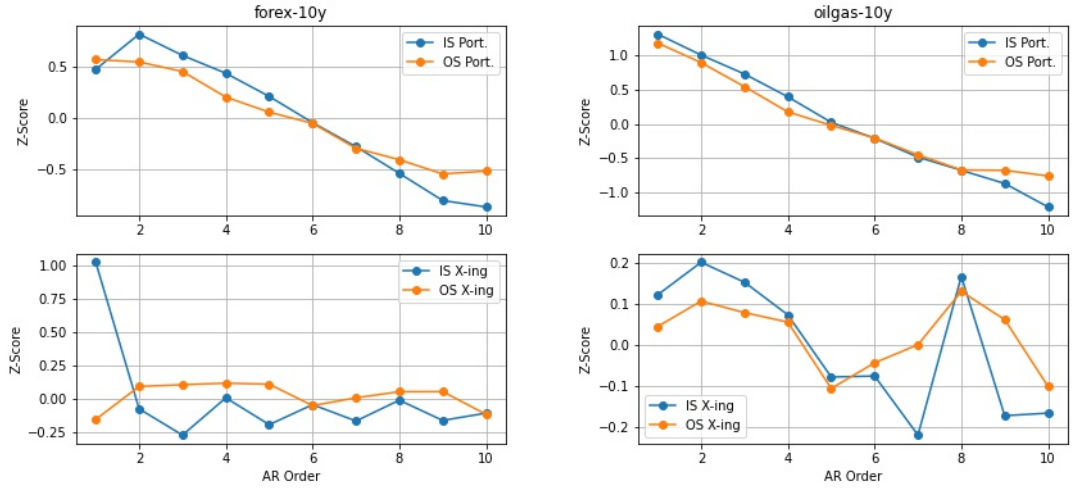

Figure 4.4. In-sample and out-of-sample standardized portmanteau and crossing statistics against model order on both data sets.

Figures 4.5 and 4.6 plot the standardized statistics against the L1 and L2 regularization parameters with model orders of $p = 1$ and $p = 5$ on the oil/gas data set. For L1 regularization, the out-of-sample portmanteau statistic appears to increase in the model order, but for L2 regularization, it tends to increase first before decreasing. There are still no meaningful trends in the crossing statistic plots, which could suggest that the crossing statistic may be a poor metric for evaluation.

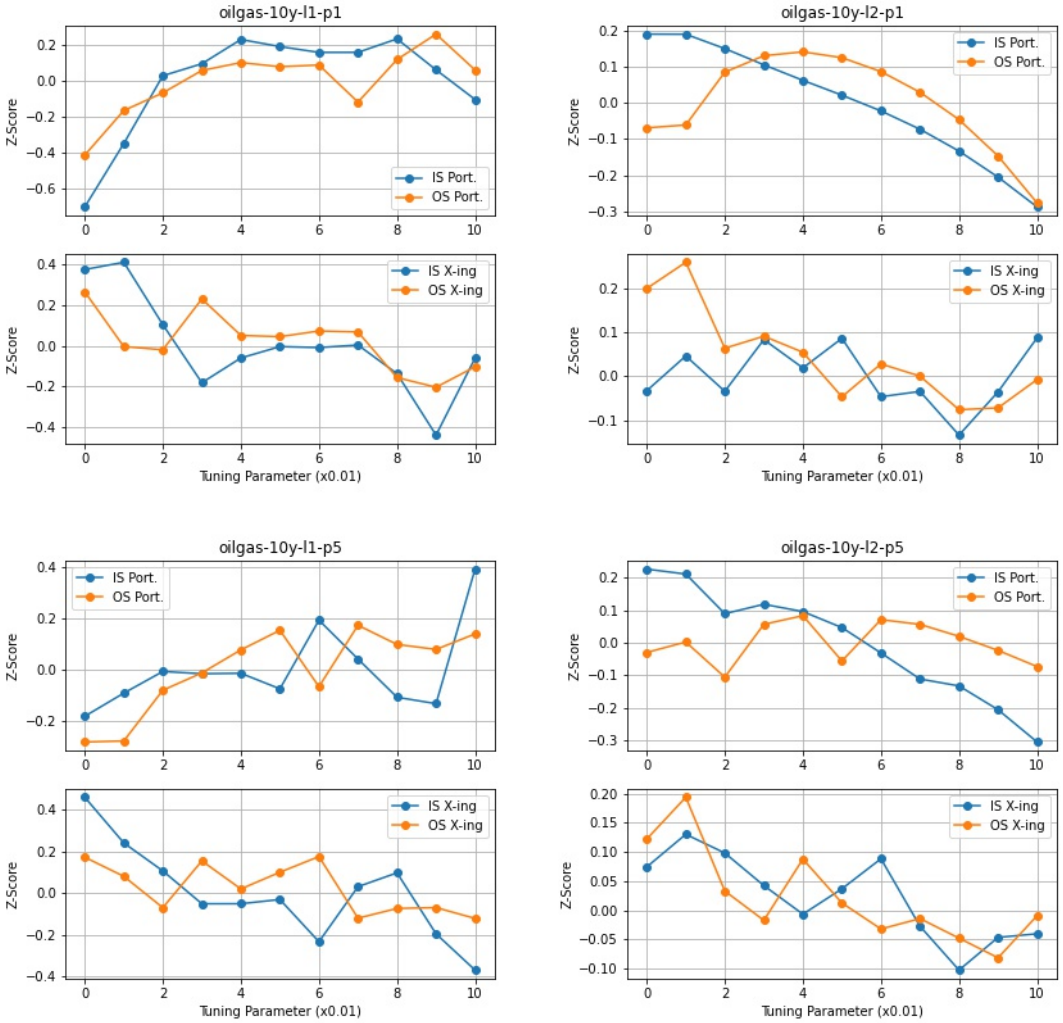

Figure 4.5. In-sample and out-of-sample standardized portmanteau and crossing statistics against L1 and L2 regularization parameters on the oil/gas data sets with model order $p = 1$ and $p = 5$.

We finally present some results specific to the L1-regularized case. As previously mentioned, L1 regularization is interesting in that it promotes sparsity and does not admit a general closed-form solution, requiring approximate solutions via iterative algorithms in our case.

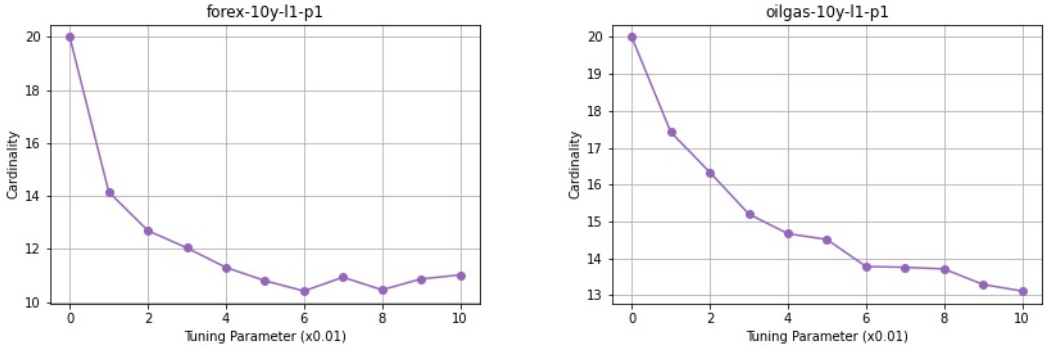

Figure 4.6. Cardinality of $w^\star$ against L1 regularization parameter for model order $p = 1$ on both data sets.

Figure 4.6 plots the average cardinality of the portfolio weights $w^\star$ selected by the simulated annealing algorithm against the L1 regularization parameter with model order $p = 1$ on both data sets, and verifies that L1 regularization promotes sparsity in our setting as well. The specific average frequency at which each asset appears in $w^\star$ is shown in Figure 4.7 for both data sets.

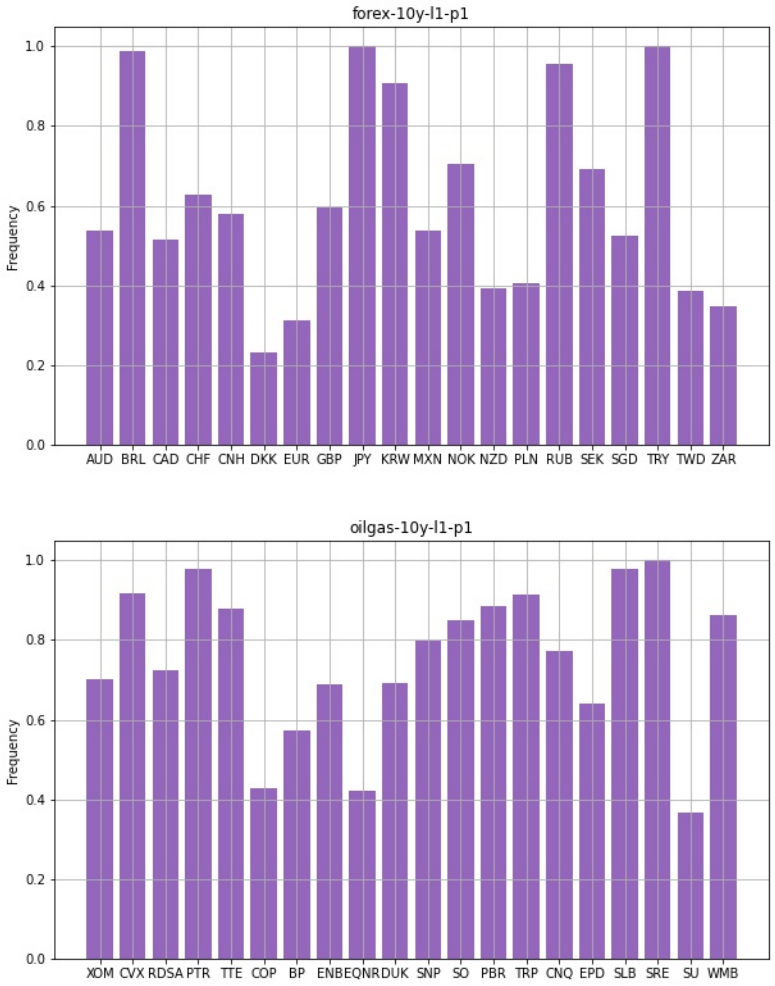

Figure 4.7. Frequency at which each asset appears in $w^\star$ against L1 regularization parameter for model order $p = 1$ on both data sets.

Over the L1 regularization parameter γ for model orders $p = 1$ and $p = 5$, we plot the average log-likelihood for each value of $\gamma = \{0, 0.01, \ldots, 0.1\}$ yielded by the $(w^\star, \phi_0^\star, \phi^\star, \sigma_\varepsilon^\star)$ outputted by different algorithms in Figure 4.8. `JML` (joint maximum likelihood) refers to the black-box methods in `scipy.optimize` for constrained optimization over all parameters, and `SA` (simulated annealing) refers to Algorithm SA described in Section 3.2. Figure 4.8 shows that Algorithm SA outperforms black-box optimization methods in attaining higher log-likelihood values, especially for higher model orders (c.f. plots for $p = 1$ and $p = 5$). The smooth lines for `SA` and jagged lines for `JML` also suggest that the former is quite stable, despite its randomized nature, and that the latter is unstable. For large N, we could reasonably expect `SA` to have even better performance relative to `JML`.

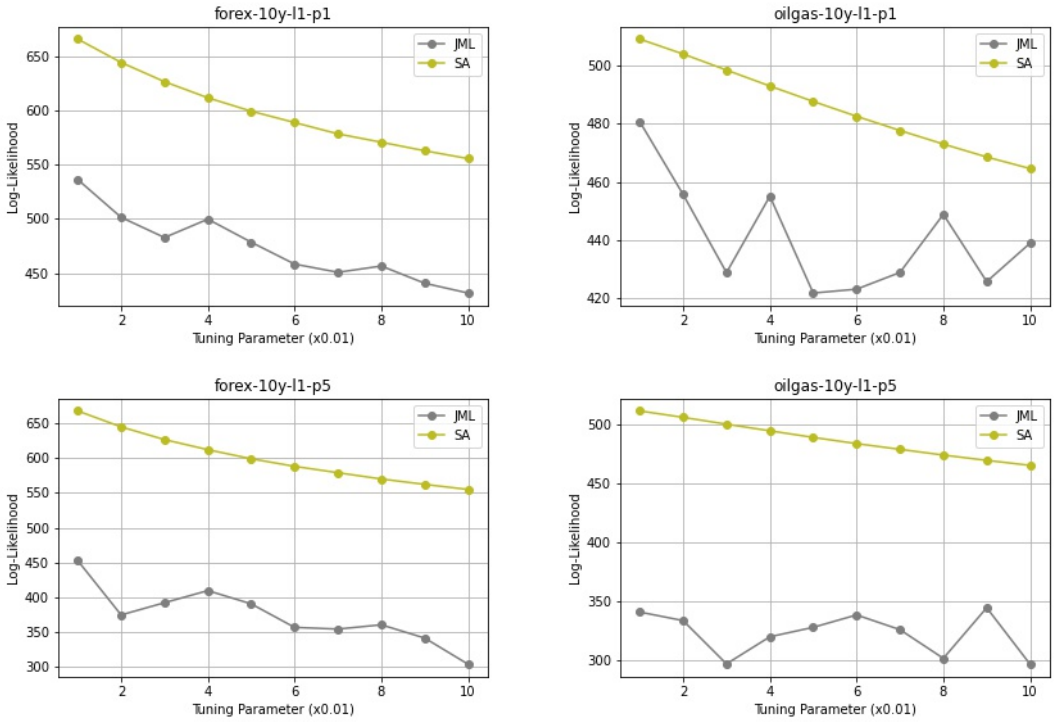

Figure 4.8. Log-likelihood of $(w^\star, \phi_0^\star, \phi^\star, \sigma_\varepsilon^\star)$ outputted by a black-box algorithm and Algorithm SA against L1 regularization parameter for model order $p = 1$ and $p = 5$ on both data sets.

5 Conclusion

We have devised algorithms to build approximate likelihood-optimal portfolios for AR(p) models of arbitrary order in the cases of no regularization, L1 regularization, and L2 regularization on the weights. The algorithms take advantage of the structure of the optimization problems, splitting the full problem into two simpler problems. The inner optimization problems have closed-form solutions or efficient methods to obtain approximate solutions, and the outer optimization problems can be solved with either black-box algorithms or simulated annealing depending on the setting. We then implement the algorithms on real-world asset price data for currencies and the oil/gas sector for varying model order, regularization type, and regularization parameter. The results suggest that moderate model order and regularization improves the model fit based on the AIC and BIC, as well as the mean reversion based on the portmanteau statistic. Directions for further exploration may include testing trading strategies on the portfolios over model order and regularization, developing methods to fit AR(p) models with fat-tailed errors (as is common for financial time series) or MA(q) models, and methods to enforce stationarity for the model coefficients.

Acknowledgements

We are grateful to Professor Zhou Fan[3] for helpful guidance, discussions, and suggestions.

[3]Department of Statistics and Data Science, Yale University

References

Barzilai, J. and Borwein, J.M. "Two-point step size gradient methods," *IMA Journal of Numerical Analysis*, 8(1), 141-148, 1988.

Box, G.E.P. and Tiao, G.C. "A canonical analysis of multiple time series," *Biometrika*, 64(2), 355-365, 1977.

Brodie, J., Daubechies, I., De Mol, C., Giannone, D., and Loris, I. "Sparse and stable Markowitz portfolios," *Proceedings of the National Academy of Sciences*, 106(30), 12267-12272, 2009.

Cuturi, M. and d'Aspremont, A. "Mean reversion with a variance threshold," *Proceedings of the 30th International Conference on Machine Learning*, 2013.

Engle, R.F. and Granger, C.W.J. "Co-integration and error correction: representation, estimation, and testing," *Econometrica*, 55(2), 251-276, 1987.

Kedem, B. and Yakowitz, S. *Time Series Analysis by Higher Order Crossings*, IEEE Press, Piscataway, 1994.

Jurek, J.W. and Yang, H. "Dynamic portfolio selection in arbitrage." *SSRN Electronic Journal*, 2007.

Leung, T. and Li, X. *Optimal Mean Reversion Trading: Mathematical Analysis and Practical Applications*, World Scientific, Singapore, 2016.

Ljung, G.M. and Box, G.E.P. "On a measure of lack of fit in time series models." *Biometrika*, 65(2), 297-303, 1978.

Ornstein, L.S. and Uhlenbeck, G.E. "On the theory of the Brownian motion," *Physical Review*, 36, 823-841, 1930.

Zhang, J., Leung, T., and Aravkin, A. "Sparse mean-reverting portfolios via penalized likelihood optimization," *Automatica*, 111, 108651, 2018.