Title
Sparse Optimization Methods and Statistical Modeling with Applications to Finance

Author
Ho, Michael

Publication Date
2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Sparse Optimization Methods and Statistical Modeling with Applications to Finance

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematics

by

Michael Ho

Dissertation Committee:
Professor Jack Xin, Chair
Associate Professor Zheng Sun
Professor Knut Sølna

2016

DEDICATION

To Tutty,Nadhirah and Alia

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Page

ACKNOWLEDGMENTS

Many people have helped me during the course of my studies at UC Irvine. First I would like to thank my advisor Professor Jack Xin for his guidance in my research. Jack has been a source of inspiration for much of the content of this dissertation. Without his help and guidance this work would not have been possible.

Secondly I would like to thank Professor Zheng Sun for offering her expertise in finance. Her suggestion to study mean-variance portfolio was the start of this research work. I truly appreciate the time and expert advice she has provided throughout the course of this dissertation work.

In addition I wish to thank Professor Knut Solna for serving as a dissertation committee member and for his advice and encouragement.

I also thank the Mathematics department for their financial support during the course of my studies. Special thanks to graduate chair Professor Patrick Guidotti and Donna McConnell for creating a positive and friendly learning environment for the graduate students. In addition to Mathematics department this work was also supported from NSF grant DMS-1211179.

Finally I would like to acknowledge my family. Without their support none of this is possible. I look forward to spending more time with you once this dissertation is complete.

Portions of chapter 4 are reprinted from the article

> M. Ho, Z. Sun, and J. Xin. Weighted elastic net penalized mean-variance portfolio design and computation. *SIAM J. Financial Mathematics*, 6:1220-1244, 2015.
>

with permission from Society for Industrial and Applied Mathematics (SIAM). I would like to thank SIAM for granting their permission to use this material. The co-authors listed in this publication directed and supervised research which forms the basis for the dissertation.

CURRICULUM VITAE

Michael Ho

EDUCATION

Doctor of Philosophy in Mathematics **2016**
University of California, Irvine *Irvine, California*

Master of Science in Mathematics **2011**
California State University, Long Beach *Long Beach, California*

Master of Science in Electrical Engineering **2002**
The Ohio State University *Columbus, Ohio*

Bachelor of Science in Electrical Engineering **2000**
Trine University *Angola, Indiana*

RESEARCH EXPERIENCE

Graduate Research Assistant **2001–2002**
The Ohio State University *Columbus, Ohio*

EMPLOYMENT

Sr. Principal Engineer **2002–2016**
Raytheon *El Segundo, CA*

ABSTRACT OF THE DISSERTATION

Sparse Optimization Methods and Statistical Modeling with Applications to Finance

By

Michael Ho

Doctor of Philosophy in Mathematics

University of California, Irvine, 2016

Professor Jack Xin, Chair

It is well known that the out-of-sample performance of Markowitz's mean-variance portfolio criterion can be negatively affected by estimation errors in the mean and covariance. In this dissertation we examine methods to address this problem through application of methods and techniques from sparse optimization and modeling. Two new techniques are developed with the aim of improving the performance of mean-variance portfolio optimization.

In the first technique a pairwise weighted elastic net penalized mean-variance criterion for portfolio design in proposed. Here we motivate the use of this penalty through a robust optimization interpretation. This interpretation is then employed to develop a bootstrap calibration technique for the pairwise elastic net. The benefit of the pairwise weighted elastic net and calibration is shown in portfolio performance results using recent U.S. stock market data.

In the second application robust Kalman filtering techniques are applied to return covariance estimation from high frequency financial price data. The methods developed address three factors which make covariance estimation from high frequency data difficult: 1) microstructure noise, 2) asynchronous trading, and 3) jumps. The performance of these robust Kalman filtering techniques are tested against simulated high frequency data and are compared with other existing covariance estimators. The results indicate that the robust Kalman filtering

techniques substantially improve covariance estimation performance versus other approaches.

Chapter 1

Introduction

Modern portfolio theory (MPT) addresses the problem of constructing an optimal investment portfolios. The cornerstone of the theory is that investors make decisions based solely on opportunity for profit and risk. An implication of this is that given two investment options with the same expected return an investor will always choose the option with the lowest risk. Hence the investor is solving an optimization problem to minimize risk subject to a given level of return.

In 1952 Harry Markowitz proposed a mean variance criterion for portfolio selection [66]. In the mean-variance criterion the mean return and return variance serve as surrogates for opportunity and risk. Thus the optimization problem becomes minimize variance such that the expected return exceeds a threshold. Mathematically this problem can be formulated as convex quadratic program with linear constraints and can solved using a wide variety of techniques [75, 12].

One drawback of the mean variance criterion for portfolio optimization is that it requires the investor to specify the expected return of each asset and the covariance of the returns of different assets. This poses a problem when the future mean and covariance matrix are not known exactly, since incorrect parameter values can lead to sub-optimal portfolio performance [70, 25]. This additional risk due to parameter uncertainty is commonly referred to as estimation risk.

Two classes of approaches to addressing this limitation are studied in this dissertation:

- Class 1: Estimate the unknown parameters using historical data
- Class 2: Augment the mean-variance objective criterion with constraints or penalties.

1.1 Class 1: Parameter Estimation

Most parameter estimation approaches use historical return data [59] to estimate mean and covariance. One approach motivated by the law of large numbers is to estimate the unknown parameters using sample averaging. This approach can be very accurate when sufficient training data is available and the returns are identically distributed.

Despite its simplicity there are difficulties in effectively implementing the sample averaging approach. The primary difficulty is that in many cases there is limited amount of relevant historical financial return data available to estimate the mean and covariance. One cause of this is that the investments' return can be non-stationary. Thus only a limited amount of past data can be used in estimating the current mean and covariance. Since the volatility of assets returns can be large, the sample average may require a large number of samples to converge. Further complicating the problem is that the covariance matrix for financial data can be ill-conditioned. This makes the portfolio weights extremely sensitive to small parameter errors. These estimation errors result in sub-optimal out of sample portfolio performance [25, 6, 49].

As an alternative to sample average estimates, Bayesian estimators for both mean and covariance have been proposed [39, 50, 55]. These estimators effectively “shrink” the sample average estimates towards a more structured estimate (via a convex combination) which takes into account prior knowledge. Prior knowledge can take the form of structured data models such as a single factor model [80] or the Fama- French three-factor model [31]. Shrinking the sample average estimates towards the more structured model reduces the variability in the parameter estimates and can improve out-of-sample portfolio performance.

Appealing to the law of large numbers covariance estimation errors can be reduced by using more data in the sample average estimate. One approach to obtain more data is to simply increase the time window size when forming the sample covariance (e.g. use 1 year of data vs 3 months of data). In order for this approach to be effective the additional data used in covariance estimation should be nearly identically distributed to future data. If the data statistics are non-stationary then increasing the window’s size to obtain more data may not improve portfolio performance as the additional data used in the covariance estimation may not be relevant to future returns.

Another approach to obtaining more data is to sample at a higher frequency [3] (e.g. 1 second update rate vs 1 day update rate) and maintain the sampling window size. This approach is less vulnerable to non-stationary statistics but presents additional challenges unique to high frequency data. For example, high frequency data is subject to market microstructure noise [14] such as bid-ask bounce which can corrupt volatility and covariance estimates. At higher frequencies the variance of the market microstructure noise can mask the true volatility of the asset returns if it is not accounted for [4, 3]. Asynchronous trading of assets observed at higher frequencies [62] further complicates covariance estimation as the standard sample average estimate assumes return data is available at each time instance.

Many approaches have been proposed for estimating covariance matrices from high frequency data in the presence of asynchronous trading and microstructure noise. For example, the refresh-time approach proposed in [5] addresses asynchronous trading by attempting to synchronize the return data by waiting for all assets to trade at least one time prior to forming an asset price vector used in covariance estimation. One disadvantage of this approach is that much of the data is ignored while waiting for all assets to trade. The pairwise refresh approach [33] uses more data by refreshing the covariance matrix element by element. Here we form a 2×1 asset price vector every time period where two assets trade. This allows for more data to be used but the resulting sample covariance matrix is not guaranteed to be positive semi-definite without applying additional corrections such as a projection method [33]. Another approach is the previous tick method employed in [90] where a fixed sampling grid is defined and trade prices are approximated on that grid as the nearest previous trade price.

To address both micro-structure noise and asynchronous returns, quasi-maximum likelihood estimators were proposed in [2, 60] that utilize pairwise refresh. A two scale realized covariance (TSCV) approach was developed in [90] where covariance estimates are obtained using both low frequency and high frequency sampling. An approach based on Kalman filtering and the EM algorithm [22], models the true unobserved log-price process and observed prices as a discrete linear normal dynamical system. Here the unobserved synchronous true price is treated as latent data and the EM algorithm is used to determine a maximum-likelihood estimate of the covariance. A Bayesian version of the Kalman-EM approach where the posterior distribution of the covariance is approximated via an augmented Gibbs sampler is proposed in [76]. This technique generates an estimate of the posterior distribution of the covariance which can then be used to obtain to a point estimate.

Each of the above techniques addressing micro-structure noise and asynchronous returns utilize a log-normal price model. However, empirical return data often exhibits heavy tails that are better explained by a jump diffusion or stochastic volatility models. Under these conditions the approaches which assume log-normal returns will yield sub-optimal results. Techniques for addressing jumps have been proposed in the literature. In [34] the authors propose wavelet techniques for detecting jumps with an application to volatility estimation. The jumps estimated using this approach are then removed from the observed data prior to volatility estimation. In [10] a jump detector is employed to selectively remove data that contain jumps from the covariance estimation samples prior to TSCV. Another technique proposed in [9] is also robust to jumps but does not address market microstructure noise.

In Chapter 3 we extend the Kalman-EM approach in [22] to discretized jump diffusion models by introducing two Kalman-ECM (KECM) approaches. In our first KECM approach we model the jumps as Laplace distributed random variables. Although the Laplace prior may seem to be an unnatural model for a jump process, we will see that the prior promotes a sparse posterior mode for the jumps by inducing an ℓ_1 norm penalty on the jumps into the complete log-likelihood function. Conditioned on other variables determining the posterior mode for the jumps is a convex ℓ_1 norm penalized quadratic program which can be solved with a variety of fast techniques [43, 11, 7]. In our second KECM approach we consider a more natural, but less tractable, spike and slab model for the jump process.

We also extend the Bayesian approach in [76] to jump models where jumps are modeled using a spike and slab prior [72, 73]. Here we use Gibbs sampling to approximate the posterior of the jumps along with the unknown covariance matrix. An estimate of the posterior mean of the covariance matrix is then obtained using the samples obtained from the posterior distribution.

1.2 Class 2: Reformulate Criterion

For Class 2, several enhancements to the mean-variance criterion have been proposed. One approach that has been shown to improve portfolio return involves regularizing the portfolio selection criterion by adding penalties to the objective function [24, 13, 36, 88] such as norm penalties on the portfolio weights. In [24] ℓ_1 and squared ℓ_2 norm constraints are proposed for the minimum variance criterion along with a method for choosing the constraints. The authors in [88, 87] propose an elastic net penalty and calibration method for use in constrained minimum variance portfolio optimization. In [36] a weighted LASSO [93] and a SCAD penalty [32] are proposed for use in minimum variance portfolios. For the weighted LASSO approach the authors propose a calibration scheme where the portfolio weights are selected according to the variability in the volatility of each asset.

The above norm constrained and penalty approaches for portfolio optimization primarily focus on the minimum variance criterion. The mean return is not taken into consideration when calibrating the penalty, making it less applicable for a mean-variance criterion. In this dissertation we propose a method which can be applied to the more general mean-variance criterion where both mean and variance are considered. In this setting we propose regularizing the objective function with a weighted elastic net (WEN) penalty. A WEN penalty is the sum of a portfolio's weighted ℓ_1 norm and the square of a portfolio's weighted ℓ_2 norm. It is shown that the WEN penalized criterion is equivalent to a robust portfolio optimization problem [42, 83]. Using this interpretation, a bootstrap method for calibrating the weights in the WEN penalty is derived.

The WEN penalized criterion has two advantages over the general robust portfolio optimization problem. First the WEN penalized portfolios can be solved using fast algorithms for ℓ_1 penalized optimization problems such as the Split-Bregman algorithm [43], the alternating direction method of multipliers (ADMM) algorithm [11] and the fast iterative shrinkage thresholding algorithm (FISTA) [7]. This is in stark contrast to the more general robust portfolio optimization problem which requires using semi-definite programming techniques [44] which do not scale well to large portfolios. Secondly the WEN penalty results in sparse portfolios which can contribute to reduced portfolio turnover and transaction costs when compared with portfolios generated with generalized robust optimization approaches. This dissertation also introduces a novel adaptive support split-Bregman approach to computing WEN penalized portfolios. This new algorithm exploits the sparse nature of WEN penalized solutions to minimize computational requirements for large portfolios. We show that this new algorithm offers significant improvements in convergence speed versus other techniques.

1.3 Outline and Contributions

The remainder of this dissertation is structured as follows. In Chapter 2, we provide an overview of various mean-variance driven portfolio optimization criteria along with commonly used parameter estimation techniques. In Chapter 3 we present new jump robust covariance estimation techniques for asynchronous, and noisy high frequency data sources. Chapter 4 introduces the weighted elastic net penalty for portfolio optimization and provides a justification for its use. Conclusions and future work are discussed in Chapter 5.

The key contributions of this dissertation are as follows:

1. Introduction and analysis of jump robust covariance matrix estimation techniques for asynchronous high frequency data with microstructure noise (Chapter 3)

2. Application of sparsity promoting priors (weighted Laplace prior, spike and slab) to improve robustness of covariance matrix estimation in jump diffusion model (Chapter 3)

3. Application of weighted elastic net (WEN) and pairwise weighted elastic net (PWEN) penalties to mean-variance portfolio selection (Chapter 4)

4. Relationship between pairwise weighted elastic net penalty and robust optimization is shown (Chapter 4)

5. Bootstrap calibration technique for weighted elastic net penalty is proposed using robust optimization interpretation (Chapter 4)

6. Development of adaptive support split-Bregman approach for solving weighted elastic net penalized portfolio problems (Chapter 4).

Chapter 2

Mean-Variance Portfolio Optimization and Parameter Estimation

This chapter reviews some background material relevant to Chapters 3 and 4. We begin by presenting several mean-variance inspired portfolio criteria. We then describe approaches to mean and covariance estimation needed to implement the various portfolio selection criteria.

2.1 Portfolio Selection Criteria

Suppose that there exists a set of N assets. Let $\{r_n\}_{n=1}^N$ be the returns of each asset for a time period T. We model $\{r_n\}_{n=1}^N$ as random variables with finite mean and covariance. For the purposes of this section we will only model the first and second order statistics of r_n.

A portfolio is defined to be a set of weights $\{w_n\}_{n=1}^N \subset \mathbb{R}$. When $w_i > 0$ a long position in the i^{th} asset is taken. A short position on the i^{th} asset is taken if $w_i < 0$. There are many criteria for how an investor chooses the portfolio. In this section we review some common criteria.

2.1.1 Mean-Variance Criteria

The mean-variance criteria was proposed by Markowitz in 1952 [66]. The objective of the mean-variance criteria is to minimize the variance of a portfolio's return such that the mean return equals a specific threshold ρ. A portfolio optimal in the mean-variance sense is said to be on the efficient frontier [18].

The motivation for the mean-variance criteria is that low risk portfolios are preferred over higher risk portfolios provided each portfolio provides the same mean return. The mean-variance criteria uses variance as a proxy for risk (high variance = high risk, low variance = low risk).

Given the returns of each asset, the portfolio return is $w^T r$. If the mean and covariance of the asset returns are known then the mean and variance of the return can be easily computed to $w^T \mathbb{E} r$ and $w^T \Gamma w$ respectively, where Γ is the covariance of r. Thus the mean variance criteria may be written as the following quadratic program

$$\begin{aligned} \min_w \;& w^T \Gamma w \\ \text{s.t. } & w^T \mu = \rho \\ & w^T \vec{1} = c. \end{aligned}$$

If short selling is not allowed then the mean variance criteria becomes

$$\begin{aligned} \min_w \;& w^T \Gamma w \\ \text{s.t. } & w^T \mu = \rho \\ & w^T \vec{1} = c \\ & w_n \geq 0 \quad \forall n. \end{aligned}$$

These quadratic programs can be efficiently solved using many different algorithms [85].

2.1.2 Sharpe Ratio Criteria

An asset is called risk free if its return is deterministic and non-negative. The excess return of each asset is defined as $r - r^f$ where r^f is the return of a risk-free asset over a time period T.

The Sharpe ratio, SR is defined as ratio between the mean excess return and the standard deviation of the excess return, i.e.

$$SR = \frac{w^T \mu}{\sqrt{w^T \Gamma w}}.$$

where μ is the mean of the excess return. For investors maximizing Sharpe ratio is desirable in that one will get the highest excess return per unit of risk (standard deviation). A portfolio that has maximal Sharpe ratio is often referred to as the tangency or market portfolio. It is of theoretical interest due to the one-fund theorem which states that any efficient portfolio is a linear combination of a risk free portfolio and the tangency portfolio [52]. Figure 2.1 shows a plot of the efficient frontier with and without a risk-free asset.

In Sharpe ratio optimization we seek to maximize SR. Thus the Sharpe ratio optimization criteria may be written as

$$\begin{aligned} &\max_w \frac{w^T \mu}{\sqrt{w^T \Gamma w}} \\ &\text{s.t. } w \neq 0 \end{aligned}$$

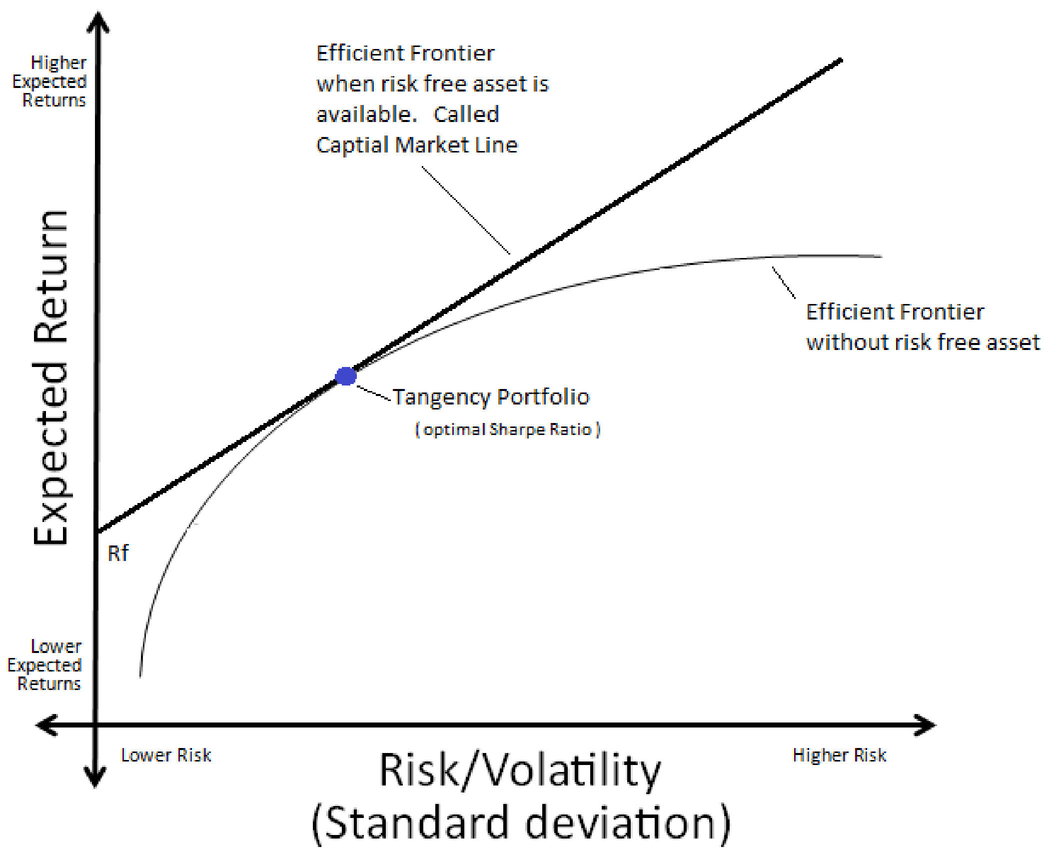

Figure 2.1: Portfolios satisfying the mean variance criteria are said to be on the efficient frontier. The portfolio with maximal Sharpe ratio is referred to as the Tangency Portfolio.

in the case short sales are allowed or

$$
\begin{aligned}
&\max_w \frac{w^T\mu}{\sqrt{w^T\Gamma w}} \\
&\text{s.t. } \; w \neq 0 \\
&\qquad w_n \geq 0 \quad \forall n
\end{aligned}
$$

if short selling is not allowed.

A solution to the Sharpe ratio optimization problem can be obtained in closed form when short selling is allowed. First note that the Sharpe ratio is unaffected by a positive scalar multiplication of the weight vector. Thus the problem can be reformulated as (up to a positive scalar multiple)

$$
\begin{aligned}
&\max_w \; w^T\mu \\
&\text{s.t. } \; w^T\Gamma w \leq c
\end{aligned}
$$

where $c > 0$ is arbitrary. If $\mu \neq 0$ the optimal objective value is positive and the constraint above will be satisfied with equality. Thus we may solve the equivalent problem

$$\min_w \; -w^T\mu + \rho\left(w^T\Gamma w\right) \tag{2.1}$$

where $\rho > 0$ is arbitrary. From first order optimality conditions the solution is (up to a positive constant multiple)

$$w = \rho^{-1}\Gamma^{-1}\mu \tag{2.2}$$

2.1.3 Minimum Variance Criterion

Estimation of parameters is necessary to implement the mean-variance criteria. It has been recognized that estimation of mean return is more difficult than covariance [69] and thus a minimum variance criterion is often advocated for in recent literature [47, 24, 36]. In the minimum variance criterion the mean of asset returns are ignored and the following criterion is used for portfolio selection

$$\begin{aligned} \min_{w} \ & w^T \Gamma w \\ \text{s.t.} \ & \sum_{i=1}^{N} w_i = 1. \end{aligned} \tag{2.3}$$

Despite ignoring all information on the mean return, the minimum variance criterion often outperforms the mean-variance criterion when judged by out-of-sample Sharpe ratio [24, 47].

2.1.4 Norm Penalized Portfolio optimization

As was stated in the introduction mean-variance portfolio optimization is sensitive to parameter estimation error. To address these concerns a number of norm penalized criteria have been proposed, primarily in the context of minimum variance optimization. Commonly used convex norm penalties include the ℓ_1 norm , squared ℓ_2 norm and elastic net penalties [88]. The ℓ_1 and squared ℓ_2 norm penalties are given as

$$\sum_{i=1}^{N} |w_i| \tag{2.4}$$

and

$$\sum_{i=1}^{N} w_i^2. \tag{2.5}$$

The elastic net penalty is a weighted sum of the ℓ_1 and squared ℓ_2 norm penalties

$$\lambda_1 \sum_{i=1}^{N} |w_i| + \lambda_2 \sum_{i=1}^{N} w_i^2 \tag{2.6}$$

where $\lambda_1, \lambda_2 > 0$. Another convex penalty is the adaptive LASSO penalty [93] which was applied to minimum variance optimization in [36]. The adaptive LASSO penalty is a weighted ℓ_1 norm given by

$$||w||_{\vec{\beta},\ell_1} = \sum_{k=1}^{N} \beta_k |w_k| \tag{2.7}$$

where $\beta_k \geq 0$. Calibration of the weighting parameters for the above penalties has primarily been studied with the goal of improving the portfolio return variance [88, 36].

Several justifications for using ℓ_1 and squared ℓ_2 norms as penalties and constraints have been given in the literature. For example in [13] it is stated that the use of an uniformly weighted ℓ_1 penalty can be motivated by the desire to obtain sparse portfolios and to regularize the mean-variance problem when the covariance is ill-conditioned. In [35] the authors show that estimation risk in the mean-variance setting due to errors in the mean return estimation is bounded above by

$$||\mu - \hat{\mu}||_\infty ||w||_{\ell_1} \tag{2.8}$$

and use that upper bound as a rationale for promoting small $||w||_{\ell_1}$. In [57] it is mentioned that a benefit of using a uniformly weighted ℓ_2 norm penalty is to stabilize the inverse covariance matrix which is often ill-conditioned in financial applications.

Non-convex penalized minimum-variance portfolio criteria were studied in [36]. One such penalty examined in [36] is the Softly Clipped Absolute Deviation (SCAD) penalty [32]. The SCAD penalty is defined as follows

$$\sum_{i=1}^{N} p_\lambda(w_i) \tag{2.9}$$

where

$$p_\lambda(x) = \begin{cases} \lambda|x| & \text{if } |x| \le \lambda \\ -\frac{x^2 - 2a_{SCAD}\lambda|x| + \lambda^2}{2(a_{SCAD}-1)} & \text{if } \lambda < |x| \le a_{SCAD}\lambda \\ \frac{(a_{SCAD}+1)\lambda^2}{2} & \text{if } |x| > a_{SCAD}\lambda \end{cases} \tag{2.10}$$

and where $a_{SCAD} > 2$. This penalty is similar to the ℓ_1 penalty and was initially proposed in context of variable selection. Calibration of the parameters a_{SCAD} and λ in (2.10) for portfolio optimization has not been fully addressed in the literature.

In chapter 4 we introduce a new weighted elastic net penalty for portfolio optimization, and compare its performance with the techniques described in this chapter.

2.2 Parameter Estimation from synchronous noise-free data

Each of the above portfolio optimization criteria require an estimate of the mean and covariance of the return. Estimation of the mean and covariance matrix is often performed using a training set of historical price data from each asset. In this section we shall assume that we observe each asset's efficient price at even spaced times $0, 1, 2, \dots T$. Denote $r_i(t)$ as the return of the i^{th} asset for time period from $(t-1)\delta$ to $t\delta$. We shall assume in this section that the training data is independent and identically distributed.

2.2.1 Sample Average and MLE

The simplest technique for obtaining estimates of mean and covariance is the sample averaging method. Here the estimate of the mean of the i^{th} asset is given as

$$\hat{\mu}_i = \frac{1}{T}\sum_{t=1}^{T} r_i(t)$$

and the estimate of the (i,j) entry of the covariance matrix is

$$\hat{\Gamma}_{i,j} = \frac{1}{T}\sum_{t=1}^{T}(r_i(t) - \hat{\mu}_i)(r_j(t) - \hat{\mu}_j).$$

Due to the strong law of large numbers these estimates converge almost surely to μ_i and $\Gamma_{i,j}$ as the number of samples T goes to infinity. It is noteworthy that for finite T the covariance estimate is biased with expected value

$$\frac{T-1}{T}\Gamma.$$

However for larger T this bias is small.

Another technique for mean and covariance estimation is the maximum likelihood estimator (MLE). Here the estimation criterion is to select μ and Γ that maximize the log-likelihood of the training data. Under some mild regularity conditions the MLE for a parameter θ is an asymptotically efficient estimator. For normally distributed returns the MLE for mean and covariance are identical to the sample average estimators.

2.2.2 Shrinkage and Bayesian Approaches

In financial applications the return statistics may be time-varying which limits the amount of revelent data available for covariance estimation. In this setting the sample average and

MLE approaches may not have sufficient number of samples to converge which can lead to unstable and unreliable portfolio performance [25].

For data starved settings shrinkage and Bayesian approaches to parameter estimation have been proposed. In both of these techniques additional structure is introduced into the parameter estimates which act to stabilize the estimates. The amount of additional structure to incorporate into the estimate is a trade-off between bias and variability in the estimate. This trade can be optimized to minimize estimation error or some other loss criterion.

For shrinkage estimators the additional structure is enforced by constraining the estimates to be a weighted sum of the sample average estimate and a pre-selected target estimate. For mean and covariance estimation we have

$$\hat{\mu} = \rho_1 \hat{\mu}_S + \rho_2 \hat{\mu}_{tgt}$$

$$\hat{\Gamma} = \psi_1 \hat{\Gamma}_S + \psi_2 \hat{\Gamma}_{tgt}$$

where $\psi_i, \rho_i > 0$. Here $\hat{\mu}_S$ and $\hat{\Gamma}_S$ are the sample average estimates and $\hat{\mu}_{tgt}$ and $\hat{\Gamma}_{tgt}$ are the target parameters. The target parameters are typically chosen to add structure into the parameter estimates and may be based on some prior knowledge on the return statistics.

Selection of the weighting factors and target return for mean estimation has been considered in [50]. There a James-Stein estimator is used for the mean where the shrinkage target is a constant vector and summation of the target and sample average is convex

$$\hat{\mu} = \rho \hat{\mu}_S + (1 - \rho)\eta \vec{1}.$$

In [50] the weighting factor is defined as

$$\rho = \min\left\{1, \frac{(N-2)}{T_{train}(\mu_S - \eta\vec{1})^T \hat{\Gamma}^{-1}(\mu_S - \eta\vec{1})}\right\}.$$

Although biased, James-Stein estimators have the surprising property of achieving lower (weighted) mean-squared error than the MLE estimator when the returns are Gaussian [28] and the number of assets is 3 or more.

For financial applications the target parameter η may be set to the global mean of the returns (i.e. averaging over all assets). This has the effect of shrinkage the sample average estimate to the global average. Thus extremely high and low sample average returns are biased towards the global average.

In [54] a shrinkage technique for estimating covariance is presented where the shrinkage target is scaled identity matrix

$$\hat{\Gamma} = \psi_1 \hat{\Gamma}_S + \psi_2 I.$$

For $\psi_2 > 0$ we see that this estimator is guaranteed to be positive definite with minimum eigenvalue $\geq \psi_2$. A criterion for choosing ψ_1 and ψ_2 is proposed in [54]. Under mild assumptions this criterion results in an asymptotically (as T goes to infinity and N/T tends to a positive constant) optimal estimate with respect to the squared Frobenius norm of the covariance error.

More sophisticated shrinkage targets for covariance can also be employed. For example in [53] the authors propose using structured covariance estimates as shrinkage targets. In particular a single factor return model is used to determine the shrinkage target. Thus the resulting covariance estimate is a weighted sum of the sample covariance and a structured single factor covariance.

Bayesian approaches can also be used to estimate mean and covariance. The Bayesian approach is similar to shrinkage techniques in that prior knowledge and assumptions are used to add stability to the parameter estimation process. In the Bayesian approach the mean and covariance are treated as a random vector and matrix with a known prior distribution. Given observed return data, estimates of the mean and covariance are chosen to minimize a loss function. One typical loss function for a parameter θ is the mean-squared error (MSE) loss function.

$$\mathbb{E}_{\theta|Y}(\theta - \hat{\theta})^2$$

where the expectation is with respect to the posterior distribution of the parameter θ given Y. Under this loss function the estimate of the parameter is equal to the conditional mean of θ given past returns, Y [56]. Another loss function is the binary loss function

$$L_{bin}(\theta, \hat{\theta}) = \begin{cases} 0 \text{ if } |\theta - \hat{\theta}| \leq \epsilon \\ 1 \text{ otherwise} \end{cases}$$

A common and convenient choice of prior distribution is one that is conjugate to the likelihood function [37, 40]. If one assumes normally distributed returns then the conjugate prior distribution for the mean and covariance is the normal-inverse-Wishart distribution. This distribution is characterized by 4 parameters $(\mu', \lambda, \Psi, \nu)$ and can be written as

$$p(\mu, \Gamma) = \mathcal{N}(\mu, \mu', \frac{1}{\lambda}\Gamma)\mathcal{W}^{-1}(\Gamma, \Psi, \nu)$$

where

- $\mathcal{N}(x, x', \Sigma)$ is the normal distribution in x with mean x' and covariance Σ
- $\mathcal{W}^{-1}(\Sigma, \Psi, \nu)$ is the inverse Wishart distribution in Σ with scale Ψ and ν degrees of freedom.

Since this distribution is conjugate to the normal likelihood function the posterior distribution is also normal-inverse-Wishart. The parameters for the posterior distribution are [52]

$$\left(\frac{T\hat{\mu}_S + \lambda\mu'}{T + \lambda}, \lambda + T, \Psi + T\hat{\Gamma}_S + \frac{T\lambda}{T + \lambda}(\hat{\mu}_S - \mu')(\hat{\mu}_S - \mu')^T, T + \nu\right).$$

If one takes as a loss function the MSE then the resulting mean and covariances are

$$\hat{\mu} = \frac{T\hat{\mu}_S + \lambda\mu'}{T + \lambda}$$

and

$$\hat{\Gamma} = \frac{1}{T + \nu - N - 1}\left(\Psi + T\hat{\Gamma}_S + \frac{T\lambda}{T + \lambda}(\hat{\mu}_S - \mu')(\hat{\mu}_S - \mu')^T\right).$$

Thus we see that the Bayesian estimate for this prior is a type of shrinkage estimator.

The choice of the hyperparameters in the prior $(\mu', \lambda, \Psi, \nu)$ can be addressed using an empirical Bayesian approaches or techniques developed for shrinkage estimators. For the empirical Bayesian approach one may assume a structured form of μ' and Ψ such as constant mean, constant variance and covariance. This reduces the number of "free" hyperparameters significantly. Then one can estimate these few number of hyperparameters from the return training data using method of moment estimators [52].

Chapter 3

Covariance Estimation from High Frequency Data

In this chapter we present jump robust methods of covariance estimation from high frequency data. The techniques address the following challenges with using high frequency data for covariance estimation

1. Asynchronous trading
2. Market microstructure noise
3. Jumps.

In section 3.1 we introduce the models which form the basis for our covariance estimation approaches. In section 3.2 and 3.3 we describe numerical algorithms for computing the covariance estimate with both the Laplace and spike and slab prior. A performance evaluation of our proposed approach is presented in section 3.4 using simulated high frequency data. The use of matrix factorization approaches for covariance estimation from high frequency data is presented in section 3.5.

3.1 High Frequency Return Modeling

Suppose that we have N assets where the true (or efficient) log price of the n^{th} asset at time t is $X_n(t)$. Let $X(t)$ denote the $N \times 1$ vector of log prices for each asset at time t and let T denote the total number of time samples. Here $X_n(t)$ can be viewed as the fundamental value of the asset in an efficient market without friction [78].

We model the dynamics of the log prices using a discrete time jump diffusion model with a drift D

$$X_i(t) = X_i(t-1) + V_i(t) + \tilde{J}_i(t) Z_i(t) + D. \tag{3.1}$$

Here we assume the following:

- $V(t)$ is multivariate normally distributed with mean 0 and covariance Γ
- $\tilde{J}_i(t)$ is normally distributed with zero mean and variance $\sigma^2_{j,i}(t)$
- $Z_i(t)$ is Bernoulli distributed, with $Pr(Z_i(t) = 0) = \zeta$
- $\tilde{J}_m(t) \perp\!\!\!\perp \tilde{J}_n(s), Z_m(t) \perp\!\!\!\perp Z_n(s) \;\; , m \neq n$ and all t, s
- $\tilde{J}, Z, V$ are jointly independent.

To simplify notation we denote the jump component as

$$J(t) = \tilde{J}(t) Z(t). \tag{3.2}$$

In many markets trading of distinct assets does not occur simultaneously. When trades occur asynchronously, current pricing data for all assets will not be observed. For prices that are observed, market microstructure noise needs to be addressed. Here transaction costs due to order processing expenses, inventory costs and adverse selection costs [14] add noise to the true efficient price. Thus the true efficient price is not directly observed.

Both asynchronous returns and microstructure noise can be captured in the following observation model

$$Y(t) = \tilde{I}(t)X(t) + W(t) \tag{3.3}$$

where

- $\tilde{I}(t)$ is a"partial" identity matrix where the rows corresponding to missing asset prices at time t are removed
- $W(t)$ is normal distributed market microstructure noise with zero mean and covariance $\Sigma_o(t) = \tilde{I}(t)\Sigma_o'\tilde{I}(t)^T$.

Here Σ_o' is a diagonal matrix $\text{diag}(\sigma_{o,1}^2, \ldots, \sigma_{o,N}^2)$. For purposes of this paper we shall assume that $\tilde{I}(t)$ is known, and that $\{W(t), X(t)\}_{t=1}^T$ are jointly independent. In section 3.4 we will test our algorithms on simulated data where the microstructure noise and price innovation are statistically dependent.

3.1.1 Conditional Distributions of Observations and Log-Prices

Now we examine the joint probability distribution of $X(1:T)$,$Y(1:T)$,$J(2:T)$. Here the notation $X(m:n)$ refers to the set $\{X(m), X(m+1), \ldots, X(n)\}$. We consider the case of when the parameters $D, \Gamma, \sigma_{o,i}^2, \zeta$ and $\sigma_{j,i}^2$ are random variables with known prior distributions. Details on our assumed priors are given in section 3.1.2.

To determine the probability distribution we first note that our model in equations (3.1) and (3.3) can be represented by the Bayesian network depicted in Figure 3.1. From the Bayesian network we see that the following conditional independence properties hold

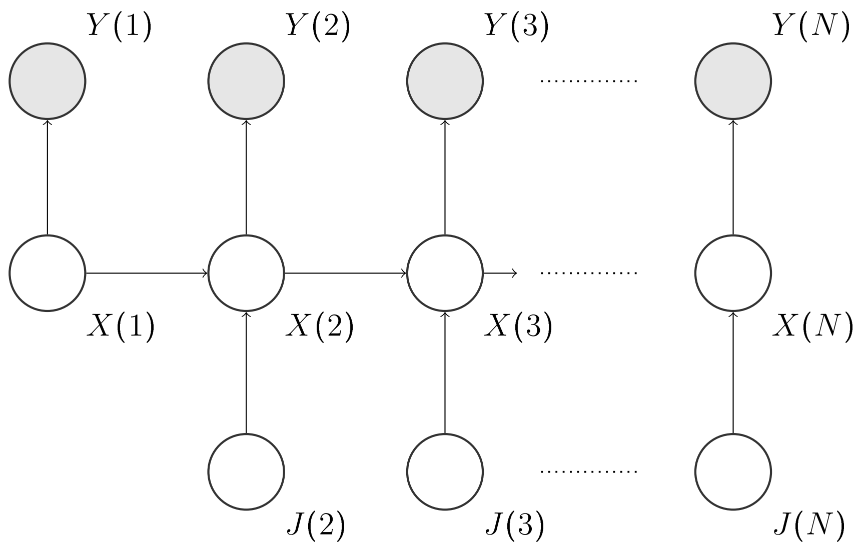

Figure 3.1: Bayesian Network Representation of (X, Y, J). Observed variables are shaded. Here the model parameters are not shown.

$$
\begin{aligned}
Y(t) \perp\!\!\!\perp J(s) | X(t) \ \ \forall s \\
Y(t) \perp\!\!\!\perp X(s) | X(t) \ \ \forall s \neq t \\
X(t) \perp\!\!\!\perp X(s) | (X(t-1), J(t)) \ \ \forall s < t-1 \\
X(t) \perp\!\!\!\perp J(s) | (X(t-1), J(t)) \ \ \forall s \neq t.
\end{aligned}
$$

From the conditional independence implied by the Bayesian network we have that the probability distribution conditioned on the parameter values may be fully characterized as follows

$$
\begin{aligned}
p(y(t)|x(1:T), \Sigma_o^2(t)) &\sim \mathcal{N}(\tilde{I}(t)x(t), \Sigma_o(t)) \\
p(x(t+1)|x(1:t), j(2:t+1), d, \Gamma) &\sim \mathcal{N}(x(t) + j(t+1) + d, \Gamma) \\
p(x(1)) &\sim \mathcal{N}(\mu, K) \\
p(j(t)|\zeta, \sigma_j^2(t)) &\sim \prod_{i=1}^{N} f(j_i(t)).
\end{aligned}
$$

Here f is the spike and slab prior

$$f(j_i(t)) = \zeta\delta_0(j_i(t)) + \frac{1-\zeta}{\sqrt{2\pi}\sigma_{j,i}(t)} \exp\left(-\frac{j_i(t)^2}{2\sigma_{j,i}^2(t)}\right) \tag{3.4}$$

with δ_0 being a point mass distribution at 0. The initial time parameters, μ and K can be chosen based on prior stock return data and will be treated as known values.

3.1.2 Prior Distribution of Parameters

To allow for more flexible modeling we shall impose prior distributions on the parameters $D, \Gamma, \sigma_{o,i}^2$ as well as the jump parameters ζ and $\sigma_{j,i}^2$. Here we take a commonly used approach of using conjugate prior distributions which facilitate calculation of conditional maximum a posteriori (MAP) parameter estimates. These priors will play an essential part in the proofs of convergence for the ECM algorithm presented in Section 3.2.

The drift parameter D is modeled as normally distributed with mean $\bar{D}$ and covariance $\sigma_D^2 I$

$$D \sim \mathcal{N}(\bar{D}, \sigma_D^2 I),$$

which is conjugate to the multivariate normal distribution given above. For the covariance matrix prior we use an inverse Wishart prior (which is also conjugate to the multivariate normal) with $\eta > N-1$ degrees of freedom and positive definite scale matrix W_o

$$\Gamma \sim \mathcal{W}^{-1}(W_o, \eta).$$

In the observation noise variance,$\sigma_{o,i}^2$, we impose a inverse gamma distribution with shape parameter $\alpha_o > 0$ and scale $\beta_o > 0$

$$\sigma_{o,i}^2 \sim IG(\alpha_o, \beta_o).$$

Finally for the jump parameters ζ and σ_j^2 we use the beta distribution and inverse gamma distribution as priors

$$\zeta \sim \text{Beta}(\alpha_\zeta, \beta_\zeta)$$

$$\sigma_{j,i}^2(t) \sim IG(\alpha_j, \beta_j).$$

We assume that ζ and $\sigma_{j,i}^2(t)$ are independent and that the parameters in each of the prior distributions is known. For each of these priors the hyperparameters may be selected to make them relatively uninformative.

3.1.3 Mixture Model Representation

We may also represent our jump model as mixture model with 2^{TN-N} components. To see this we condition on $Z(1:T)$ and obtain the following

$$\begin{aligned} p(y(t)|x(1:T), \Sigma_o(t)) &\sim \mathcal{N}(\tilde{I}(t)x(t), \Sigma_o(t)) \\ p(x(t)|x(1:t-1), z(2:t), d, \Gamma) &\sim \mathcal{N}(x(t-1) + d, \Gamma + \text{Diag}(t, z(t))) \\ p(x(1)) &\sim \mathcal{N}(\mu, K) \\ p(z(t)|\zeta) &\sim \zeta^{TN-T_J}(1-\zeta)^{T_J} \end{aligned}$$

where T_J is the total number of jumps

$$T_J = \sum_{i,t} z_i(t)$$

and where $\mathrm{Diag}(t, z(t))$ is the diagonal matrix

$$\mathrm{Diag}(t, z(t)) = \begin{pmatrix} z_1(t)\sigma_{j,1}^2(t) & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & z_N(t)\sigma_{j,N}^2(t) \end{pmatrix}.$$

Here we see that the covariance is time-varying and at time t is equal to $\Gamma + \mathrm{Diag}(t, Z(t))$. Thus our model is equivalent to a large switching state space model [41] with a log-price posterior distribution consisting of 2^{TN-N} components.

3.1.4 Laplace Prior Approximation

Recall from the previous section that jump model is equivalent to a switching state space model. Inference in switching state space models becomes intractable as the number of states increase [41]. For example estimation of the posterior distribution of X given Y involves marginalizing out the 2^{TN-N} possible states for Z, which is an intractable integral. Maximum a posteriori (MAP) estimation of Z is also difficult due to the multimodal structure of $p(z)$.

In this section we approximate the distribution of J using a Laplace distribution. We denote the Laplace distribution for J as $g(j)$

$$p(j) \approx g(j|\lambda) \doteq \prod_{i,t} \frac{\lambda_i(t)}{2} \exp\left(-\lambda_i(t)|j_i(t)|\right)$$

where $\lambda_i(t) > 0$.

There are two advantages to taking this approximation. First the log-likelihood of a Laplace distribution is concave in its parameter. This aids in conditional MAP estimation of J. Secondly, the Laplace distribution is desired in that it promotes sparse MAP estimates of J

[79, 67, 1] making it a good approximation to infrequent jumps. We illustrate this with the following example.

Example 3.1. *Suppose that κ is a Laplace distributed random variable with parameter 2 and $\eta = \kappa + q$ where q is independent of κ and normally distributed with mean 0 and variance 1. Suppose η is observed to be 0.5. Then the likelihood of κ given η is $\mathcal{N}(0.5, 1)$ but the posterior of κ has its mode at 0 as shown in Figure 3.2.*

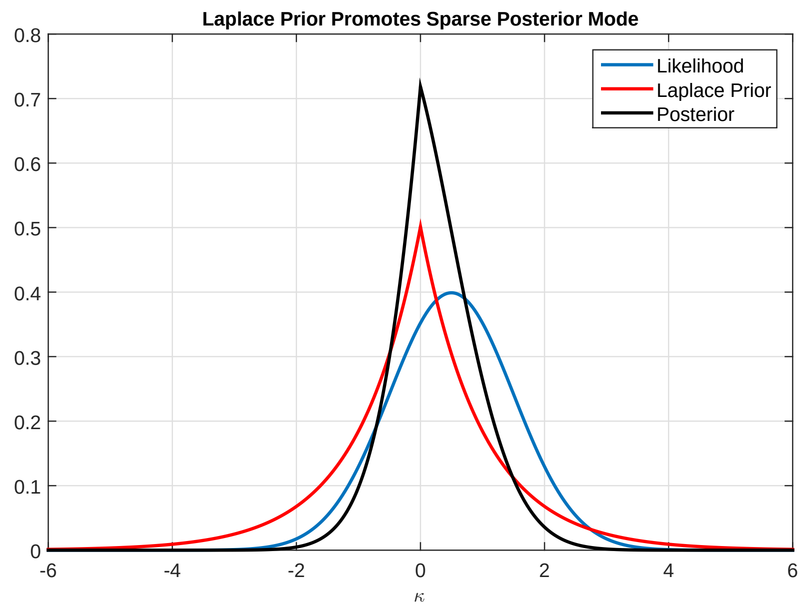

Figure 3.2: Here we show an example of a Laplace prior promoting a posterior mode at 0.

To make the model more robust we will not assume that each $\lambda_i(t)$ is known. Instead we will estimate $\lambda_i(t)$ from the data. Since the problem of estimating both $J_i(t)$ and $\lambda_i(t)$ is ill-posed we regularize it by introducing a prior distribution on each $\lambda_i(t)$ which we denote as $q(\lambda)$.

We wish to design the prior distribution q such that it induces a similar level of sparseness that is induced by the spike and slab prior f. To develop a criterion for designing q we first define a notion of similarity between $g(j|\lambda)$ and $f(j|, \zeta, \sigma_j^2)$.

Definition 1. *Let V be a zero-mean normal random variable with variance σ_v^2 and let $J_1 \sim$*

Laplace(λ') *and* $J_2 \sim SpikeSlab(\zeta', \sigma_j^{2'})$ *which are independent of* V. *Define*

$$\begin{aligned} Y_1 &= J_1 + V \\ Y_2 &= J_2 + V. \end{aligned}$$

Then $Laplace(\lambda')$ *is* σ_v^2***-equivalent*** *to* $SpikeSlab(\zeta', \sigma_j^{2'})$ *(denoted* $\lambda' \sim_{\sigma_v^2} (\zeta', \sigma_j^{2'})$ *) if*

$$\mathbb{E}_{p(y_2|J_2=0)} Pr(J_2 = 0|Y_2) = \mathbb{E}_{p(y_1|J_1=0)} Pr(\bar{J}_1 = 0) \tag{3.5}$$

where $\bar{J}_1$ *is the mode of* $p(j_1|Y_1)$.

To interpret the above definition assume that a jump has not occurred. Then $\lambda' \sim_{\sigma_v^2} (\zeta', \sigma_j^{2'})$ if the probability of falsely declaring a jump under the *Laplace*(λ') model (with MAP criterion) equals the average posterior probability of a jump under the spike and slab prior with parameters ζ' and $\sigma_j^{2'}$. Here σ_v^2 can be interpreted as the squared volatility of the diffusion component of the asset returns. Note that for each triplet $(\sigma_v^2, \zeta', \sigma_j^{2'})$ there is a unique λ' such that $\lambda' \sim_{\sigma_v^2} (\zeta', \sigma_j^{2'})$.

Since $(\sigma_v^2, \zeta', \sigma_j^{2'})$ are random and unobserved we cannot directly select a λ' such that $\lambda' \sim_{\sigma_v^2} (\zeta', \sigma_j^{2'})$. However the distribution of $(\sigma_v^2, \zeta_o', \sigma_j^{2'})$ induces a distribution on λ through the mapping $\sim_{\sigma_v^2}$. The resulting distribution can then be used as a prior $q(\lambda)$. The following section presents an example on how to construct a distribution for λ.

3.1.5 Procedure for selecting $q(\lambda)$

In this section we outline the method for selecting the distribution $q(\lambda)$ for a special case of when the prior distribution of volatility of each asset is identical. Suppose the squared volatility of each asset return is inverse gamma distributed with scale c and shape η. Let σ_v^2

be distributed as $IG(c,\eta)$ and be statistically independent of ζ' and σ_j^2.

To determine an appropriate prior distribution of λ we first obtain samples of λ, $(\tilde{\lambda}_1,\ldots,\tilde{\lambda}_{M_\lambda})$ by the performing the following steps

1. For $k = 1,\ldots,M_\lambda$

2. Draw independent samples from the distribution of $(\sigma_v^{2'},\zeta',\sigma_j^{2'})$. This is relatively straight forward using standard statistical functions due to the independence assumptions.

3. Determine a λ' such that $\lambda' \sim_{\sigma_v^{2'}} (\zeta',\sigma_j^{2'})$. This can be done via Monte Carlo integration as shown below.

    - For a large number L draw a sample $v_1 \ldots v_L$ from the distribution $\mathcal{N}(0,\sigma_v^2)$.
    - Compute $P_i = Pr(J = 0|J + V = v_i)$, where $J \sim SpikeSlab(\zeta',\sigma_j^{2'})$. The value of P_i is
    $$\frac{\frac{\zeta'}{\sqrt{\sigma_v^{2'}}}\exp(-v_i^2/(2\sigma_v^2))}{\frac{\zeta'}{\sqrt{\sigma_v^{2'}}}\exp(-v_i^2/(2\sigma_v^{2'})) + \frac{1-\zeta'}{\sqrt{\sigma_v^{2'}+\sigma_j^{2'}}}\exp(-v_i^2/(2(\sigma_v^{2'}+\sigma_j^{2'})))}.$$
    - Compute the simulated empirical mean $\bar{P} = \frac{1}{L}\sum_{i=1}^{L} P_i$.
    - Choose λ' such that (3.5) is satisfied with $\mathbb{E}_{p(y_2|J_2=0)}Pr(J_2 = 0|Y_2)$ approximated as $\bar{P}$. This value is given below
    $$\lambda' = \frac{\operatorname{erf}^{-1}(\bar{P})\sqrt{2\sigma_v^{2'}}}{\sigma_v^{2'}}$$
    where $\operatorname{erf}^{-1}()$ is the inverse error function.

4. Set $\tilde{\lambda}_k = \lambda'$

5. Goto step 1

Examples of histograms of samples obtained using the above procedures are shown in Figures 3.3 - 3.5. Once we obtain samples of λ we fit a smooth distribution to the sampled data. Since the gamma distribution is a conjugate prior to the Laplace distribution a gamma distribution is a convenient choice for $q(\lambda)$. Furthermore examination of Figures 3.3 -3.5 indicate that a gamma distribution is a reasonable approximation. Thus we choose

$$q(\lambda) = \frac{\beta_\lambda^{\alpha_\lambda}}{\Gamma_f(\alpha_\lambda)} \lambda^{\alpha_\lambda - 1} \exp(-\lambda \beta_\lambda)$$

where $\Gamma_f()$ is the gamma function. Here α_λ and β_λ can be selected using maximum likelihood or method of moments.

Since $q(\lambda)$ develops a singularity near zero for large values of β_λ we shall impose a prior of λ^{-1} rather than λ. We denote this prior as $q_{inv}(\lambda^{-1})$. Since λ is gamma distributed with shape α_λ and rate β_λ it follows that $q_{inv}(\lambda^{-1})$ is the inverse gamma distribution with shape α_λ and scale β_λ

$$q_{inv}(\lambda^{-1}) = \frac{\beta_\lambda^{\alpha_\lambda}}{\Gamma_f(\alpha_\lambda)} (\lambda^{-1})^{-\alpha_\lambda - 1} \exp\left(-\frac{\beta_\lambda}{\lambda^{-1}}\right).$$

3.2 KECM Approach to estimation of Γ

Maximum a posteriori (MAP) estimation of Γ with Kalman ECM (KECM) techniques is investigated in this section. The first ECM approach is an approximate technique where the prior distribution on the jumps is modeled as a Laplace distribution. The advantage of this approximation is that the conditional maximization steps in the ECM approach result in global (conditional) optimal solutions can be obtained. The disadvantage is that we are approximating the true spike and slab jump model. The second approach uses the spike and slab model for jumps, which is a true representation of the model presented in Section

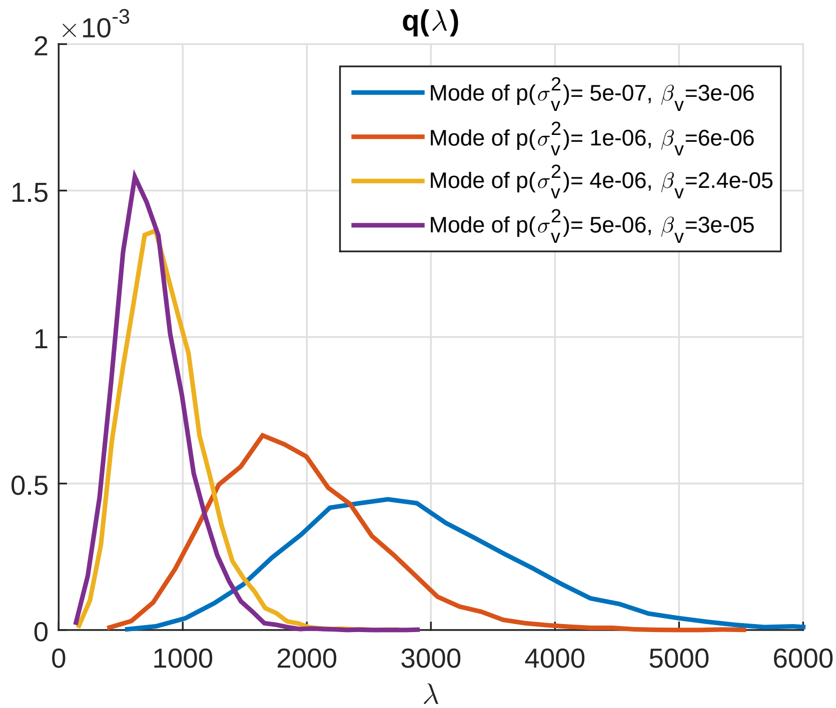

Figure 3.3: Normalized histograms of λ samples. In all experiments $\sigma_j^2 \sim IG(10, 0.0011)$,$\zeta \sim \text{Beta}(5, 1.0201)$, $\sigma_v^2 \sim IG(5, \beta_v)$.

3.1. However we will see that using the spike and slab jump model results in a non-concave optimization problem in the conditional M-step for J.

3.2.1 KECM algorithm for Laplace Distribution

First we consider a KECM approach to estimating Γ when J_i is approximated by a Laplace distributed random variable. We define

$$\Theta = [\Theta_1, \Theta_2, \Theta_3, \Theta_4, \Theta_5]$$

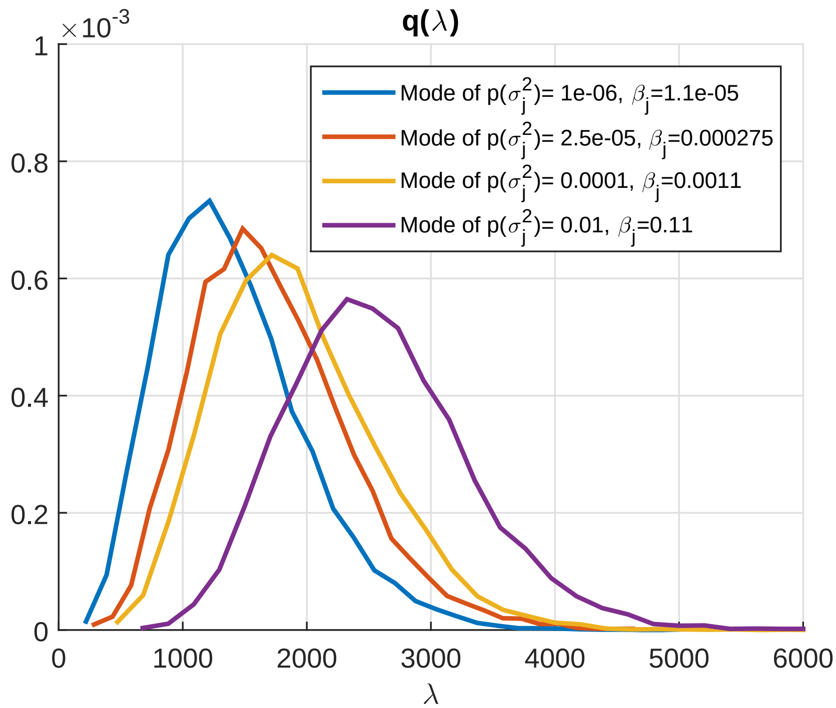

Figure 3.4: Normalized histograms of λ samples. In all experiments $\sigma_j^2 \sim IG(10, \beta_j)$,$\zeta \sim \text{Beta}(5, 1.0201)$, $\sigma_v^2 \sim IG(5, 6e-6)$.

where

$$\begin{aligned}
\Theta_1 &= D \\
\Theta_2 &= \Gamma \\
\Theta_3 &= \sigma_{o,i}^2, 1 \le i \le N \\
\Theta_4 &= J(2:T) \\
\Theta_5 &= \{\lambda_i(t)^{-1}\}_{1\le i\le N, 2\le t\le T}
\end{aligned}$$

as our vector of unknown parameters and $X(1:T)$ as the latent variables.

The KECM approach is an iterative algorithm that can be applied to the following problem

$$\Theta^* = \arg\max_{\theta} L(\theta)$$

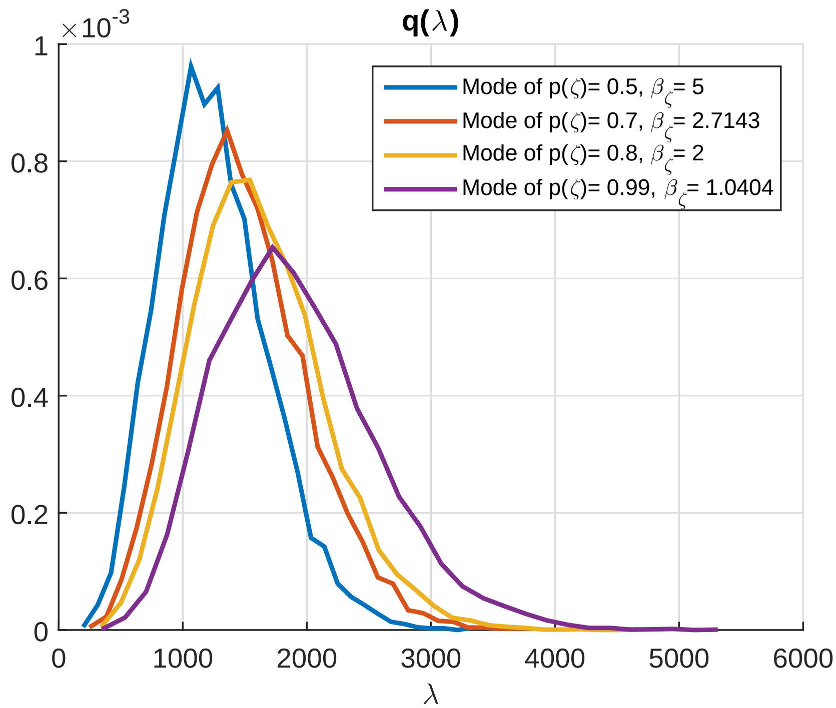

Figure 3.5: Normalized histograms of λ samples. In all experiments $\sigma_j^2 \sim IG(10, 0.0011)$,$\zeta \sim \text{Beta}(5, \beta_\zeta)$, $\sigma_v^2 \sim IG(5, 6e-6)$.

where $L(\theta)$ is the log posterior of Θ. In the KECM algorithm we iterate over E-steps and conditional M-steps to arrive at an estimate of Θ.

The E-step in the KECM algorithm involves computing the expected value of

$$\log p(X(1:T), y(1:T)|\theta)p(\theta)$$

with respect to $p(x(1:T)|y, \Theta^{(k)})$

$$\mathcal{G}(\theta, \Theta^{(k)}) = \mathbb{E}_{p(x|y,\Theta^{(k)})} \log p(X(1:T), y(1:T)|\theta) + \log(p(\theta))$$

where $\Theta^{(k)}$ is an estimate of Θ at the k^{th} iteration and where $p(\theta)$ is the prior distribution of parameters

$$p(\theta) = p(\theta_1)p(\theta_2)p(\theta_3)g(\theta_4, |\lambda)q_{inv}(\lambda^{-1}).$$

Here the complete log-likelihood is

$$\begin{aligned}
\log p(x, y|\theta) \quad = \quad & -0.5 \sum_{t=1}^{T} \log(|\Sigma_o(t)|) - \frac{1}{2} \sum_{t=1}^{T} \|y(t) - \tilde{I}(t)x(t)\|^2_{\mathrm{diag}(\Sigma_o(t)^{-1}),\ell_2} \\
& - \frac{T-1}{2} \log(|\Gamma|) \\
& - \frac{1}{2} \sum_{t=2}^{T} r(t)^T \Gamma^{-1} r(t) \\
& + const
\end{aligned}$$

where

$$r(t) = x(t) - x(t-1) - d - j(t).$$

and where

$$\|q\|^2_{\beta,\ell_2} = \sum_i \beta_i q_i^2.$$

It is well known that the function $\mathcal{G}(\theta, \Theta^{(k)})$ serves as a lower bound to $\log p(\theta, y)$ and that $\log p(\Theta^{(k)}, y) = \mathcal{G}(\Theta^{(k)}, \Theta^{(k)})$ [26].

The EM approach prescribes a joint maximization of $\mathcal{G}(\theta, \Theta^{(k)})$ with respect to θ. This is difficult due to the coupling of variables and the non-concavity of the problem. Conditional maximization of each parameter in turn is more tractable. Thus we apply conditional maximization as in the ECM [68] algorithm. The conditional M-steps involves a coordinate-wise maximization of $\mathcal{G}$. Here the conditional M-steps are

$$\begin{aligned}
\Theta_1^{(k+1)} &= \arg\max_{\theta_1} \mathcal{G}\left(\left[\theta_1, \Theta_2^{(k)}, \Theta_3^{(k)}, \Theta_4^{(k)}, \Theta_5^{(k)}\right], \Theta^{(k)}\right) \\
\Theta_2^{(k+1)} &= \arg\max_{\theta_2} \mathcal{G}\left(\left[\Theta_1^{(k+1)}, \theta_2, \Theta_3^{(k)}, \Theta_4^{(k)}, \Theta_5^{(k)}\right], \Theta^{(k)}\right) \\
\Theta_3^{(k+1)} &= \arg\max_{\theta_3} \mathcal{G}\left(\left[\Theta_1^{(k+1)}, \Theta_2^{(k+1)}, \theta_3, \Theta_4^{(k)}, \Theta_5^{(k)}\right], \Theta^{(k)}\right) \\
\Theta_4^{(k+1)} &= \arg\max_{\theta_4} \mathcal{G}\left(\left[\Theta_1^{(k+1)}, \Theta_2^{(k+1)}, \Theta_3^{(k+1)}, \theta_4, \Theta_5^{(k)}\right], \Theta^{(k)}\right) \\
\Theta_5^{(k+1)} &= \arg\max_{\theta_5} \mathcal{G}\left(\left[\Theta_1^{(k+1)}, \Theta_2^{(k+1)}, \Theta_3^{(k+1)}, \Theta_4^{(k+1)}, \theta_5\right], \Theta^{(k)}\right). \qquad (3.6)
\end{aligned}$$

Each of these problems can be readily solved as we will show later.

E-step of KECM

The posterior $p(x|y,\Theta^{(k)})$ needed to perform the E-step is normal and can be computed using a Kalman smoother [81]. By normality and the Markov property the posterior is completely defined by the following posterior moments for $m = T$

$$\bar{X}(t|m) \doteq \mathbb{E}(X(t)|y(1:m))$$

$$P(t|m) \doteq cov(X(t),X(t)|y(1:m))$$

$$P(t,t-1|m) \doteq cov(X(t),X(t-1)|y(1:m))$$

where $cov(:,:)$ refers to the covariance function. Equations for these quantities are derived in [82] and are stated in Appendix A.1. The expected value of log-posterior distribution with respect to the posterior of $X(1:T)$ can be shown to be

$$\begin{aligned}
\mathcal{G}(\theta,\Theta^{(k)}) &= \mathbb{E}_{p(x|y,\Theta^{(k)})}\log p(X(1:T),y(1:T)|\theta)+\log(p(\theta)) \\
&= -\frac{T-1}{2}\log(|\Gamma|)-\frac{1}{2}\mathrm{tr}(\Gamma^{-1}(C-B-B^T+A)) \\
&\quad -0.5\sum_{t=1}^{T}\log(|\Sigma_o(t)|) \\
&\quad -\frac{1}{2}\sum_{t=1}^{T}||y(t)-\tilde{I}(t)\bar{X}(t)||^2_{\mathrm{diag}(\Sigma_o(t)^{-1}),\ell_2}+\mathrm{tr}(P(t|T)\tilde{I}(t)^T\Sigma_o(t)^{-1}\tilde{I}(t)) \\
&\quad +\log(p(\theta))+const
\end{aligned} \tag{3.7}$$

where

$$A=\sum_{t=2}^{T}\left(P(t-1|T)+\bar{X}(t-1|T)\bar{X}(t-1|T)^T\right)$$

$$B=\sum_{t=2}^{T}\left(P(t,t-1|T)+(\bar{X}(t|T)-D^{(k)}-J^{(k)}(t))\bar{X}(t-1|T)^T\right)$$

$$C = \sum_{t=2}^{T} \left(P(t|T) + (\bar{X}(t|T) - D^{(k)} - J^{(k)}(t))(\bar{X}(t|T) - D^{(k)} - J^{(k)}(t))^T \right).$$

These equations are derived in Appendix A.2. For notational convenience the dependence of $P(t|m)$ and $P(t,t-1|m)$ on the iteration number has been dropped.

Conditional M-steps of KECM

For the conditional M-step it can be shown using standard conjugate prior relationships [37] that

$$D^{(k+1)} = F\left(\frac{1}{\sigma_D^2}\bar{D} + \Gamma^{(k)^{-1}} \sum_{t=2}^{T} \bar{X}(t|T) - \bar{X}(t-1|T) - J^{(k)}(t) \right) \tag{3.8}$$

and

$$\Gamma^{(k+1)} = \frac{1}{T-1+\eta}\left(A + C^{(k)} - B^{(k)} - B^{(k)T} \right) + \frac{1}{T-1+\eta} W \tag{3.9}$$

where

$$F = \left((T-1)\Gamma^{(k)^{-1}} + \sigma_D^{-2} I \right)^{-1}$$

$$B^{(k)} = \sum_{t=2}^{T} \left(P(t,t-1|T) + (\bar{X}(t|T) - D^{(k+1)} - J(t)^{(k)})\bar{X}(t-1|T)^T \right)$$

and

$$\begin{aligned} C^{(k)} &= \sum_{t=2}^{T} P(t|T) \\ &\quad + \sum_{t=2}^{T} (\bar{X}(t|T) - D^{(k+1)} - J(t)^{(k)})(\bar{X}(t|T) - D^{(k+1)} - J(t)^{(k)})^T. \end{aligned}$$

The conditional M-step for the observation noise variance is

$$\sigma_{o,i}^{2,(k+1)} = \frac{2\beta_o + \sum_{t\in\mathcal{T}_i} (y(t) - \tilde{I}(t)\bar{X}(t|T))^2_{\eta(i,t)} + (P(t|T))_{i,i}}{2\alpha_o + 2 + M_i}. \tag{3.10}$$

Here $\mathcal{T}_i$ is the set of times where the price of asset i is observed and M_i is the total number of prices observed for asset i. The subscript $\eta(i,t)$ is the row number of $\tilde{I}(t)$ such that $\tilde{I}(t)_{\eta(i,t),i} = 1$.

For each conditional M-step $P(t,T)$, $P(t,t-1|T)$ and $\bar{X}(t|T)$ are evaluated with respect to $p(X(1:T)|Y,\Theta^{(k)})$.

To compute the conditional M-step for J we denote

$$Q(j) \doteq \mathcal{G}([\Gamma^{(k+1)}, D^{(k+1)}, \{\sigma^2_{o,i}\}_{1\le i\le N}, j, \{\lambda_i(t)^{-1}\}^{(k)}_{1\le i\le N, 2\le t\le T}], \Theta^{(k)}).$$

Then up to a constant not depending on j

$$\begin{aligned}
Q(j) &= -\frac{1}{2}\sum_{t=2}^{T}(\bar{X}(t|T) - j(t) - D^{(k+1)})^T(\Gamma^{(k+1)})^{-1}(\bar{X}(t|T) - j(t) - D^{(k+1)}) \\
&\quad + \sum_{t=2}^{T}(\bar{X}(t|T) - j(t) - D^{(k+1)})^T(\Gamma^{(k+1)})^{-1}\bar{X}(t-1|T) \\
&\quad + \log(g(j(2:T)|\{\lambda_i(t)\}^{(k)}_{1\le i\le N, 2\le t\le T})) + const_1 \\
&= -\frac{1}{2}\sum_{t=2}^{T}(\bar{X}(t|T) - j(t) - D^{(k+1)})^T(\Gamma^{(k+1)})^{-1}(\bar{X}(t|T) - j(t) - D^{(k+1)}) \\
&\quad + \sum_{t=2}^{T}(\bar{X}(t|T) - j(t) - D^{(k+1)})^T(\Gamma^{(k+1)})^{-1}\bar{X}(t-1|T) \\
&\quad - \sum_{t=2}^{T}\|j(t)\|_{\lambda(t),\ell_1} + const_2.
\end{aligned}$$

where $\|j(t)\|_{\lambda(t),\ell_1} = \sum_{n=1}^{N}\lambda_n(t)|j_n(t)|$. By rearranging terms we can express $Q(j)$ as a quadratic function of j

$$\begin{aligned}
Q(j) &= -\frac{1}{2}\sum_{t=2}^{T} j(t)^T(\Gamma^{(k+1)})^{-1}j(t) \\
&\quad + \sum_{t=2}^{T}(\bar{X}(t|T) - D^{(j+1)} - \bar{X}(t-1|T))^T(\Gamma^{(k+1)})^{-1}j(t) \\
&\quad - \sum_{t=2}^{T}\|j(t)\|_{\lambda(t),\ell_1} + const_3.
\end{aligned}$$

Referring to equation (3.6) we see that $J^{(k+1)}(t)$ is the solution of the following ℓ_1 penalized quadratic program

$$J^{(k+1)}(t) = \arg\min_j \frac{1}{2} j^T (\Gamma^{(k+1)})^{-1} j - j^T (\Gamma^{(k+1)})^{-1} \Delta^{(k+1)} + \|j(t)\|_{\lambda(t),\ell_1} \tag{3.11}$$

where

$$\Delta^{(k)}(t) = \bar{X}(t|T) - D^{(k)} - \bar{X}(t-1|T). \tag{3.12}$$

This problem can be solved with a variety of fast algorithms such as ADMM [11] and FISTA [7].

Now we determine $\{\lambda_i(t)^{-1}\}_{1 \le i \le N, 2 \le t \le T}$ which depends only on $q_{inv}(\lambda^{-1})$ and $p(j|\lambda)$. Using conjugate prior relationships we have $p(\lambda_i(t)^{-1}|j_i(t))$ is inverse gamma distributed with shape $\alpha_\lambda + 1$ and scale $\beta_\lambda + |j_i(t)|$. Thus the conditional MAP estimate is

$$\lambda_i(t)^{-1} = \frac{|J_i^{(k+1)}(t)| + \beta_\lambda}{\alpha_\lambda + 2}. \tag{3.13}$$

which implies that

$$\lambda_i(t) = \frac{\alpha_\lambda + 2}{|J_i^{(k+1)}(t)| + \beta_\lambda}. \tag{3.14}$$

An outline of the KECM algorithm for Laplace jump models is given below.

Algorithm 1 KECM Algorithm for estimation of Γ under Laplace Prior

Initialize: $\Theta^{(0)}, k = 0$

while not converged **do**

Compute $\bar{X}(t|T), P(t|T), P(t,t-1|T)$ using Kalman smoothing equations for $\Theta^{(k)}$ using equations (A.1)-(A.6)

Compute $D^{(k+1)}, \Gamma^{(k+1)}$, and $\sigma_{o,i}^{2,(k+1)}$ using equations (3.8),(3.9), and (3.10) respectively

Compute $J^{(k+1)}$ by solving (3.12)

Compute $\{\lambda_i(t)\}_{1\le i\le N, 2\le t\le T}$ by solving (3.14)

$k = k + 1$

end while

Convergence results for this algorithm are given in Appendix A.3.

Remark. Since of value of $\{\lambda_i(t)\}_{1\le i\le N, 2\le t\le T}$ changes with each iteration we see that we effectively reweight the ℓ_1 penalty in (3.11) after each iteration. Reweighting of the ℓ_1 norm has been proposed in several papers and has been shown to have improved performance in compressive sensing problems versus a fixed set of weights [17].

3.2.2 KECM approach for the Spike and Slab Jump Prior

Now we present a KECM for the spike and slab jump prior. As with the Laplace prior we treat X as a latent variable. Let us denote the unknown parameters as Φ where

$$
\begin{aligned}
\Phi_1 &= D \\
\Phi_2 &= \Gamma \\
\Phi_3 &= \sigma_{o,i}^2 \\
\Phi_4 &= Z(2:T), \tilde{J}(2:T) \\
\Phi_5 &= \zeta \\
\Phi_6 &= \left\{\sigma_{j,i}^2(t)\right\}_{i=1,\dots,N,t=1,\dots,T}.
\end{aligned}
$$

Here we allow for distinct σ_j^2 values for each time and asset.

The E-step as well as the conditional M-steps for Φ_1, Φ_2, Φ_3 are identical to the KECM algorithm for Laplace priors. The differences for this section are in the conditional M-steps for J, ζ and σ_j^2.

First we address the conditional M-step for $J^{(k+1)}$. Here we need to solve

$$
\begin{aligned}
[J^{(k+1)}(t), Z^{(k+1)}(t), \tilde{J}^{(k+1)}(t)] &= \arg\min_{j,z,\tilde{j}} \frac{1}{2} j^T(\Gamma^{(k+1)})^{-1} j \\
&\quad - j^T(\Gamma^{(k+1)})^{-1}\Delta^{(k+1)} - \sum_{i=1}^{N} \log(f(\tilde{j}_i, z_i)) \\
&\quad \text{s.t. } j_i = \tilde{j}_i z_i
\end{aligned}
\tag{3.15}
$$

where

$$
\log f(\tilde{j}_i, z_i) = \log\left(\zeta 1_{z_i=0} + (1-\zeta)1_{z_i=1}\right) + \log\left(\frac{1}{\sqrt{2\pi\sigma_{j,i}^2}} \exp\left(-\frac{\tilde{j}_i^2}{2\sigma_{j,i}^2}\right)\right). \tag{3.16}
$$

Here we dropped the notation for time dependence. When restricted to $j_i = \tilde{j}_i z_i$, $-\log(\tilde{j}_i, z_i)$ induces a penalty on j_i. which is a weighted sum of an ℓ_0 and squared ℓ^2 norm. A plot of this penalty is shown in Figure 3.6.

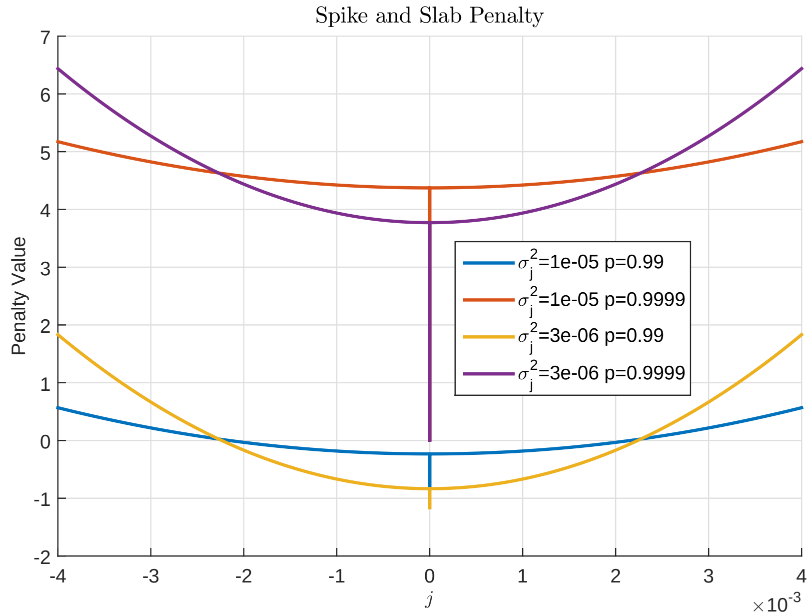

Figure 3.6: Spike and slab penalty function for various parameter values. Here we see that the penalty is a weighted sum of ℓ_0 and squared ℓ_2 norms.

The term $-\log(\tilde{j}_i, z_i)$ is non-convex and complicates the conditional M-step (3.15). Hence we seek an approximate maximization through coordinate descent. Here we divide the problem into tractable 1-dimensional optimization problems with respect to one asset at a time. The method and equations for implementing coordinate descent are derived in Appendix A.4.2 and described below. For ease of notation we drop the notation denoting dependence on k.

Let us define the following conditional mean and variance

$$a(i) = \Delta_i(t) + \Gamma_{i,-i}\Gamma^{-1}_{-i,-i}(j_{-i}(t) - \Delta_{-i}(t)) \tag{3.17}$$

and

$$b^2(i) = \Gamma_{i,i} - \Gamma_{i,-i}\Gamma^{-1}_{-i,-i}\Gamma_{-i,i} \tag{3.18}$$

where the subscript $-i$ is to be interpreted as all indices except i. Then the following rule determines the MAP optimal value of $z_i(t)$ conditioned on $j_{-i}(t)$

$$z_{i|-i}(t) = \begin{cases} 0 \text{ if } \frac{\zeta}{1-\zeta}\mathcal{N}(0, a(i), b^2(i)) > \mathcal{N}(0, a(i), b^2(i) + \sigma_{j,i}^2(t)) \\ 1 \text{ else} \end{cases} \tag{3.19}$$

where $\mathcal{N}(0, a(i), b^2(i))$ is the normal PDF with mean $a(i)$ and variance $b^2(i)$ evaluated at 0. An optimal value of $\tilde{J}_{i|-i}(t)$ is then given as

$$\tilde{J}_{i|-i}(t) = \begin{cases} \frac{a}{1+b^2\sigma_{j,i}^{-2}(t)} \text{ if } z_{i|-i}(t) \neq 0 \\ 0 \text{ else} \end{cases} . \tag{3.20}$$

The mapping defined by equations (3.19) and (3.20) is a combination of a thresholding step followed by a shrinkage operation

$$\begin{aligned} J_{i|-i}(t) &= SpikeSlabShrink(a, b^2) \\ &\doteq \begin{cases} 0 & \text{if } \frac{\zeta\mathcal{N}(0,a(i),b^2(i))}{(1-\zeta)\mathcal{N}(0,a(i),b^2(i)+\sigma_{j,i}^2(t))} > 1 \\ \frac{a(i)}{1+b^2(i)\sigma_{j,i}^{-2}(t)} & \text{else} \end{cases} . \end{aligned} \tag{3.21}$$

This spike and slab shrinkage is illustrated in Figure 3.7. As the plots indicate the shrinkage is discontinuous and large values are shrunk more than smaller values.

Equation (3.21) is cycled through all $i = 1 \ldots N$. Multiple cycles may also be performed to obtain an improved estimate of J. A summary of the algorithm for the conditional M-step for J is given below in Algorithm 2.

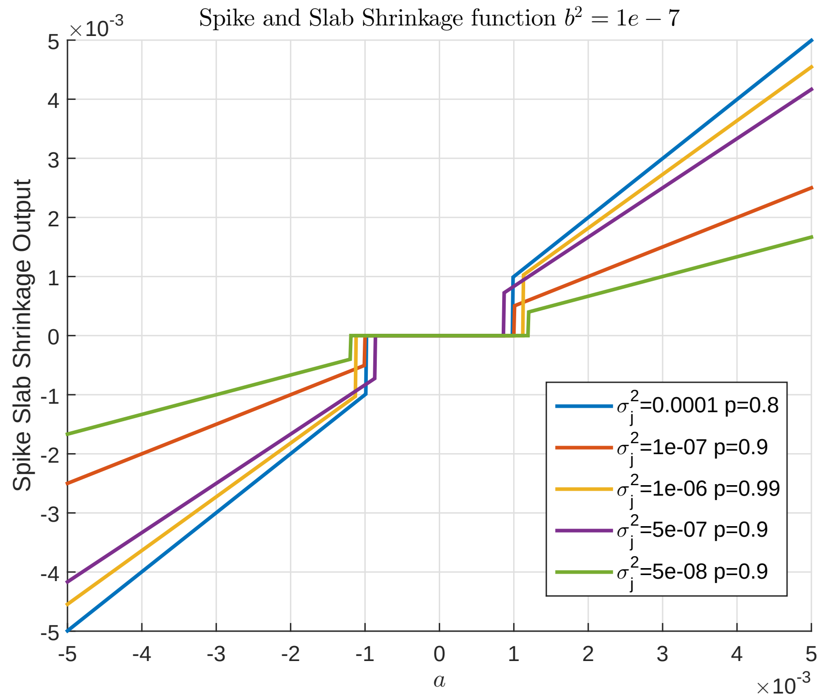

Figure 3.7: Spike and slab shrinkage function for various parameter values

---

Algorithm 2 Coordinate Descent for Determination of $Z^{(k+1)}(t)$,$\tilde{J}^{(k+1)}(t)$, and $J^{(k+1)}(t)$

Initialize: Set $J^{(k+1)}(t) = J^{(k)}(t)$, it=0, $L > 0$

while $it \leq L$ **do**

  $it = it + 1$

  $i = 0$

  while $i < N$ **do**

    $i = i + 1$

    Compute $Z_i^{(k+1)}(t)$ using equations (3.17), (3.18), and (3.19)

    Compute $\tilde{J}_i^{(k+1)}(t)$ using equations (3.17), (3.18), and (3.20)

    Set $J_i^{(k+1)}(t) = Z_i^{(k+1)}(t)\tilde{J}_i^{(k+1)}(t)$

  end while

end while

return $J^{(k+1)}(t)$

Although this method is not guaranteed to solve (3.15) it will not increase the value of the objective function compared with $J^k(t)$.

Once $J^{(k+1)}$ is obtained, values for $\zeta^{(k+1)}$ and $\sigma_j^{2,(k+1)}$ are easily computed through conjugate prior relationships. First let N_Z be number of zero values in $J(2:T)^{(k+1)}$. Then by conjugate prior relationships the conditional M-steps for ζ and σ_j^2 are

$$\zeta^{(k+1)} = \frac{\alpha_\zeta + N_Z}{N(T-1) + \beta_\zeta + \alpha_\zeta} \tag{3.22}$$

and

$$\sigma_{j,i}^{2,(k+1)}(t) = \frac{\beta_j + 0.5(J_i(t))^2}{\alpha_j + 1 + 0.5(Z_i(t))}. \tag{3.23}$$

The KECM algorithm for spike and slab models is summarized in Algorithm 3.

Algorithm 3 KECM Algorithm for estimation of Γ under Spike and Slab Prior

---

Initialize: $\Phi^{(0)}, k = 0$

while not converged **do**

Compute $\bar{X}(t|T), P(t|T), P(t,t-1|T)$ using Kalman smoothing equations for $\Theta^{(k)}$ using equations (A.1)-(A.6)

Compute $D^{(k+1)}, \Gamma^{(k+1)}$, and $\sigma_{o,i}^{2,(k+1)}$ using equations (3.8), (3.9), and (3.10) respectively

For all t, compute $\tilde{J}^{(k+1)}(t)$, $Z^{(k+1)}(t)$ using Algorithm 2

Set $J_i^{(k+1)}(t) = Z_i^{(k+1)}(t)\tilde{J}_i^{(k+1)}(t)$

Compute $\zeta^{(k+1)}$ using equation (3.22)

Compute $\sigma_{j,i}^{2,(k+1)}(t)$ using equation (3.23) for all i, t

$k = k + 1$

end while

Note that although $J(t)$ is only approximately maximized in each conditional M-step this is still an ECM algorithm. To see this we can simply redefine Φ as

$$\left[D, \Gamma, \sigma_o^2, J_1(2), \ldots, J_N(2), \ldots, J_1(T), \ldots, J_N(T), \zeta, \sigma_j^2\right].$$

Then the above algorithm is an ECM algorithm for the redefined parameter vector. The convergence of Algorithm 3 is similar to the proof of the convergence of Algorithm 1 in Appendix A.3.

Remark. A comparison of the spike and slab shrinkage function with the shrinkage function of the b^2-*equivalent* Laplace prior is shown in Figure 3.8. The Laplace shrinkage function (with parameter λ) is defined as

$$LaplaceShrink(a, b^2) \doteq \begin{cases} a - \lambda b^2 & \text{if } a > \lambda b^2 \\ a + \lambda b^2 & \text{if } a < -\lambda b^2 \\ 0 & \text{else} \end{cases}.$$

The graphs illustrate advantages and disadvantages of the Laplace prior. One notable disadvantage is that for large σ_j^2 the Laplace prior has a large bias relative to spike and slab priors. However for small σ_j^2 and large values of a we see that the Laplace prior is less biased than the spike and slab. This can be attributed to the quadratic penalty induced by the spike and slab prior which penalizes large jumps more heavily than the Laplace prior.

Remark. *The use of Laplace priors and ℓ_1 penalties has been applied in context of robust Kalman filtering and smoothing in [67, 1]. Here the authors considered the problem of non-gaussian heavy tailed observation noise rather than process noise.*

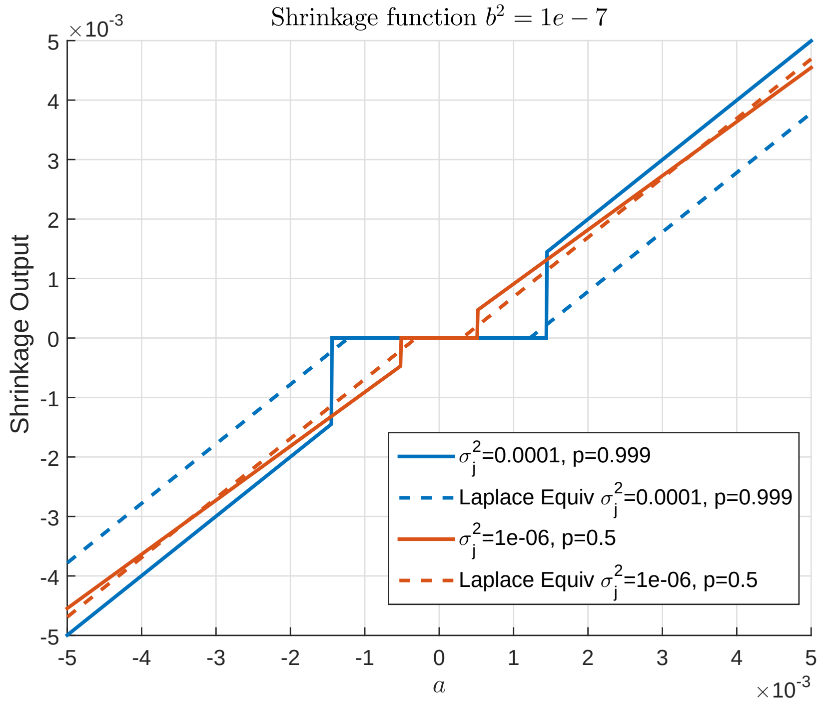

Figure 3.8: Shrinkage Functions of the spike and slab and the corresponding $b^2 - equivalent$ Laplace prior

3.3 Bayesian Approach using MCMC

In this section we consider a fully Bayesian approach where we estimate the posterior distribution of Γ. The advantages of the fully Bayesian approach to this problem are

1. Uncertainty in nuisance parameters such as J and σ_o^2 are averaged out rather than relying on MAP point estimates

2. Estimate of the posterior distribution of Γ is obtained which provides more information than a posterior mode

3. Estimates of uncertainty in covariance estimate can be obtained.

To describe the Bayesian approach to estimation of Γ, let Φ represent the unknown parameters

$$
\begin{aligned}
\Phi_0 &= Y_{miss} \\
\Phi_1 &= X(1:T) \\
\Phi_2 &= J(2:T) \\
\Phi_3 &= D \\
\Phi_4 &= \{\sigma_{o,i}^2\}_{1\leq i\leq N} \\
\Phi_5 &= \zeta \\
\Phi_6 &= \left\{\sigma_{j,i}^2(t)\right\}_{i=1,\ldots,N,t=1,\ldots,T}
\end{aligned}
\tag{3.24}
$$

where Y_{miss} are the unobserved prices. Unlike the KECM approaches in the previous section we sample the missing observations Y_{miss}. One advantage of sampling Y_{miss} is that the covariance of $X(t)|X(t-1), X(t+1), Y(t), Y_{miss}(t)$ is the same for all $2 \leq t \leq T-1$, where as the covariance of $X(t)|X(t-1), X(t+1), Y(t)$ depends on t. This simplification allows for faster numerical simulation in the Gibbs sampler.

In the Bayesian approach given the data $Y(t)$ we wish to compute the posterior distribution of Γ

$$
p(\gamma|y) = \frac{\int p(y|\phi,\gamma)p(\phi,\gamma)d\phi}{\int\int p(y|\phi,\gamma')p(\phi,\gamma')d\phi d\gamma'}. \tag{3.25}
$$

Once the posterior is obtained the posterior mean of Γ can be obtained via

$$
\mathbb{E}_{|y}\Gamma = \int \gamma p(\gamma|y)d\gamma. \tag{3.26}
$$

The posterior mean which is optimal in a minimum mean squared error (MMSE) sense can be used as an estimate of Γ.

Evaluating the integrals in (3.25) and (3.26) are intractable, however we can obtain samples from the posterior distribution using a Markov Chain Monte Carlo (MCMC) technique such as Gibbs sampling. These samples can then be used to obtain an estimate of $\mathbb{E}_{|y}\Gamma$. A Gibbs sampling approach for estimating $\mathbb{E}_{|y}\Gamma$ is described in the next section.

3.3.1 Gibbs Sampling approach

Gibbs sampling [19] is an MCMC approach for generating samples from a multivariable distributions such as $p(\phi, \gamma|y)$. In this application Gibbs sampling may be implemented as follows to generate samples of $\Gamma^{(1)}, \dots, \Gamma^{(M_G)}$ from $p(\gamma|y)$.

1. Initialize the first samples,$\Phi^{(0)}$,$\Gamma^{(0)}$
2. for $k = 1$ to M_G
   - Sample $Y_{miss}^{(k)}$ from the conditional distribution $p(Y_{miss}|X^{(k-1)}, \sigma_o^2)$
   - for $t = 1$ to T
     - Sample $X(t)^{(k)}$ from

$$p(x(t)|y, Y_{miss}^{(k)}, \Phi_{2:6}^{(k-1)}, X^{(k)}(1:t-1), X^{(k-1)}(t+1:T), \Gamma^{(k-1)})$$

   - for $l = 1$ to L
     - for $t = 1$ to T,$n = 1$ to N
       * Sample $J_n(t)^{(k-1+l/L)}$ from

$$p(j_n(t)|y, \Phi_{0,1}^{(k)}, J_{1:n-1}^{(k-1+l/L)}(t), J_{n+1:N}^{(k-1+(l-1)/L)}(t), \Gamma^{(k-1)})$$

   - Sample D from $p(d|y, \Phi_{0:2}^{(k)}, \Phi_{4:6}^{(k-1)}, \Gamma^{(k-1)})$
   - Sample $\Gamma^{(k)}$ from $p(\gamma|y, \Phi_{0:3}^{(k)}, \Phi_{4:6}^{(k-1)})$

- Sample $\sigma_o^{2,(k)}$ from $p(\sigma_o^2|y, \Phi_{0:3}^{(k)}, \Phi_{5:6}^{(k-1)}, \Gamma^{(k)})$
- Sample $\zeta^{(k)}$ from $p(\zeta|y, \Phi_{0:4}^{(k)}, \Phi_6^{(k-1)}, \Gamma^{(k)})$
- Sample $\sigma_{j,i}^{2,(k)}(t)$ from $p(\sigma_{j,i}^2(t)|y, \Phi_{0:5}^{(k)}, \Gamma^{(k)})$ for all i, t

where $\Phi_{i:j}$ refers to $[\Phi_i, \ldots, \Phi_j]$ and where

$$\Phi_{-n}(t) = [\Phi_0(t), \ldots, \Phi_{n-1}(t), \Phi_{n+1}(t), \ldots, \Phi_6(t)]$$

Each of these steps draws from conditional distributions can be implemented easily as shown in Appendix A.4.

It can be shown using well known results on Markov chains that the samples produced by the above Gibbs sampler form a Markov chain [77] with a limiting stationary distribution $p(\phi, \gamma|y)$.

3.3.2 Estimation of Γ

The samples of $\Gamma^{(k)}$ are used as an estimate of the posterior distribution of Γ. Using the estimated posterior distribution the posterior mean of Γ is the sample average of $\Gamma^{(k)}$ (where we discarded earlier samples to allow for the samples to converge)

$$\hat{\Gamma} = \frac{1}{M_G - k + 1} \sum_{m=k}^{M_G} \Gamma^{(m)}.$$

Another technique to estimate Γ is Rao-Blackwellization which reduces the variance in the covariance estimate [61]. Here we take the sample of average of the conditional means to arrive at an estimate of the posterior mean of Γ

$$\hat{\Gamma} = \frac{1}{M_G - k + 1} \sum_{m=k}^{M_G} \mathbb{E}(\Gamma|\Phi^{(k)}).$$

The numerical experiments presented in the next section use Rao-Blackwellization for the posterior mean estimation.

3.4 Numerical Results

In this section we evaluate the performance of the following algorithms

1. KEM [22]
2. KECM Laplace(section 3.2)
3. KECM Spike and Slab (section 3.2)
4. MCMC approach (section 3.3)
5. Pairwise refresh with TSCV [33, 90]
6. Pairwise refresh with TSCV and jump correction [10]

for determining a covariance matrix from high frequency data. The performance is evaluated using a Monte Carlo approach with simulated high frequency return data.

3.4.1 Performance Assessment Methodology

We track two performance measures for the covariance estimate, $\hat{\Gamma}$, in this study. For the first performance measure we compute the minimum variance portfolio

$$\begin{aligned} \tilde{w} &= \arg\min_{w} w^T \hat{\Gamma} w \\ \text{s.t.} &\sum_i w_i = 1. \end{aligned}$$

The variance of this portfolio's return is then computed as a figure of merit. The variance of the portfolio return is given below

$$\tilde{w}^T \Gamma \tilde{w}.$$

For the second performance measure we compute the relative Frobenius norm of the error between the true and estimated covariance

$$\frac{\sqrt{\sum_{i,j} |\Gamma_{i,j} - \hat{\Gamma}_{i,j}|^2}}{\sqrt{\sum_{i,j} |\Gamma_{i,j}|^2}}.$$

3.4.2 Algorithm Initialization and other considerations

In each study we initialize the algorithms in the same way. The hyper-parameters for the prior distribution are listed in Table 3.1. For the KEM and KECM algorithms the initial covariance estimate is computed using the time refresh method in [5]. The initialization of drift and jump estimate of each algorithm is set to zero. For the MCMC algorithm we take the output of the KECM spike and slab algorithm as the first sample.

In the KECM algorithms we employ one additional initialization step to avoid being trapped in an over-smoothed local solution. This step involves using a forward Kalman filter rather than a smoother to approximate the posterior distribution of $X(t)$ in the first 10 iteration of the KECM algorithms. After 10 iterations we revert to the approaches described in Section 3.2 which use the Kalman smoother.

The stability of the covariance estimate forms the basis for a stopping criterion in the KECM algorithms. The KECM algorithms are terminated at iteration n when the relative difference between the current and previous covariance estimate is less than 0.001

$$\frac{\sqrt{\sum_{i,j}|\hat{\Gamma}_{i,j}^{(n)} - \hat{\Gamma}_{i,j}^{(n-1)}|^2}}{\sqrt{\sum_{i,j}|\hat{\Gamma}_{i,j}^{(n-1)}|^2}} < 0.001.$$

For the MCMC algorithm we generate 10000 samples and discard the first 2000 samples to allow for convergence of the Markov chain.

Since jumps cannot be predicted an ambiguity occurs if there is no observation of the price at the time the jump occurs. Thus to prevent ambiguity we assume jumps in the i^{th} asset price can only occur if an observation of the i^{th} price is made. We believe that this is a mild assumption given that in many markets jumps in the efficient price will be traded upon almost immediately. This assumption is built into the KECM and MCMC approaches by setting $\lambda = \infty$ and $\zeta = 1$ when an observation does not occur.

3.4.3 Simulated Data Jump Model

For the data study we simulated 30 minutes of data from 20 assets according to equations (3.1) and (3.3) at 1 second intervals. Here 50 data sets were generated to test our algorithms. Taking motivation from factor models for U.S. stock returns we set our covariance Γ according to the following 5 factor model

$$\Gamma = \sum_{i=1}^{5} \beta_{v_i} v_i v_i^T + \epsilon I.$$

Here we compute a new covariance for each Monte Carlo data set. We draw v_1 from a multivariate normal distribution with mean $\frac{1}{\sqrt{2}}$ and covariance $0.5I$. For $i > 1$, we draw v_i from a multivariate normal distribution with zero mean and covariance I. The factor variance β_{v_i} is modeled as gamma distributed with shape 2 and mean $\frac{0.7*0.02^2}{23400}$ for $i = 1$ and mean $\frac{0.3/4*0.02^2}{23400}$ for $i \neq 1$. The ϵ term is defined to be $\frac{0.02^2}{23400*100}$. With these settings each simulated asset will on average have a daily return volatility of approximately 2 percent.

	Value	Comment
α_ζ	10×0.995	
β_ζ	$10 - \alpha_\zeta$	prior mean of ζ is 0.995
α_j	10	
β_j	$0.01^2(\alpha_j + 1)$	prior mode of σ_j^2 is 1e-4
α_o	5	
β_o	$(\alpha_o + 1) \times 0.0001^2$	prior mode of σ_o^2 is 1e-8
η	$N + 5$	
W_o	$\frac{0.02^2(\eta+N+1)}{23400} I$	Corresponds to 0.02% daily volatility
α_λ	5.6	Obtained using method in Section 3.1.4
β_λ	5e-04	Obtained using method in Section 3.1.4

Table 3.1: Parameters used in KEM, KECM and MCMC algorithms

For the D parameter we use a random number generator for each data set. The value for D was drawn from a multi-variate normal distribution with mean 0 and covariance $\left(\frac{0.01}{23400}\right)^2 I$. The observation noise variance of each asset was set to a random number drawn from a gamma distribution with shape 2 and mean 0.0002^2. For a stock price of \$25 this corresponds to a mean noise standard deviation of about \$0.005. The jump parameters ζ and σ_j^2 were varied parametrical over several values.

Both the KECM and MCMC algorithms require hyperparameters to be specified for the prior distributions. For these experiments we choose hyperparameters which would result in diffuse priors in order to minimize bias. For the hyperparameters of the Laplace prior in the KECM algorithm we used the technique described in Section 3.1.5. A listing of all the hyperparameters used in the algorithms are shown in Table 3.1.

The probability that any given price is observed is set to be commensurate with the price innovation. This is consistent with empirical observations that trading volume can be positively correlated with volatility [51]. To model this association the probability that the m^{th} asset price will be observed at time t is simulated as

$$p_{obs,m}(t) = \frac{|X_m(t) - X_{m-1}(t) - D_m|}{|X_m(t) - X_{m-1}(t) - D_m| + \nu}$$

ζ	σ_j^2	KEM	KECM Laplace	KECM Spike & Slab	MCMC	Pairwise Refresh	Pairwise Refresh (jump)
1	N/A	1.2e-10	1.3e-10	1.3e-10	1.3e-10	1.8e-10	2.3e-10
0.9999	6.25e-06	1.5e-10	1.4e-10	1.4e-10	1.5e-10	1.8e-10	3.3e-10
0.9999	0.0001	1.6e-10	1.4e-10	1.4e-10	1.5e-10	2.6e-10	3.5e-10
0.9995	6.25e-06	1.6e-10	1.3e-10	1.3e-10	1.3e-10	2.4e-10	3.5e-10
0.9995	0.0001	3e-10	1.3e-10	1.2e-10	1.3e-10	7.9e-10	4.4e-10
0.999	6.25e-06	2.4e-10	1.6e-10	1.6e-10	1.7e-10	4.7e-10	4.1e-10
0.999	2.5e-05	4.5e-10	1.7e-10	1.7e-10	1.8e-10	9.8e-10	4.3e-10
0.999	0.0001	8.2e-10	1.6e-10	1.6e-10	1.7e-10	1.7e-09	6.4e-10

Table 3.2: Portfolio variance for jump model, best performance highlighted in green.

where

$$\nu = \frac{\sqrt{2\Gamma_{m,m}}}{\pi}\left(\frac{1}{p_{Obs}} - 1\right).$$

This choice of ν ensures that when the innovation achieves its mean absolute value ,$\sqrt{\frac{2\Gamma_{m,m}}{\pi}}$, the probability of an observation will be p_{Obs}. We set $p_{Obs} = 0.3$ in our numerical experiments.

The performance results for different values of the jump parameters are shown Tables 3.2 and 3.3. For the majority of cases we see that the KECM approaches outperform the other methods when jumps are present. In Figure 3.9 we show the Kalman estimate of the true price for various algorithms. The figure highlights the disadvantage of the KEM algorithm in the presence of jumps, namely that it over smoothes prices near jumps.

3.4.4 Simulated Data from GARCH(1,1)-jump model

In addition to the jump diffusion model we also evaluate the algorithms against a multivariate GARCH(1,1)-jump pricing model [21, 65, 8], where the effect of jumps persists in the price volatility. Using the GARCH(1,1)-jump model the log- price data is generated as

$$X_i(t) = X_i(t-1) + \sqrt{h_i}V_i(t) + J_i(t)Z_i(t) + D$$

ζ	σ_j^2	KEM	KECM Laplace	KECM Spike & Slab	MCMC	Pairwise Refresh	Pairwise Refresh (jump)
1	N/A	0.2	0.2	0.2	0.22	0.48	0.51
0.9999	6.25e-06	0.22	0.22	0.22	0.24	0.47	0.52
0.9999	0.0001	0.73	0.21	0.21	0.22	0.89	0.56
0.9995	6.25e-06	0.29	0.21	0.21	0.22	0.55	0.58
0.9995	0.0001	3.5	0.18	0.18	0.2	2.9	0.57
0.999	6.25e-06	0.36	0.21	0.21	0.22	0.67	0.63
0.999	2.5e-05	1.1	0.21	0.21	0.22	1.5	0.68
0.999	0.0001	4.8	0.2	0.2	0.21	4.6	0.73

Table 3.3: Average covariance error for jump model, best performance highlighted in green. Large errors highlighted in red.

$$h_i(t+1) = b_i h_i(t) + a_i(X_i(t) - X_i(t-1) - D)^2 + c_i$$

$$h_i(0) = \Gamma_{i,i}$$

where a_i, b_i, c_i are non-negative with $b_i + a_i < 1$ and $c_i = \Gamma_{i,i}(1 - a_i - b_i)$. Here $V(t)$ is modeled as multivariate normal with

- $V_i(t) \sim \mathcal{N}(0, 1)$
- $\mathbb{E}V_i(t)V_j(t) = \frac{\Gamma_{i,j}}{\sqrt{\Gamma_{i,i}\Gamma_{j,j}}}$
- $\mathbb{E}V_i(t_1)V_j(t_2) = 0$ for $t_1 \neq t_2$.

The value of c ensures that in the absence of jumps, the long term average volatility for the i^{th} asset will be $\sqrt{\Gamma_{i,i}}$. We also see that the correlation coefficient between any two assets is constant [8].

In these experiments $a_i = 0.3$ and $b_i = 0.5$. This allows for volatility clustering which has been observed in many empirical stock return data. All other parameters such as the covariance matrix are identical to the previous experiment.

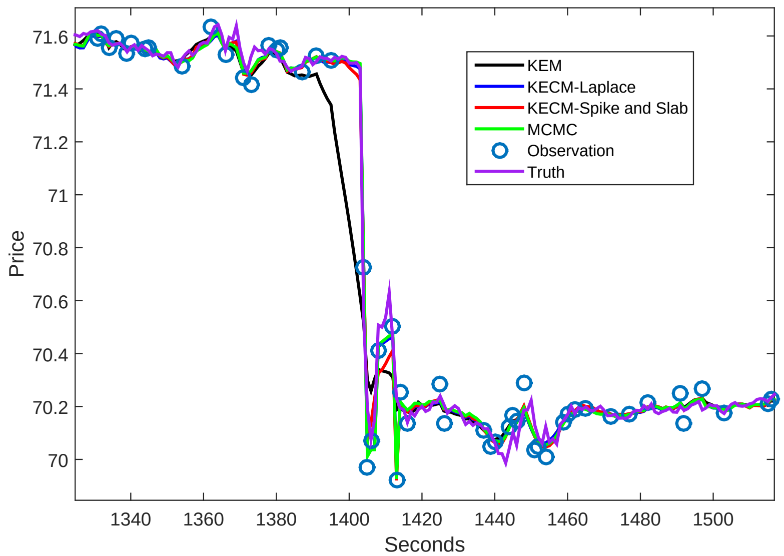

Figure 3.9: Price estimate example from the KEM, KECM, and Gibbs sampling. This is an example of the KEM algorithm over-smoothing near a small jump in price

The results for the GARCH(1,1)-jump model are shown in Tables 3.4 - 3.5. From these tables we see that the KECM and MCMC algorithms are robust to the volatility clustering exhibited in GARCH models.

3.4.5 Simulated Data from GARCH(1,1)-jump Model and stochastic microstructure variance

In this section we test our algorithms under a GARCH(1,1)-jump model with stochastic microstructure variance. This microstructure noise model accounts for a positive correlation between the bid-ask spread and the squared innovation. This models an empirical phenomena that has been observed in many markets [91]. Here we assume the same efficient price innovation as the GARCH(1,1)-jump model but now we allow for time-varying variance in the microstructure noise. In this model the variance of the microstructure noise at time t

ζ	σ_j^2	KEM	KECM Laplace	KECM Spike & Slab	MCMC	Pairwise Refresh	Pairwise Refresh (jump)
1	N/A	1.3e-10	1.3e-10	1.3e-10	1.4e-10	2.5e-10	4e-10
0.9999	6.25e-06	1.6e-10	1.6e-10	1.5e-10	1.6e-10	2.4e-10	3.1e-10
0.9999	0.0001	1.6e-10	1.3e-10	1.3e-10	1.3e-10	4.4e-10	4.2e-10
0.9995	6.25e-06	2e-10	1.5e-10	1.5e-10	1.5e-10	4.4e-10	4.6e-10
0.9995	0.0001	3.7e-10	1.3e-10	1.4e-10	1.3e-10	1e-09	3.9e-10
0.999	6.25e-06	2.6e-10	1.4e-10	1.4e-10	1.4e-10	5.8e-10	4.5e-10
0.999	2.5e-05	5.5e-10	1.5e-10	1.7e-10	1.6e-10	1.4e-09	7e-10
0.999	0.0001	1.1e-09	1.6e-10	1.5e-10	1.5e-10	2e-09	6.8e-10

Table 3.4: Portfolio variance for GARCH(1,1)-jump model, best performance highlighted in green.

ζ	σ_j^2	KEM	KECM Laplace	KECM Spike & Slab	MCMC	Pairwise Refresh	Pairwise Refresh (jump)
1	N/A	0.37	0.37	0.38	0.41	0.5	0.52
0.9999	6.25e-06	0.43	0.37	0.38	0.41	0.58	0.55
0.9999	0.0001	3.3	0.39	0.4	0.43	1.7	0.55
0.9995	6.25e-06	0.88	0.42	0.43	0.44	0.81	0.62
0.9995	0.0001	18	0.65	0.49	0.46	8.5	0.61
0.999	6.25e-06	1.4	0.48	0.51	0.49	1.2	0.64
0.999	2.5e-05	7.7	0.64	0.62	0.52	4.5	0.69
0.999	0.0001	36	1.4	0.67	0.55	16	0.71

Table 3.5: Average covariance error for GARCH(1,1)-jump model, best performance highlighted in green. Large errors highlighted in red.

ζ	σ_j^2	KEM	KECM Laplace	KECM Spike & Slab	MCMC	Pairwise Refresh	Pairwise Refresh (jump)
1	N/A	1.5e-10	1.6e-10	1.6e-10	1.6e-10	2.2e-10	4.7e-10
0.9999	6.25e-06	1.6e-10	1.6e-10	1.6e-10	1.6e-10	3e-10	2.8e-10
0.9999	0.0001	2e-10	1.6e-10	1.6e-10	1.6e-10	5.1e-10	4.6e-10
0.9995	6.25e-06	2.6e-10	1.9e-10	1.9e-10	1.9e-10	4.6e-10	5.1e-10
0.9995	0.0001	5.1e-10	1.8e-10	1.8e-10	1.7e-10	1.5e-09	7.4e-10
0.999	6.25e-06	2.3e-10	1.5e-10	1.5e-10	1.5e-10	5e-10	5.4e-10
0.999	2.5e-05	5.6e-10	1.7e-10	1.7e-10	1.7e-10	1.4e-09	1e-09
0.999	0.0001	9e-10	2e-10	1.6e-10	1.5e-10	2.3e-09	7.4e-10

Table 3.6: Portfolio variance for GARCH(1,1)-jump model with stochastic microstructure noise variance, best performance highlighted in green

for i^{th} asset is

$$\left(0.1\frac{(X_i(t) - X_i(t-1) - D)^2}{\Gamma_{i,i}} + 0.9\right)\tilde{\sigma}^2_{o,i}$$

which is the sum of fixed variance and time varying term which is dependent on the efficient price innovation. Here we see that when the squared innovation equals the variance then the observation noise variance equals $\tilde{\sigma}^2_{o,i}$. As in the previous simulations, $\tilde{\sigma}^2_{o,i}$ is chosen to be a realization of a gamma distributed random variable with shape 2 and mean 0.0002^2.

The results for this model are shown in Tables 3.6 and 3.7. A comparison of the covariance errors is shown in Table 3.8. From the comparison table we see that the covariance errors are larger for the non-stationary microstructure noise model. Here the KECM-Laplace model is especially sensitive to the stochastic microstructure noise variance for $\sigma_j^2 = 1e-4$. In some cases the covariance error increased by about a factor of 10. The KECM-spike and slab and MCMC approaches were not as sensitive to the stochastic noise variance.

ζ	σ_j^2	KEM	KECM Laplace	KECM Spike & Slab	MCMC	Pairwise Refresh	Pairwise Refresh (jump)
1	N/A	0.37	0.37	0.38	0.42	0.57	0.6
0.9999	6.25e-06	0.51	0.42	0.42	0.47	0.55	0.56
0.9999	0.0001	21	1.5	0.38	0.41	2.6	0.56
0.9995	6.25e-06	0.78	0.4	0.41	0.44	0.82	0.55
0.9995	0.0001	75	3.3	0.47	0.44	9.5	0.67
0.999	6.25e-06	1.2	0.41	0.44	0.43	1	0.6
0.999	2.5e-05	13	0.48	0.51	0.46	3.5	0.79
0.999	0.0001	1.3e+02	13	2.7	0.45	13	2.2

Table 3.7: Average covariance error for GARCH(1,1)-jump model with stochastic microstructure noise variance, best performance highlighted in green. Large errors highlighted in red.

ζ	σ_j^2	KECM Laplace Model 1	KECM Laplace Model 2	KECM Spike &Slab Model 1	KECM Spike &Slab Model 2	MCMC Model 1	MCMC Model 2
1	N/A	0.37	0.37	0.38	0.38	0.41	0.42
0.9999	6.25e-06	0.37	1.5	0.38	0.38	0.41	0.41
0.9999	0.0001	0.39	0.42	0.4	0.42	0.43	0.47
0.9995	6.25e-06	0.42	3.3	0.43	0.47	0.44	0.44
0.9995	0.0001	0.65	0.4	0.49	0.41	0.46	0.44
0.999	6.25e-06	0.48	13	0.51	2.7	0.49	0.45
0.999	2.5e-05	0.64	0.48	0.62	0.51	0.52	0.46
0.999	0.0001	1.4	0.41	0.67	0.44	0.55	0.43

Table 3.8: Average covariance error comparison for GARCH-jump model (model 1) and GARCH-jump model with stochastic microstructure noise variance (model 2)

ζ	σ_j^2	KEM (sec)	KECM Laplace (sec)	KECM Spike Slab(sec)	MCMC (sec)	Pairwise Refresh (sec)	Pairwise Refresh jump (sec)
1	N/A	24.9	77.6	57.5	182.5	0.7	7.1
0.9999	6.25e-06	28.7	76.4	58.0	182.4	0.7	7.1
0.9999	0.0001	48.9	83.5	59.8	184.0	0.7	7.0
0.9995	6.25e-06	46.2	88.5	61.5	185.1	0.7	7.1
0.9995	0.0001	95.1	109.9	64.4	193.8	0.7	7.1
0.999	6.25e-06	51.9	83.1	62.0	187.4	0.8	7.0
0.999	2.5e-05	86.9	99.9	66.1	177.5	0.8	7.1
0.999	0.0001	90.0	122.9	72.8	200.9	0.8	7.1

Table 3.9: Run-time (seconds) for GARCH-jump model with stochastic microstructure noise variance

3.4.6 Timing

Average MATLAB timing of the algorithms for the GARCH(1,1)-jump model with stochastic microstructure noise variance are shown in Table 3.9. The machine running the simulation has the Windows 7 operating system and an Intel i7-3740 processor with 32.0 GB of RAM. The table shows that the pairwise refresh methods are the least computationally costly, while the MCMC method requires the most run time. The data also indicates that the KEM, KECM, and MCMC algorithms take longer to converge when larger and more frequent jumps are present.

3.4.7 Numerical Results Summary

The following are key observations from the numerical simulation results:

1. Both KECM and MCMC approaches outperform KEM in the presence of jumps.

2. Laplace prior underperforms spike and slab models for large jumps.

3. Spike and slab models are more robust to stochastic microstructure noise variance than

the Laplace prior model.

The first observation is not surprising since both the KECM and MCMC approaches explicitly account for jumps. The second and third observations may be the result of a large jump estimation bias that can occur when using the Laplace prior for large σ_j^2.

3.5 Extensions to Matrix Factorization Approaches

In this section we examine the use of a robust nuclear norm formulation to the recover the missing unobserved true prices. This technique is inspired by robust PCA [15] and matrix completion [16]. In the application to covariance estimation the recovered missing prices could be used as an input to a covariance estimation algorithm.

Intuitively in the presence of missing data the ability of the KECM algorithms to recover the true unobserved price depends on the approximate low rank structure of the covariance matrix. If the covariance is low rank then missing observations can be inferred from changes in the prices of observed assets. However if assets returns are nearly uncorrelated (i.e. high rank covariance) then returns on one asset provide little information on the returns of other assets. An illustration of this is shown in Figures 3.10 and 3.11. Here we see that for a low rank covariance we are able to accurately recover missing data whereas in a high rank covariance the missing data cannot be recovered accurately.

Mathematically we can observe this behavior by considering a toy two asset problem with zero-mean normal log-returns with covariance

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Let $x_1(t)$ and $x_2(t)$ denote the log returns. Suppose the price at time 0 is observed for both assets, but only the price of stock 2 is observed at time 1. Then the posterior distribution of the price of asset 1 at time 1 is normal with mean

$$x_1(0) + \rho(x_2(1) - x_2(0))$$

and variance

$$1 - \rho^2.$$

Here we see that for approximately rank deficient covariance matrix (ρ close to 1) we have a posterior variance of nearly 0. Thus we can accurately recover the missing observation $x_1(1)$ by setting its value to the posterior mean. If the covariance is well conditioned (ρ close to 0), then the posterior variance is close to 1 which means the posterior mean may not be a reliable estimate.

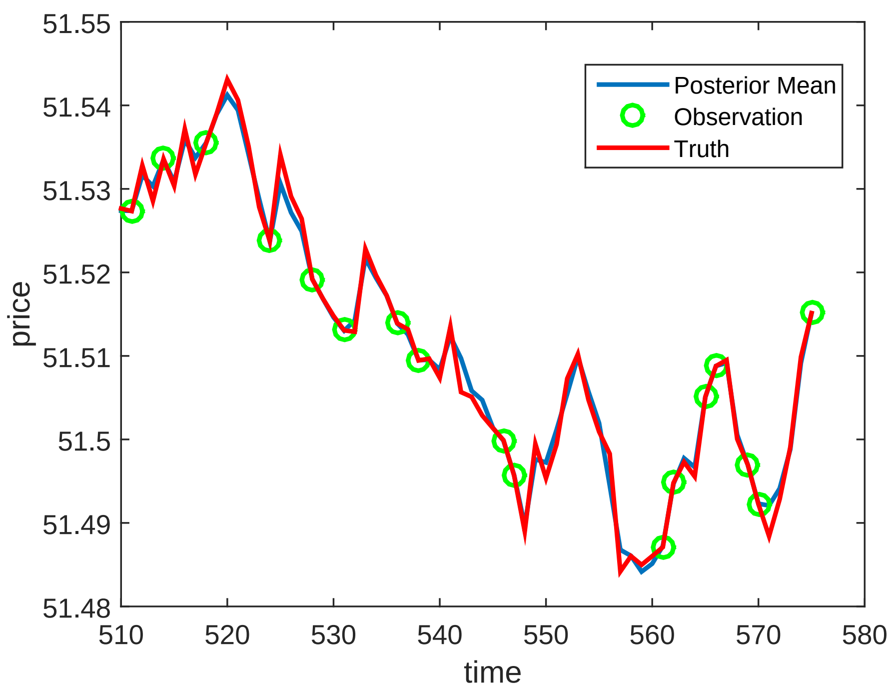

Figure 3.10: Missing data is accurately recovered using KECM when covariance matrix is low rank. Here the percentage of missing observation is 70% and asset returns from 20 stocks is generated from a 5 factor model with small residual term.

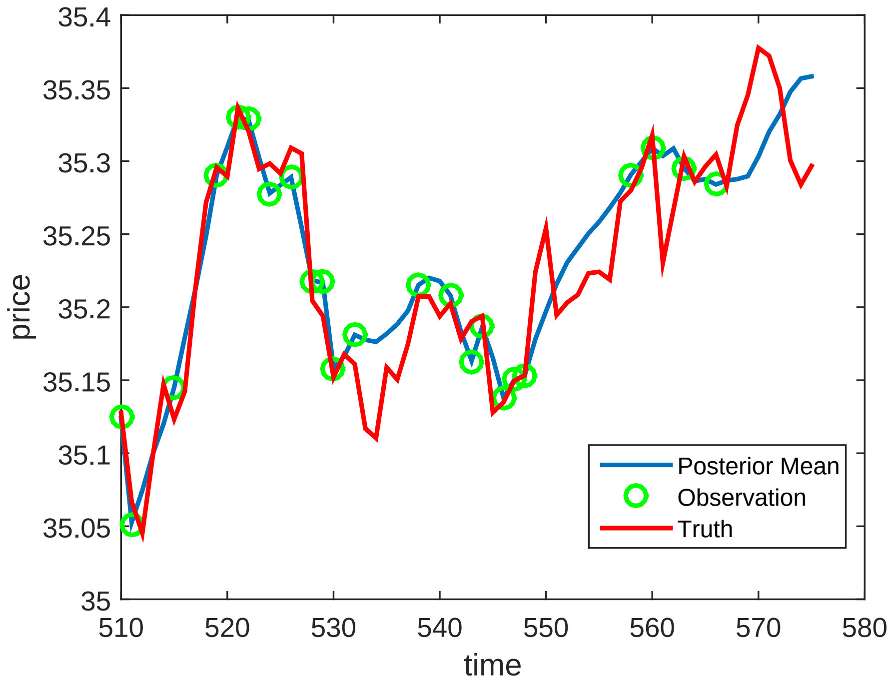

Figure 3.11: Missing data is not accurately recovered using KECM when covariance matrix is high rank. Here the percentage of missing observation is 70% and asset returns from 20 stocks are generated from a 5 factor model with large residual term.

Since a large number of financial asset returns have a low rank structure it is natural to exploit this structure directly in order to recover the missing underling true price. In this section we show how nuclear norm minimization can be used for this. First we define the following notation:

- $R_{i,t}$:unobserved low rank diffusion component of log-return for asset i at time $t > 1$. For $t = 1$ let $R_{i,t}$ be the unobserved true log asset price at time $t = 1$.
- $J_{i,t}$:unobserved sparse jump component of log return for asset i at time $t > 1$. Let $J_{i,1} = 0$.
- Y_{i_k,t_k} : observed (noisy) log price of asset i_k at time t_k.
- S: discrete time integration (in time) operator (rectangular method).

Here S has a upper triangular matrix representation

$$S = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 0 & 1 & 1 & \dots & 1 \\ \vdots & 0 & 1 & \dots & 1 \\ & \vdots & 0 & \dots & 1 \\ & & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Since the data returns consist of a low rank diffusion and sparse jump component one criterion for recovery of the true price $X(t)$ is

$$\text{argmin}_{j,r,x} \quad rank(r) + \lambda_1 \sum_k (((r+j)S)_{i_k,t_k} - Y_{i_k,t_k})^2 + \lambda_2 \|j\|_{\ell_0} \text{ s.t. } j_{i,1} = 0 \quad \forall i; \quad x = (r+j)S$$

for some $\lambda_1, \lambda_2 > 0$. This criterion promotes low rank + sparse representations of the return such that they are consistent with the observed noisy prices. Unfortunately solving this problem is non-convex and NP-hard due to the rank and ℓ_0 component of the objective.

One alternative to the non-convex criterion is to use the following convex relaxation

$$\text{argmin}_{j,r,x} \quad \|r\|_* + \lambda_1 \sum_k (((r+j)S)_{i_k,t_k} - Y_{i_k,t_k})^2 + \lambda_2 \|j\|_{\ell_1} \text{ s.t. } j_{i,1} = 0 \quad \forall i; \quad x = (r+j)S \tag{3.27}$$

where $\|r\|_*$ represents the nuclear norm which is the sum of the singular values of r. The nuclear norm term promotes a low rank estimate of R while the ℓ_1 term promotes a sparse estimate of the jumps [15].

The criterion in (3.27) can be solved using accelerated proximal gradient (APG) descent methods [58] with convergence rate of $O(1/k^2)$. For strongly convex objective functions the APG algorithm converges at a linear rate [20, 74]. Unfortunately the objective in (3.27) is not strongly convex. However assuming the noisy price at time 0 is observed for each asset, convergence can be accelerated to a linear rate by considering the following strongly convex variant of (3.27)

$$\begin{aligned} \operatorname{argmin}_{x,j,r} \quad & \|r\|_* + \lambda_1 \sum_k \left(((r+j)S)_{i_k,t_k} - Y_{i_k,t_k}\right)^2 + \lambda_2 \|j\|_{\ell_1} + \\ & + \epsilon_1 \sum_{t>1} \sum_m r_{m,t}^2 + \epsilon_2 \sum_{t>1} \sum_m j_{m,t}^2 \quad \text{s.t. } j_{i,1} = 0 \quad \forall i; \;\; x = (r+j)S \end{aligned} \tag{3.28}$$

where ϵ_1 and ϵ_2 are positive valued. Here the addition of the ℓ_2^2 penalty can be interpreted as imposing a zero-mean normal prior on the distribution of the returns and jumps. An application of the APG algorithm to solving (3.28) is given in Algorithm 4.

Algorithm 4 APG algorithm solving (3.28)

Initialize: $R^{(0)}, J^{(0)}, k = 0$

Define

$$f(r,j) = \lambda_1 \sum_k \left(((r+j)S)_{i_k,t_k} - Y_{i_k,t_k}\right)^2 + \epsilon_1 \sum_{t>1}\sum_m r_{m,t}^2 + \epsilon_2 \sum_{t>1}\sum_m j_{m,t}^2$$

Compute maximum eigenvalue of Hessian of $f(r,j)$. Denote as λ_{max}.

Compute minimum eigenvalue of Hessian of $f(r,j)$. Denote as λ_{min}.

Set $s = \frac{\alpha}{\lambda_{max}}$ for some $\alpha < 1$

Set $\tau = \frac{\sqrt{\lambda_{max}} - \sqrt{\lambda_{min}}}{\sqrt{\lambda_{max}} + \sqrt{\lambda_{min}}}$

Set $\tilde{R}^{(0)} = R^{(0)}$, $\tilde{J}^{(0)} = J^{(0)}$

Set $k = 0$

while not converged **do**

  Compute $\nabla_r f(\tilde{R}^{(k)}, \tilde{J}^{(k)})$,$\nabla_j f(\tilde{R}^{(k)}, \tilde{J}^{(k)})$

  Set $R^{(k+1)} = shrinkNuc(\tilde{R}^{(k)} - s\nabla_r f, s)$

  Set $J^{(k+1)} = shrink_{\ell_1}(\tilde{J}^{(k)} - s\nabla_j f, s\lambda_2)$

  Set $\tilde{R}^{(k+1)} = R^{(k+1)} + \tau(R^{(k+1)} - R^{(k)}$

  Set $\tilde{J}^{(k+1)} = J^{(k+1)} + \tau(J^{(k+1)} - J^{(k)}$

  Set $k = k + 1$

end while

Here the nuclear norm shrinkage is given as

$$shrinkNuc(R,s) = U\text{diag}((\sigma_1 - s) \vee 0, \ldots, (\sigma_M - s) \vee 0)V^T$$

where R has singular value decomposition $U\text{diag}(\sigma_1, \ldots, \sigma_M)V^T$. The ℓ_1 norm shrinkage is

$$(shrink_{\ell_1}(J,s))_{m,n} = \begin{cases} J_{m,n} - s \text{ if } J_{m,n} > s \\ 0 \text{ else} \\ J_{m,n} + s \text{ if } -J_{m,n} > s. \end{cases}$$

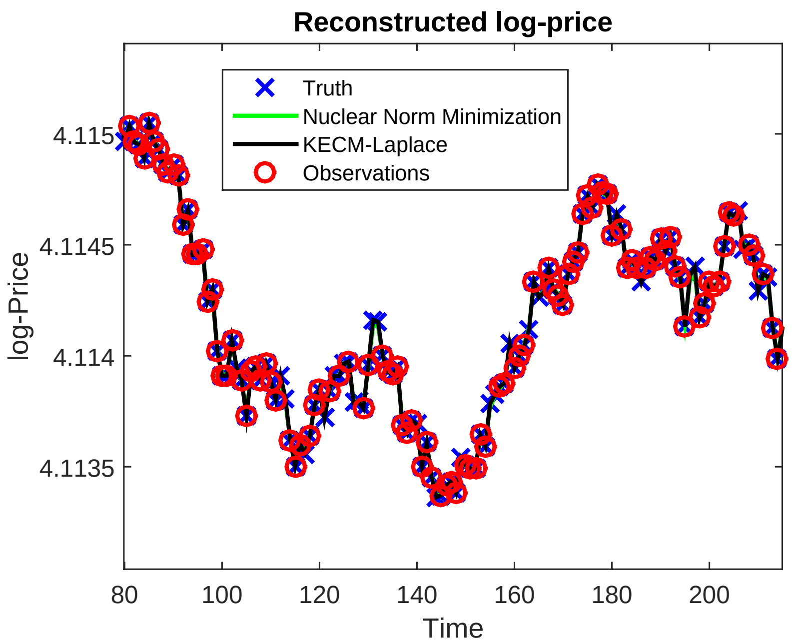

Figure 3.12: Here percentage of observed prices is 80% and there is no observation noise. The recovery by nuclear norm minimization is nearly identical to KECM-Laplace. Parameter settings in (3.28) are $\lambda_1 = 2000, \lambda_2 = 0.8, \epsilon_1 = 88, \epsilon_2 = 0.88$. 20 assets, 250 time samples. The squared ℓ_2 norm of the reconstruction error is 2.9e-7 for KECM-Laplace and 2.5e-6 for nuclear norm minimization.

3.5.1 Nuclear Norm Simulation Examples

In Figures 3.12 to 3.19 we show some example plots of the nuclear norm reconstruction criterion in (3.28) for various levels of missing data and noise. In Figures 3.12 and 3.13 we see that for a large percentage of observations and without noise the criterion in (3.28)

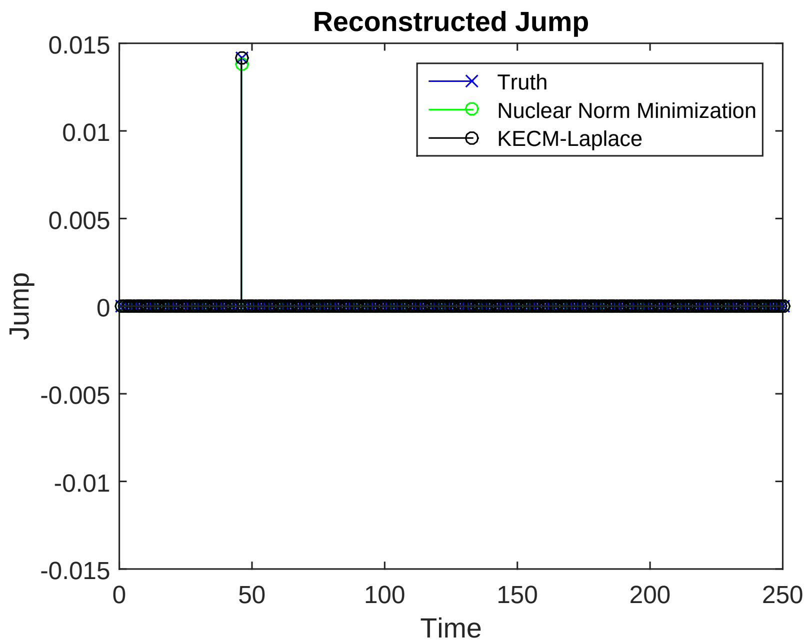

Figure 3.13: Here percentage of observed prices is 80% and there is no observation noise. The jump recovery by nuclear norm minimization is slightly biased towards 0. Parameter settings in (3.28) are $\lambda_1 = 2000, \lambda_2 = 0.8, \epsilon_1 = 88, \epsilon_2 = 0.88$. 20 assets, 250 time samples.

is comparable with KECM-Laplace. However as shown in Figures 3.14 to 3.19 the nuclear norm minimization technique performs poorly when a larger percentage of data is missing or if there is large observation noise.

Table 3.10 shows covariance matrix estimation accuracy for 20 assets and 300 time samples under a GARCH(1,1)-jump model. Here the observation noise variance is 0.0001^2. All other data parameters are identical to section 3.4. To estimate the covariance for the nuclear norm minimization approach we use sample averaging of the recovered prices using the criterion in (3.28). The results indicate that the covariance matrix estimates obtained using the recovered prices from matrix factorization are less accurate than those obtained with the KECM and MCMC approaches.

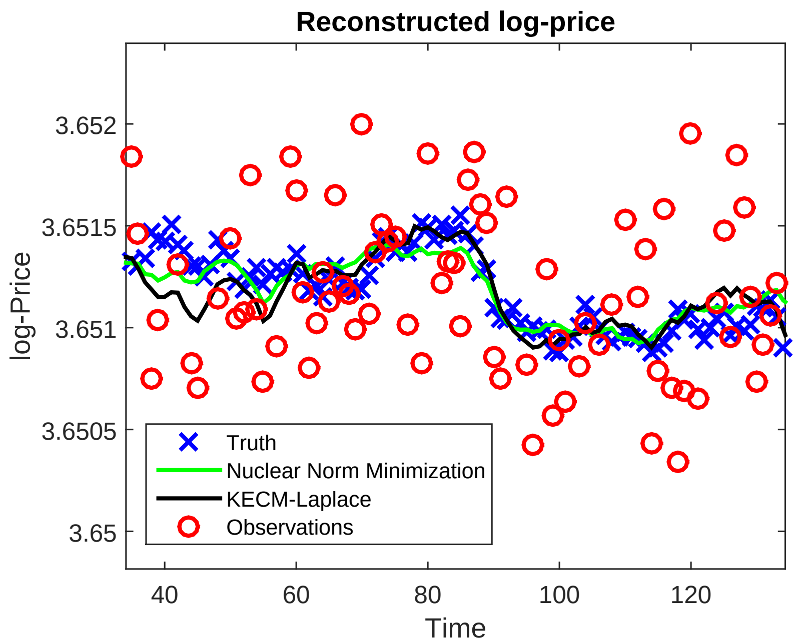

Figure 3.14: Here percentage of observed prices is 80% and there is observation noise with variance 0.0004^2. Parameter settings in (3.28) are $\lambda_1 = 20, \lambda_2 = 0.2, \epsilon_1 = 88, \epsilon_2 = 0.88$. 20 assets, 250 time samples. The squared ℓ_2 norm of the reconstruction error is 1.1e-4 for KECM-Laplace and 2.1e-4 for nuclear norm minimization.

3.5.2 Timing

Timing results for the examples in section 3.5.1 are shown in Table 3.11. The computer and software is identical to the description in Section 3.4.6.

Here we see that nuclear norm minimization requires significantly more computation time then KECM. The primary contributor to the long computation time is computing the singular value decomposition which accounted for 80% of the computation time in our experiments. When examining the no noise case we see that the nuclear norm minimization takes longer to run. For those cases the λ_1 penalty value was increased to enforce a better fit to the observed data. This increases the condition number of the objective function which in turn increases the number of iterations needed for convergence.

ζ	σ_j^2	KEM	KECM Laplace	KECM Spike & Slab	MCMC	Pairwise Refresh	Pairwise Refresh (jump)	Nuclear Norm Min.
1	N/A	0.45	0.44	0.45	0.56	0.55	0.57	0.91
0.9999	0.0001	2.4	0.43	0.43	0.56	2.4	2.1	0.91
0.9999	6.25e-06	0.63	0.39	0.39	0.48	0.66	0.66	0.91
0.9995	0.0001	12	0.66	0.46	0.6	13	11	0.95
0.9995	6.25e-06	0.79	0.38	0.39	0.45	0.82	0.78	0.91
0.999	0.0001	24	0.99	2.7	0.58	15	11	0.94
0.999	2.5e-05	6	0.6	0.73	0.53	7.5	4.4	0.93
0.999	6.25e-06	3.1	0.46	0.64	0.65	1.4	1.2	0.91

Table 3.10: Average covariance error for GARCH(1,1)-jump model, best performance highlighted in green,$\sigma_o^2 = 0.0001^2$. 300 time samples

Percentage Missing	Noise	Time KECM-Laplace (sec)	Time Nuclear Norm Min (sec) APG descent
20 %	Yes	8.5	15.1
20 %	No	8.7	148
70 %	Yes	8.2	16.2
70 %	No	8.4	150

Table 3.11: Run time (seconds) for nuclear norm model examples in Section 3.5.1

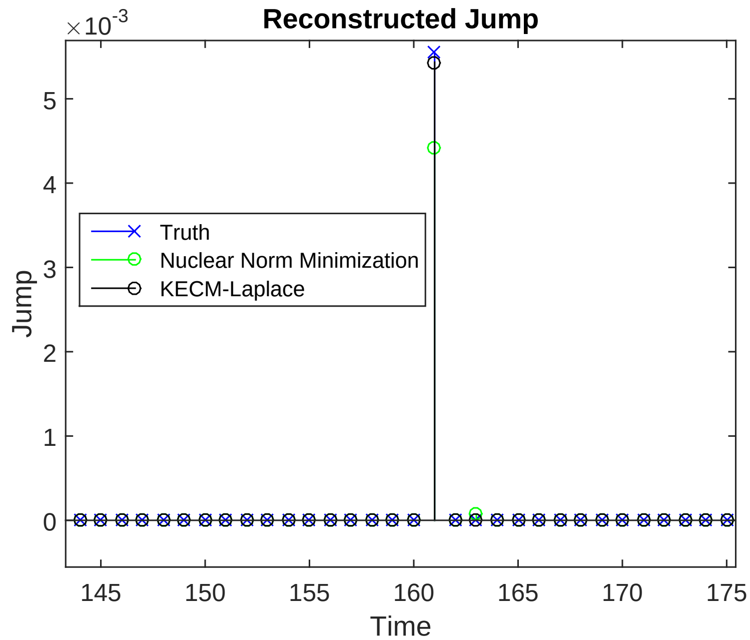

Figure 3.15: Here percentage of observed prices is 80% and there is observation noise with variance 0.0004^2. Parameter settings in (3.28) are $\lambda_1 = 20, \lambda_2 = 0.2, \epsilon_1 = 88, \epsilon_2 = 0.88$. 20 assets, 250 time samples

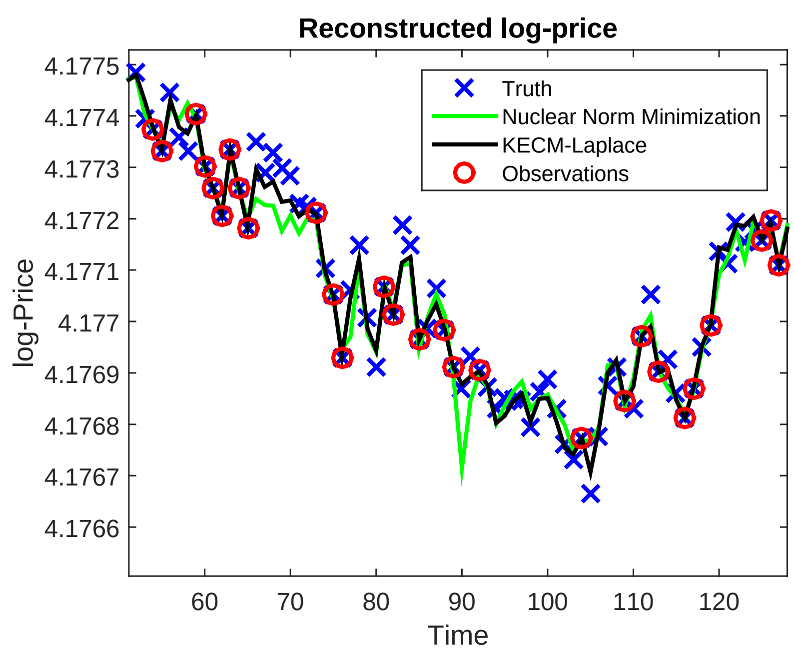

Figure 3.16: Here percentage of observed prices is 30% and there is no observation noise. Parameter settings in (3.28) are $\lambda_1 = 2000, \lambda_2 = 0.8, \epsilon_1 = 37, \epsilon_2 = 0.37$. 20 assets, 250 time samples. The squared ℓ_2 norm of the reconstruction error is 1.7e-5 for KECM-Laplace and 4.5e-5 for nuclear norm minimization.

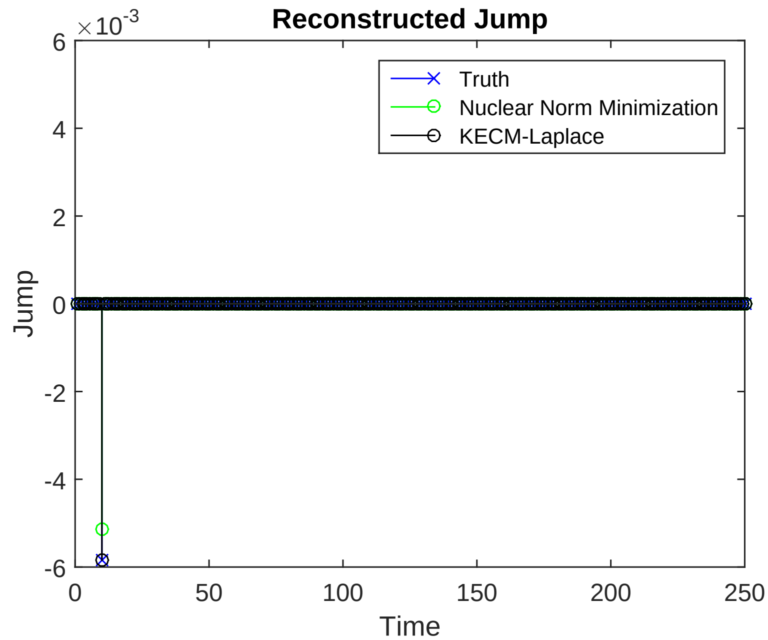

Figure 3.17: Here percentage of observed prices is 30% and there is no observation noise. Parameter settings in (3.28) are $\lambda_1 = 2000, \lambda_2 = 0.8, \epsilon_1 = 37, \epsilon_2 = 0.37$. 20 assets, 250 time samples

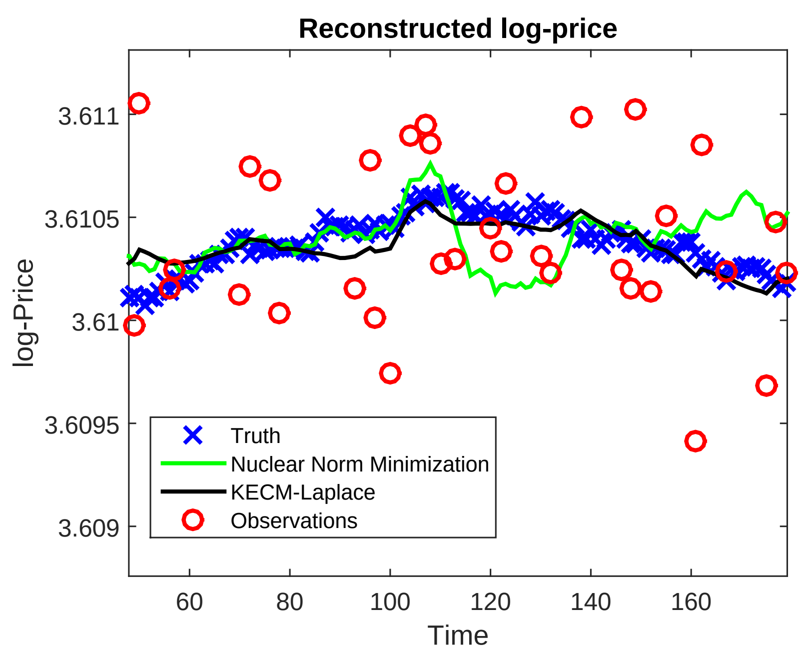

Figure 3.18: Here percentage of observed prices is 30% and there is observation noise with variance 0.0004^2. Parameter settings in (3.28) are $\lambda_1 = 20, \lambda_2 = 0.2, \epsilon_1 = 37, \epsilon_2 = 0.37$. 20 assets, 250 time samples. The squared ℓ_2 norm of the reconstruction error is 2.2e-4 for KECM-Laplace and 7.2e-4 for nuclear norm minimization.

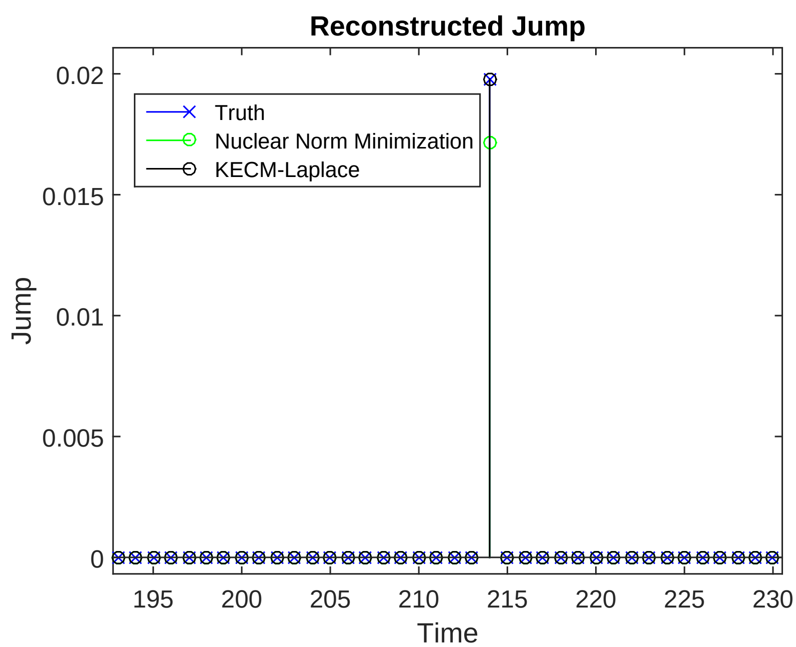

Figure 3.19: Here percentage of observed prices is 30% and there is observation noise with variance 0.0004^2. Parameter settings in (3.28) are $\lambda_1 = 20, \lambda_2 = 0.2, \epsilon_1 = 37, \epsilon_2 = 0.37$. 20 assets, 250 time samples

Chapter 4

Weighted Elastic Net Penalized Portfolios

In this chapter we present a new penalization approach to improve the robustness of mean-variance portfolio optimization. We also present efficient numerical techniques to solve the new penalized problem.

Portions of this chapter reuse material from the article [46](© 2015 Society for Industrial and Applied Mathematics. Reprinted with permission. All rights reserved).

4.1 Weighted Elastic Net (WEN)

The norm penalties presented in Chapter 2 are derived and calibrated primarily from a minimum variance perspective. In this chapter we extend the above methods for minimum variance portfolio design to mean-variance portfolios. Here we propose augmenting the mean-variance criterion with two penalty terms. The first term is a weighted ℓ_1 norm of the portfolio weights and the second term is the square of a weighted ℓ_2 norm of the portfolio weights. The sum of these two penalty terms will be referred to as a weighted elastic net (WEN) which was studied in the context of variable selection in [94].

Let $\{\alpha_i\}_{i=1}^N$ and $\{\beta_i\}_{i=1}^N$ be positive real numbers. Then the WEN penalty for portfolio weights w is given by

$$||w||_{\vec{\beta},\ell_1} + ||w||^2_{\vec{\alpha},\ell_2}$$

where

$$||w||_{\vec{\beta},\ell_1} = \sum_{k=1}^N \beta_k |w_k|$$

and

$$||w||^2_{\vec{\alpha},\ell_2} = \sum_{k=1}^N \alpha_k |w_k|^2.$$

With these definitions the WEN penalized objective function is

$$\begin{aligned} \Psi(w) &= w^T\hat{\Gamma}w - w^T\hat{\mu} + ||w||_{\vec{\beta},\ell_1} + ||w||^2_{\vec{\alpha},\ell_2} \qquad (4.1) \\ &= w^T Rw - w^T\hat{\mu} + ||w||_{\vec{\beta},\ell_1} \end{aligned}$$

where $\hat{\Gamma}$ and $\hat{\mu}$ are estimates of Γ and μ respectively. Here $R = \hat{\Gamma} + D_\alpha$ where

$$D_\alpha \doteq \begin{pmatrix} \alpha_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \alpha_N \end{pmatrix}. \qquad (4.2)$$

Thus the WEN penalized criterion is

$$\min_w \quad \Psi(w) \qquad (4.3)$$

Since R is positive semidefinite the objective function Ψ is strictly convex and the solution to (4.3) is unique.

4.1.1 Robustness of the WEN penalized portfolio

Motivation for augmenting the mean-variance criterion with a WEN penalty can be obtained by considering a robust portfolio optimization problem. Recall from Chapter 1 that the out-of-sample performance of the mean-variance portfolio can degrade significantly when there are errors in the estimates of mean and covariance. This parameter estimation risk can be reduced by accounting for it in the optimization criterion.

One way to approach parameter estimation risk is to assume that the true covariance and mean belong to uncertainty sets such as the ones below

$$A = \left\{R : R_{i,j} = \hat{\Gamma}_{i,j} + e_{i,j}; |e_{i,j}| \leq \Delta_{i,j}; R \geq 0\right\}$$

$$B = \{v : v_i = \hat{\mu}_i + c_i; |c_i| \leq \beta_i\}.$$

Here the matrix Δ is symmetric and diagonally dominant with $\Delta_{i,j} \geq 0$ for all i, j which ensures that a matrix, R, of the form

$$\begin{aligned} R_{i,j} &= \begin{cases} \hat{\Gamma}_{i,i} + \Delta_{i,i} \text{ if } i = j \\ \hat{\Gamma}_{i,j} \pm \Delta_{i,j} \text{ if } i \neq j \end{cases} \\ R_{i,j} &= R_{j,i} \end{aligned}$$

is positive semi-definite (i.e. $R \in A$).

Robustness to parameter uncertainty can be gained by selecting a portfolio to optimize the worse case performance over the above uncertainty sets. This can be written as a robust optimization problem [42]

$$\min_{w} \max_{R \in A, v \in B} w^T R w - v^T w. \tag{4.4}$$

The main result in this section is given below.

Theorem 4.1. *The WEN penalized problem in* (4.3) *is equivalent to the robust optimization problem in* (4.4), *when*

$$\Delta = D_\alpha \tag{4.5}$$

Proof. Note that for a fixed R and v this problem is convex in w. Since the pointwise maximum of convex functions remains convex we have that

$$\max_{R \in A, v \in B} w^T R w - v^T w \tag{4.6}$$

is convex in w. Performing the inner maximization with respect to v reduces the problem to

$$\min_{w} \max_{R \in A} \quad w^T R w + \sum_{i=1}^{N} \left(-\hat{\mu}_i + \beta_i \mathrm{sgn}(w_i) \right) w_i \tag{4.7}$$

where

$$\mathrm{sgn}(w_i) = \begin{cases} \frac{w_i}{|w_i|} & \text{if } w_i \neq 0 \\ 0 & \text{else.} \end{cases}$$

This can be re-written as

$$\min_{w} \max_{R \in A} \ \mathrm{tr}(R w w^T) - w^T \hat{\mu} + \|w\|_{\vec{\beta}, \ell_1},$$

and the inner maximization with respect to R can be solved in closed form to give

$$\min_w w^T \hat{\Gamma} w - w^T \hat{\mu} + |w|^T \Delta |w| + \|w\|_{\vec{\beta},\ell_1} \tag{4.8}$$

where the vector $|w|$ is defined as

$$|w|_i = |w_i|. \tag{4.9}$$

This is a pairwise weighted elastic net (PWEN) [63] penalized criterion. When Δ equals the diagonal matrix D_α the criterion simplifies to the WEN penalized problem defined in problem (4.3). □

4.1.2 Bootstrap Calibration

In this section we address the problem of calibrating the weighting parameters α and β. A guideline for choosing these parameters is implied by Theorem 4.1 which states that problems (4.3) and (4.4) are equivalent. This implies that α and β represent the level of uncertainty in the mean and variance of each asset. Thus a natural guideline for selecting α and β is that they should be proportional to the amount of uncertainty in the parameter estimates.

Since the amount of error in the parameter estimates is unknown, we will need to estimate the errors prior to choosing α and β. One approach to estimate the amount of error is the bootstrap method [30]. Bootstrapping is a non-parametric approach that has been applied to portfolio optimization [71] and calibration of robust portfolio optimization problems [83]. One advantage of bootstrapping is that it does not require specification of a distribution of the return data or normality assumptions on the parameter estimation error.

Bootstrapping can be used to calibrate the penalty factors as follows.

1. For $k = 1$ to K

   - Resample T_{train} samples from the training data
   - Estimate mean and covariance using this sample, denote as $\tilde{\mu}$ and $\tilde{\Gamma}$
   - Compute absolute error between the resampled estimates and $\hat{\mu}$ and $\hat{\Gamma}$

$$\mu_{i,err}(k) = |\tilde{\mu}_i - \hat{\mu}_i|$$

$$\Gamma_{i,err}(k) = \left|\tilde{\Gamma}_{i,i} - \hat{\Gamma}_{i,i}\right|$$

2. Compute empirical distributions, f_i, of $\mu_{i,err}$
3. Compute empirical distributions, g_i, of $\Gamma_{i,err}$
4. Set α_i to p_1 percentile of f_i and β_i to p_2 percentile of g_i

An economic interpretation of the percentile parameters p_1 and p_2 is that of model estimation risk aversion factors. Here p_1 represents the aversion to squared volatility estimation risk and p_2 is the aversion to mean estimation risk. A percentile value of 0 corresponds to no aversion to estimation risk whereas a value of 1 corresponds to a high aversion to estimation risk. Note that a higher aversion to estimation risk will increase the weights in the elastic net.

4.2 Computational Aspects

In this section we present an adaptive support split-Bregman approach for solving (4.3). This technique improves upon the speed of the split-Bregman [43] algorithm by exploiting the sparse nature of WEN penalized portfolio weights.

4.2.1 Preliminaries

First we derive approximate optimality conditions for the WEN penalized criterion (4.3). Recall that the objective function in the WEN criterion, $\Psi(w)$, is convex. Thus w^* minimizes Ψ if and only if

$$0 \in \partial\Psi(w^*) \tag{4.10}$$

where $\partial\Psi(w)$ is the sub-gradient of Ψ evaluated at w [12]. Since Ψ is strictly convex there is a unique solution to (4.3).

In most cases we are only interested in portfolios that are approximately optimal. Thus we can relax the above optimality conditions to derive a stopping criterion. Before introducing our relaxed conditions we define the support of a portfolio w as

$$\operatorname{supp}(w) = \{i : |w_i| > 0\}$$

and define the smallest variance uncertainty as

$$\alpha_o = \min\{\alpha_i : 0 \le i \le N\}. \tag{4.11}$$

With the above definitions we have the following theorem which establishes an approximate optimality condition.

Theorem 4.2. *Let w^* be the solution of* (4.3)*. Suppose that $\tilde{w}$ satisfies*

$$\sum_{i \in \operatorname{supp}(\tilde{w})} \left(\frac{\partial}{\partial w_i} \left(w^T R w - w^T \hat{\mu} + \|w\|_{\vec{\beta},\ell_1} \right) \Big|_{w=\tilde{w}} \right)^2 \le 2\epsilon\alpha_o \tag{4.12}$$

and

$$-\beta_i \le \frac{\partial}{\partial w_i} \left(w^T R w - w^T \hat{\mu} \right) \Big|_{w=\tilde{w}} \le \beta_i \tag{4.13}$$

for all $i \notin \operatorname{supp}(\tilde{w})$. *Then*

$$\Psi(\tilde{w}) \le \Psi(w^*) + \epsilon \tag{4.14}$$

Proof. See Section 4.2.3. □

In a numerical algorithm it may happen that none of the portfolio weights are exactly 0, although they may be extremely close to zero. Thus the above theorem may not be practical for use as a stopping criterion. For this reason let us separate the small portfolio weights from the larger portfolio weights. To that end we define

$$\operatorname{supp}_\epsilon(w) = \{i \in \operatorname{supp}(w) : |w_i| < \epsilon\}.$$

With this definition we have the following corollary which suggests a more practical stopping rule than Theorem 4.2.

Theorem 4.3. *Let* $M \ge 2\|R\|_{\ell_2}$ *and let* $\epsilon > 0$ *be given. Choose* $\eta < \frac{\epsilon \wedge \sqrt{\epsilon \alpha_o}}{\sqrt{N}M}$. *Let* w^* *be the solution the of* (4.3). *Suppose that* $\tilde{w}$ *satisfies*

$$\sum_{i \in \operatorname{supp}(\tilde{w}) \setminus \operatorname{supp}_\eta(\tilde{w})} \left(\frac{\partial}{\partial w_i} \left(w^T R w - w^T \hat{\mu} + \|w\|_{\vec{\beta}, \ell_1} \right) \Big|_{w = \tilde{w}} \right)^2 \le 2\epsilon\alpha_o \tag{4.15}$$

and

$$-\beta_i + \epsilon \le \frac{\partial}{\partial w_i} \left(w^T R w - w^T \hat{\mu} \right) \Big|_{w = \tilde{w}} \le \beta_i - \epsilon \tag{4.16}$$

for $i \in \operatorname{supp}_\eta(\tilde{w}) \cup \overline{\operatorname{supp}(\tilde{w})}$. *Then*

$$\Psi(\zeta) \le \Psi(w^*) + \frac{(\sqrt{2}+1)^2}{2}\epsilon \tag{4.17}$$

where

$$\zeta_i = \begin{cases} 0 \ \ \textit{if } i \in \operatorname{supp}_\eta(\tilde{w}) \\ \tilde{w}_i \ \ \textit{else}. \end{cases}$$

Proof. See Section 4.2.3. □

4.2.2 Solution via adaptive support split-Bregman Algorithm

The weighted elastic net problem can be reformulated as a quadratic program and solved using general purpose solvers. However the reformulation involves adding an additional N primal variables as well as $2N$ dual variables. Thus this approach may not be applicable to large scale problems.

An algorithm better suited to handle problems like (4.3) is the split-Bregman algorithm. The split-Bregman algorithm was introduced in [43] for problems involving ℓ_1 regularization such as (4.3). When using the split-Bregman method to solve (4.3) we solve an equivalent problem

$$\begin{aligned} &\min_{w,d} \; w^T R w - w^T \hat{\mu} + \|d\|_{\ell_1} \\ &\text{s.t. } \; d = \psi(w) \end{aligned} \tag{4.18}$$

where $R = \rho\hat{\Gamma} + D_\alpha$ and where $\psi(w) = (\beta_1 w_1, \ldots, \beta_N w_N)$. The split-Bregman algorithm applied to (4.18) is

Algorithm 5 Split-Bregman algorithm for solving (4.18)

Initialize: $k = 1,\ b^k = 0, w^k = 0, d^k = 0$

while $\|w^k - w^{k-1}\|_{\ell_2} > tol$ **do**

$w^{k+1} = \arg\min_w w^T R w - w^T \hat{\mu} + \frac{\lambda}{2}\|d^k - \psi(w) - b^k\|_{\ell_2}^2$

$d^{k+1} = \arg\min_d \frac{\lambda}{2}\|d - \psi(w^{k+1}) - b^k\|_{\ell_2}^2 + \|d\|_{\ell_1}$

$b_i^{k+1} = b_i^k + \beta_i w_i^{k+1} - d_i^{k+1}$

$k = k + 1$

end while

Both inner optimization problems in Algorithm 5 have closed form solutions. The first problem is an unconstrained strictly convex quadratic program and the second problem can be solved using the shrinkage operator

$$d_j^{k+1} = shrink(\beta_j w_j^{k+1} + b_j^k, \frac{1}{\lambda})$$

where

$$shrink(x, \gamma) = \frac{x}{|x|} \cdot \max(|x| - \gamma, 0).$$

The stopping criterion in Algorithm 5 does not ensure that the objective value is within a desired tolerance. A modification to the algorithm can be made to ensure that this occurs. One such modification uses Theorem 4.3 to derive a stopping criterion.

Algorithm 6 Modified split-Bregman Algorithm for solving (4.18)

Initialize: $k = 0$, $b^k, w^k, d^k = |w^k|, tol > 0$

while w^k does not satisfy conditions of Theorem 4.3 for $\epsilon = \frac{2}{(\sqrt{2}+1)^2} tol$ and $\tilde{w} = w^k$ **do**

$w^{k+1} = \arg\min_w w^T R w - w^T \hat{\mu} + \frac{\lambda}{2} \|d^k - \psi(w) - b^k\|_{\ell_2}^2$

$d^{k+1} = \arg\min_d \frac{\lambda}{2} \|d - \psi(w^{k+1}) + b^k\|_{\ell_2}^2 + \|d\|_{\ell_1}$

$b_i^{k+1} = b_i^k + \beta_i w_i^{k+1} - d_i^{k+1}$

$k = k + 1$

end while

Output ζ and d^k where ζ is defined as in Theorem 4.3 using $\epsilon = \frac{2}{(\sqrt{2}+1)^2} tol$ and $\tilde{w} = w^k$.

By Theorem 4.3 this algorithm ensures that the objective value is within *tol* of the optimal value.

The first sub-problem in Algorithms 5 and 6 involves solving a $N \times N$ system of equations. When the number of assets is large completing this step becomes computational expensive.

This is especially true for financial data where the covariance matrix is ill-conditioned and dense. Thus Algorithms 5 and 6 may be impractical in applications where real-time results are required or computational performance is limited.

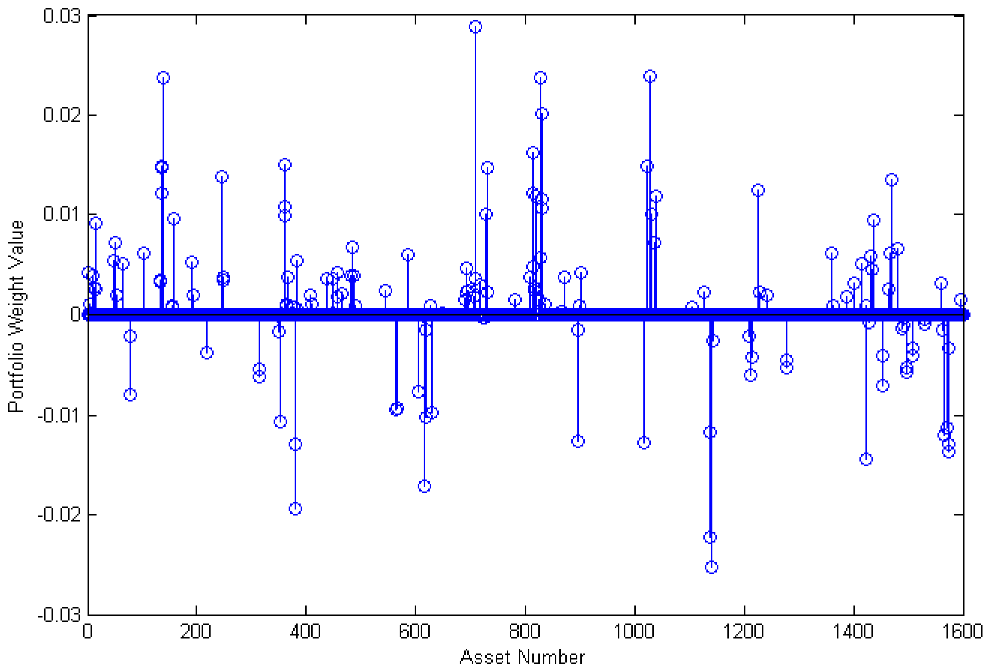

Figure 4.1: Elastic net penalty promotes sparsity in the portfolio weights

It is well known [13] that portfolio optimization problems with an ℓ_1 regularization term can result in sparse portfolios i.e. the solution of (4.3) is only non-zero in a small number of indices. Figure 4.1 illustrates this behavior by showing the portfolio weights for 1600 assets obtained using the criterion in 4.3. For this example less than 10% of the assets have a non-zero weight.

Sparsity of the portfolio weights can be exploited to reduce computational complexity. To see this suppose w^* solves (4.3) and $I = \text{supp}(w^*)$ is known a priori (before computing the solution). Then the problem (4.3) can be relaxed to the equivalent problem

$$\min_{w} \; w^T R_{|I} w - w^T \hat{\mu}_{|I} + \|w\|_{\vec{\beta},\ell_1}$$

where $R_{|I}$ and $\mu_{|I}$ represent the covariance and mean restricted to I. This problem is of dimension $|I|$ and requires fewer operations to compute per iteration. This suggests that an adaptive support split-Bregman Algorithm which attempts to solve (4.3) on smaller subspaces, I, where $\mathrm{supp}(w^*) \subset I$ can save computational time.

To develop an effective algorithm we first derive an optimality condition which can be used as a stopping criterion.

Lemma 4.1. *w^* solves (4.3) if and only if $|(2Rw^*)_i - \hat{\mu}_i| \leq \beta_i$ for all $i \notin \mathrm{supp}(w^*)$ and $(2Rw^*)_i - \hat{\mu}_i + \beta_i \mathrm{sign}(w_i^*) = 0$ for all $i \in \mathrm{supp}(w^*)$.*

Proof. Suppose w^* solves (4.3) and let $i \in \mathrm{supp}(w^*)$. Then since w^* is optimal and $w_i^* \neq 0$ the partial derivative of the objective function with respect to w_i exists and is equal to 0. Thus

$$\begin{aligned} 0 &= \frac{\partial}{\partial w_i}\Psi(w)|_{w=w^*} \\ &= 2(Rw^*)_i - \hat{\mu}_i + \beta_i \mathrm{sign}(w_i^*). \end{aligned}$$

Now suppose $i \notin \mathrm{supp}(w^*)$. Now the partial derivative of the objective function does not exist. However by optimality we have

$$0 \in \partial\Psi(w^*)$$

Thus

$$\lim_{h \downarrow 0} \frac{\Psi(w^* + h\delta_i) - \Psi(w^*)}{h} \geq 0$$

and

$$\lim_{h \uparrow 0} \frac{\Psi(w^* + h\delta_i) - \Psi(w^*)}{h} \leq 0$$

which imply

$$(2Rw^*)_i - \hat{\mu}_i \geq -\beta_i$$

and

$$(2Rw^*)_i - \hat{\mu}_i \leq \beta_i.$$

For the converse suppose that $|(2Rw^*)_i - \hat{\mu}_i| \leq \beta_i$ for all $i \notin \text{supp}(w^*)$ and $(2Rw^*)_i - \hat{\mu}_i + \beta_i \text{sign}(w_i^*) = 0$ for all $i \in \text{supp}(w^*)$. Choose $\epsilon = \min\{|w_i| : i \in \text{supp}(w)\}$. Then for any w such that $||w - w^*||_\infty < \epsilon$

$$\begin{aligned}\Psi(w) - \Psi(w^*) &\geq \sum_{i \in \text{supp}(w^*)} \left((2Rw^*)_i - \hat{\mu}_i + \beta_i \text{sgn}(w_i^*)\right)(w_i - w_i^*) + \\ &\quad + \sum_{i \notin \text{supp}(w^*)} \left((2Rw^*)_i - \hat{\mu}_i\right) w_i + \beta_i |w_i| \\ &\geq 0.\end{aligned}$$

Thus w^* is locally optimal which implies global optimality. □

Lemma 4.1 can be used to derive a criterion for determining which indices in a portfolio, x, belong in the support. For example, suppose that $i \notin \text{supp}(x)$, and $|(2Rx)_i - \hat{\mu}_i| > \beta_i$. Then the objective function in (4.3) can be reduced by adding i into $\text{supp}(x)$. Thus x is not optimal and we should incorporate i into $\text{supp}(x)$.

Next we look at how to prolongate the adaptive support split-Bregman variables (w, d, b) from a lower dimensional space to a higher dimensional space. Prolongation of w and d can be achieved through simple zero filling. Prolongation of b is more delicate. The following Lemma suggests an effective prolongation.

Lemma 4.2. *Suppose (w^*,d^*) is the solution of* (4.18) *obtained with Algorithm 5. Then*

$$\lim_{k \to \infty} b_i^k = -(2Rw^* - \hat{\mu})_i / (\beta_i \lambda). \tag{4.19}$$

Proof. By Algorithm 5 we have for all k

$$2(Rw^{k+1})_i - \hat{\mu}_i - \lambda(d^k - \psi(w^{k+1}) - b^k)_i\beta_i = 0.$$

Since $\lim_{k\to\infty} w^k = w^*$ and $\lim_{k\to\infty} d^k = d^*$ and $d^* = \psi(w_i^*)$ we have

$$\lim_{k\to\infty} 2(Rw^{k+1})_i - \hat{\mu}_i + \lambda(b^k)_i\beta_i = 0$$

which implies that

$$\lim_{k\to\infty}(b^k)_i = \frac{\hat{\mu}_i - 2(Rw^*)_i}{\beta_i\lambda}.$$

□

This suggests that the prolongation of b can be defined from equation (4.19). For example suppose $(\tilde{w}, \tilde{d}, \tilde{b})$ solves (4.18) on a restricted domain $I \subset \{1, 2, \ldots N\}$ and let w and d represent the prolongation of $\tilde{w}$ and $\tilde{d}$ to a set $J \supset I$ i.e.

$$w_j = \begin{cases} \tilde{w}_j & \text{if } j \in I \\ 0 & \text{if } j \in J - I \end{cases} \tag{4.20}$$

$$d_j = \begin{cases} \tilde{d}_j & \text{if } j \in I \\ 0 & \text{if } j \in J - I \ . \end{cases} \tag{4.21}$$

Then taking a cue from equation (4.19) the prolongation of $\tilde{b}$ may be defined as

$$b_i = (-2R_{|J}w + \hat{\mu}_{|J})_i / (\beta_i\lambda). \tag{4.22}$$

The adaptive support split-Bregman Algorithm for solving (4.18) is given below.

Algorithm 7 Adaptive support split-Bregman algorithm for solving 4.18

Initialize: $k = 0, w^0 = 0, d^0 = 0, b^0 = 0, \epsilon > 0, M > 0$

Define $D^0 = 2Rw^0 - \hat{\mu}$

while $|D_i^k| > \beta_i$ for any $i \notin \text{supp}(w^k)$ AND $k < N$ **do**

Define the set $J^k = \{D_i^k : i \notin \text{supp}(w^k)\}$

Set $K = M \vee (k + 1 - |\text{supp}(w^k)|)$

Set $\tilde{J}^k$ equal to the largest K elements in J^k

Set $I^k = \tilde{J}^k \cup \text{supp}(w^k)$

Run Algorithm 6 on I^k with initialization $w^k_{|I^k}, b^k_{|I^k}, d^k_{I^k}$ and tolerance ϵ

Set (w^{k+1}, d^{k+1}) to the prolongation of output of previous step

Set $b_i^{k+1} = -2(Rw^{k+1} - \hat{\mu})_i/(\beta_i\lambda)$,

Set $D^{k+1} = 2Rw^{k+1} - \hat{\mu}$

$k = k + 1$

end while

The next theorem shows that Algorithm 7 converges.

Theorem 4.4. *Let w^* be the optimal solution to* (4.3) *and let w' be a solution produced by Algorithm 7 for $\epsilon = tol$. Then*

$$\Psi(w') \leq \Psi(w^*) + tol. \tag{4.23}$$

Proof. By design the algorithm terminates after at most N iterations. Suppose the algorithm terminates in $k < N$ iterations. Let $I^{(k)}$ be the support in iteration k of the adaptive support split-Bregman algorithm. Then by the proof of Theorem 4.3, w' satisfies the conditions of Theorem 4.2 with $\epsilon = tol$. Thus by Theorem 4.2 $\Psi(w') < \Psi(w^*) + tol$. Now suppose the algorithm terminates in N iterations. Since $I^{(N-1)}$ contains all asset indices it follows by the design of Algorithm 6 that $\Psi(w') < \Psi(w^*) + tol$. □

To evaluate the execution speed of adaptive support split-Bregman algorithm we performed a comparison with the following fast algorithms described in the literature: split-Bregman algorithm (Algorithm 6), FISTA [7] and Multilevel Iterated-Shrinkage [84]. To the best of our knowledge these algorithms are considered state of the art for large-scale ℓ_1-penalized quadratic programs. For the multi-level algorithm proposed in [84] we use the FISTA [7] algorithm for all relaxations and lowest level solvers. To make a fair comparison we have used the same error tolerance of 10^{-6} for each algorithm.

Tables 4.1 and 4.2 presents MATLAB run times for solving (4.3) for a large and small basket of US stocks. The machine running the simulation has the Windows 7 operating system and an Intel i7-3740 processor with 32.0 GB of RAM.

Table 4.1: Adaptive support split-Bregman converges quickly to a solution for sparse portfolios

Dimension	Sparsity Level	Adaptive Support Split-Bregman	Split-Bregman	FISTA	Multi-level FISTA [84]
2000	88	0.1 sec	20.6 sec	0.4 sec	0.2 sec
2000	142	0.2 sec	14.5 sec	0.8 sec	0.2 sec
2000	450	0.9 sec	14.6 sec	3.6 sec	1.5 sec
2000	853	4.8 sec	23.0 sec	8.8 sec	9.2 sec
2000	1692	10.4 sec	38.0 sec	21.4 sec	22.7 sec
3000	237	0.3 sec	48.2 sec	12.9 sec	2.7 sec
3000	805	1.3 sec	49.9 sec	55.7 sec	24.6 sec
4000	234	0.5 sec	107.6 sec	24.6 sec	2.2 sec

In Table 4.1 we see that the adaptive support split-Bregman algorithm converges much faster than both split-Bregman, FISTA and Multi-Level FISTA for sparse portfolios taken from a large set of assets. On the other hand Tables 4.1 and 4.2 show that the advantage of the adaptive support split-Bregman algorithm decreases when the cardinality of the asset set is small or when the support of the portfolio is large.

Table 4.2: Benefit of adaptive support split-Bregman decreases when dimensionality is small

Dimension	Sparsity Level	Adaptive support split-Bregman	Split-Bregman	FISTA	Multi-level FISTA [84]
500	53	0.03 sec	0.8 sec	0.02 sec	0.02 sec
500	150	0.09 sec	0.6 sec	0.04 sec	0.03 sec
500	261	0.2 sec	0.5 sec	0.2 sec	0.2 sec

4.2.3 Proofs of Technical Results

In this section we provide proofs for Theorems 4.2 and 4.3. To facilitate the proof we will first reformulate the criterion in (4.3) as a quadratic program.

Problem (4.3) can be reformulated as a quadratic program with linear inequality constraints by introducing an auxiliary variable d,

$$\begin{aligned} \min_{w,d} \ & \Phi(w,d) \\ \text{s.t.} \ & -d_i \le w_i \\ & -d_i \le -w_i \end{aligned} \tag{4.24}$$

where $\Phi(w,d) = w^T R w - w^T \hat{\mu} + \sum_{i=1}^{N} \beta_i d_i$ and where $R = \hat{\Gamma} + D_\alpha$. The Lagrangian for this problem

$$L(w,d,\lambda) = w^T R w - w^T \hat{\mu} + \sum_{i=1}^{N} \beta_i d_i + \sum_{i=1}^{N} \lambda_i(-d_i - w_i) + \sum_{i=1}^{N} \lambda_{i+N}(-d_i + w_i) \tag{4.25}$$

plays an important role in our subsequent analysis in the next section.

Now we prove Theorems 4.2 and 4.3 using the quadratic program reformulation (4.24). Our first task is to derive a lower bound on the Lagrangian for a fixed λ and when $d = |w|$. First note that R is symmetric positive definite whose smallest eigenvalue is $\geq \alpha_o$ where

$$\alpha_o = \min\{\alpha_i : 1 \leq i \leq N\}.$$

Thus for $d_i = |w_i|$,$\tilde{d}_i = |\tilde{w}_i|$ and $\lambda > 0$ we have

$$\begin{aligned}
\Phi(w,d) &\geq L(w,d,\lambda) \\
&= L(\tilde{w},\tilde{d},\lambda) + \nabla_w L(\tilde{w},\tilde{d},\lambda)^T(w-\tilde{w}) \\
&\qquad + \nabla_d L(\tilde{w},\tilde{d},\lambda)^T(d-\tilde{d}) + (w-\tilde{w})^T H_w(\tilde{w},\tilde{d},\lambda)(w-\tilde{w}) \\
&\geq L(\tilde{w},\tilde{d},\lambda) + \nabla_w L(\tilde{w},\tilde{d},\lambda)^T(w-\tilde{w}) + \nabla_d L(\tilde{w},\tilde{d},\lambda)^T(d-\tilde{d}) \\
&\qquad + \alpha_o \|w-\tilde{w}\|_{\ell_2}^2 \\
&\geq L(\tilde{w},\tilde{d},\lambda) + \nabla_w L(\tilde{w},\tilde{d},\lambda)^T(w-\tilde{w}) + \nabla_d L(\tilde{w},\tilde{d},\lambda)^T(d-\tilde{d}) \\
&\qquad + \frac{1}{2}\alpha_o \|w-\tilde{w}\|_{\ell_2}^2 + \frac{1}{2}\alpha_o \|d-\tilde{d}\|_{\ell_2}^2
\end{aligned} \tag{4.26}$$

where H_w is the Hessian of L w.r.t to the w variables.

We now present two lemmas which will be useful in deriving a stopping criterion. Our first lemma gives an upper bound for L when the gradient of L is small.

Lemma 4.3. *Suppose $d_i = |w_i|$ for all i and $\|\nabla_{w,d} L(\tilde{w},\tilde{d},\lambda)\|_{\ell_2} \leq \sqrt{2\epsilon\alpha_o}$. Then $L(\tilde{w},\tilde{d},\lambda) \leq \Phi(w^*,d^*) + \epsilon$ where w^* solves (4.3) and $d_i^* = |w_i^*|$ for all i.*

Proof. By equation (4.26) we have

$$\begin{aligned}\Phi(w^*,d^*) \geq L(w^*,d^*,\lambda) \quad \geq \quad & L(\tilde{w},\tilde{d},\lambda) + \nabla_w L(\tilde{w},\tilde{d})^T(w^* - \tilde{w}) + \nabla_d L(\tilde{w},\tilde{d})^T(d^* - \tilde{d}) \\ & + \frac{1}{2}\alpha_o \|w^* - \tilde{w}\|_{\ell_2}^2 + \frac{1}{2}\alpha_o \|d^* - \tilde{d}\|_{\ell_2}^2.\end{aligned}$$

The righthand side is minimized by substituting $-\frac{1}{\alpha_o}\nabla_d L(\tilde{w},\tilde{d},\lambda)$ for $(d^* - \tilde{d})$ and substituting

$$-\frac{1}{\alpha_o}\nabla_w L(\tilde{w},\tilde{d},\lambda)$$

for $(w^* - \tilde{w})$. With these substitutions we obtain

$$\begin{aligned}\Phi(w^*,d^*) \quad & \geq \quad L(\tilde{w},\tilde{d},\lambda) - \frac{1}{2\alpha_o}\|\nabla_{w,d} L(\tilde{w},\tilde{d},\lambda)\|_{\ell_2}^2 \\ & \geq \quad L(\tilde{w},\tilde{d},\lambda) - \epsilon.\end{aligned}$$

□

The next lemma can be verified easily.

Lemma 4.4. *Suppose* $|a| \leq b$. *Then there exist* $x_1, x_2 \geq 0$ *such that*

$$\begin{aligned} x_1 + x_2 &= b \\ -x_1 + x_2 &= a. \end{aligned}$$

Proof of Theorem 4.2

We are now ready to prove Theorem 4.2 which establishes a condition for approximate optimality of a portfolio under the weighted elastic net criterion (4.3).

Proof. of Theorem 4.2

Choose d^* and $\tilde{d}$ such that $d_i^* = |w_i^*|$ and $\tilde{d}_i = |\tilde{w}_i|$. For $i \in \operatorname{supp}(\tilde{w})$ define λ such that

$$\lambda_i = \begin{cases} 0 \text{ if } w_i > 0, i \in \operatorname{supp}(\tilde{w}) \\ \beta_i \text{ if } w_i < 0, i \in \operatorname{supp}(\tilde{w}) \end{cases}$$

and for $i \in \operatorname{supp}(\tilde{w})$,define $\lambda_{i+N} = \beta_i - \lambda_i$.

For $i \notin \operatorname{supp}(\tilde{w})$ we want to define λ_i and λ_{i+N} such that $\lambda_i \geq 0$, $\lambda_{i+N} \geq 0$,

$$\lambda_i + \lambda_{i+N} = \beta_i \tag{4.27}$$

and

$$-\lambda_i + \lambda_{i+N} = -\frac{\partial}{\partial w_i}\left(w^T R w - w^T \hat{\mu}\right)\Big|_{w=\tilde{w}}. \tag{4.28}$$

By Lemma 4.4, equation (4.13) implies that such a λ_i, λ_{i+N} exists.

Let us form the Lagrangian $L(w, d, \lambda)$ as in equation (4.25). Then for $i \in \operatorname{supp}(\tilde{w})$

$$\frac{\partial}{\partial w_i} L(w, d, \lambda)_{|(\tilde{w}, \tilde{d})} = \frac{\partial}{\partial w_i}\left(w^T R w - w^T \hat{\mu} + \|w\|_{\vec{\beta}, \ell_1}\right)\Big|_{w=\tilde{w}}$$

and

$$\frac{\partial}{\partial d_i} L(w, d, \lambda)_{|(\tilde{w}, \tilde{d})} = 0.$$

For $i \notin \operatorname{supp}(\tilde{w})$ we have by equation (4.28)

$$\frac{\partial}{\partial w_i} L(w, d, \lambda)_{|(\tilde{w}, \tilde{d})} = 0$$

and by equation (4.27)

$$\frac{\partial}{\partial d_i} L(w, d, \lambda)_{|(\tilde{w}, \tilde{d})} = 0.$$

It then follows from equation (4.12) that

$$\|\nabla_{w,d} L(\tilde{w}, \tilde{d}, \lambda)\|_{\ell_2} \leq \sqrt{2\epsilon\alpha_o}$$

and so by Lemma 4.3 and our choice of λ we have that

$$\begin{aligned} \Phi(\tilde{w}, \tilde{d}) &= L(\tilde{w}, \tilde{d}, \lambda) \\ &\leq \Phi(w^*, d^*) + \epsilon. \end{aligned}$$

This clearly implies that

$$\Psi(\tilde{w}) \leq \Psi(w^*) + \epsilon.$$

$\square$

Proof of Theorem 4.3

Now we prove Theorem 4.3 which can be used to establish a more practical convergence criterion than Theorem 4.2.

Proof. of Theorem 4.3

By construction $\|\zeta - \tilde{w}\|_{\ell_\infty} \leq \|\zeta - \tilde{w}\|_{\ell_2} \leq \frac{\epsilon \wedge \sqrt{\epsilon\alpha_o}}{M}$. It follows that

$$\sum_{i \epsilon \mathrm{supp}(\zeta)} \left(\frac{\partial}{\partial w_i} \left(w^T R w - w^T \hat{\mu} + \|w\|_{\vec{\beta},\ell_1} \right) \Big|_{w=\zeta} \right)^2 \leq (\sqrt{2} + 1)^2 \alpha_o \epsilon$$

and

$$-\beta_i \leq \frac{\partial}{\partial w_i} \left(w^T R w - w^T \hat{\mu} \right) \Big|_{w=\zeta} \leq \beta_i$$

for all $i \notin \mathrm{supp}(\zeta)$. So by Theorem 4.2 we have that ζ satisfies (4.17). $\square$

4.3 Extension to pairwise weighted elastic net (PWEN)

As a further generalization one can also consider the pairwise weighted elastic net (PWEN) (4.8). As was shown in the proof of Theorem 4.1, PWEN also has a robust optimization motivation. Thus the bootstrap method can be used to calibrate the weights in PWEN. Here for the robust optimization interpretation to hold the matrix Δ must be chosen such that matrices of the form

$$\begin{aligned} R_{i,j} &= \begin{cases} \hat{\Gamma}_{i,i} + \Delta_{i,i} \text{ if } i = j \\ \hat{\Gamma}_{i,j} \pm \Delta_{i,j} \text{ if } i \neq j \end{cases} \\ R_{i,j} &= R_{j,i} \end{aligned}$$

are positive semi-definite. This can be guaranteed by selecting Δ to be symmetric,diagonally dominant and non-negative. In practice diagonal dominance may be too restrictive given the bootstrap estimated uncertainties. For those cases the Δ matrix does not need to be diagonal dominant, however the robust optimization interpretation will not hold. As we show in section 4.4 this will not necessarily prevent one from obtaining good portfolio performance.

Application of split-Bregman techniques to PWEN is complicated due to the cross products of $|w_i|$ and $|w_j|$ induced through Δ. As an alternative numerical solution of the PWEN criterion can be performing via coordinate descent as described in [63]. Coordinate descent converges for the PWEN problem provided Δ is positive definite [45].

4.4 Performance Evaluation

In this section we quantify the performance benefit of using a weighted elastic net penalty by testing our criterion in (4.3) on daily return data from 630 U.S. stocks collected between

January 1, 2001 and July 1, 2014 with market capitalization greater than 4 billion US dollars. The results are then compared with other portfolio selection criteria described in Section 2.1 and the naive equal-weighted portfolio.

In our experiments we compute new portfolios every 63 trading days using daily returns from the prior 252 trading days as training data for parameter estimation and calibration of the elastic net weights. Our criteria for evaluating the portfolio performance is the out-of-sample Sharpe ratio of the daily portfolio returns. Sharpe ratio is defined as the portfolio's excess return divided by its standard deviation. The formula used for computing the Sharpe ratio is given below

$$SR = \frac{\frac{1}{\tau}\sum_{i=1}^{\tau} w(t_i)^T r(t_i)}{\sqrt{\frac{1}{\tau}\sum_{i=1}^{\tau}\left(w(t_i)^T r(t_i) - \frac{1}{\tau}\left(\sum_{j=1}^{\tau} w(t_j)^T r(t_j)\right)\right)^2}} \tag{4.29}$$

where τ is the total number of trading days in our 13.5 year data set. Here $w(t_i)$ is the portfolio on day t_i, which is computed from the previous set of training data and remains fixed over intervals of 63 trading days.

4.4.1 Parameter Selection

Due to the large number of assets and small amount of training data, estimation of the covariance and mean in our experiments is performed using shrinkage techniques [29]. We estimate the covariance matrix using the technique described in [54]. In that paper the following shrinkage estimator for Γ is proposed

$$\hat{\Gamma} = \rho_1 \Gamma_S + \rho_2 I \tag{4.30}$$

where Γ_S is the sample average covariance obtained from the training data and where ρ_1, ρ_2 are > 0. In our experiments we use the optimal values of $\rho_1 > 0$ and $\rho_2 > 0$ which are derived in [54]. Note that this choice of shrinkage target guarantees that $\hat{\Gamma}$ will be positive definite.

Since the weighted elastic net penalty consists of a squared weighted ℓ_2 norm, the shrinkage in (4.30) may appear to be redundant when applied with the weighted elastic net regularization in Section 4.1. However, this is not the case since the weights on the weighted elastic net and the shrinkage parameters in (4.30) are adaptively selected according to different criteria. Thus the covariance shrinkage target becomes a combination of the bootstrap derived target and the target derived according to [54]. One benefit of this approach is that there will always be some level of ℓ_2 regularization regardless of what the bootstrap criterion derives.

For estimation of the mean we employ a James-Stein estimator [27, 48] which was proposed for portfolio optimization in [50]. When applying the James-Stein approach we compute the estimate of μ using the equation

$$\hat{\mu} = (1 - \rho)\mu_S + \rho\eta\vec{1}. \tag{4.31}$$

Here μ_S is the sample mean vector and η is the maximum of average of the sample means and the daily historical return of the US stock market between 1928 and 2000 [23]

$$\eta = \left(\frac{1}{N}\sum_{i=1}^{N}\mu_{S,i}\right) \vee 0.0004. \tag{4.32}$$

The value of ρ is set according to [50] as

$$\rho = \min\left\{1, \frac{(N-2)}{T_{train}(\mu_S - \eta\vec{1})^T\hat{\Gamma}^{-1}(\mu_S - \eta\vec{1})}\right\}. \tag{4.33}$$

The weights for the weighted elastic net penalty are calibrated using the bootstrap technique described in Section 4.1.2 with identical estimation risk aversion factors for mean and squared volatility i.e. $p_1 = p_2$. Calibration of the weighted LASSO penalty is performed using the technique described in [36]. Since the weighted LASSO calibration in [36] is only defined up to a constant we perform a parametric study for various constants. Calibration of the elastic net penalty is handled using the technique described in section 1.6.2 of [87]. The calibration method in [87] only determines the sum $\lambda_1 + \lambda_2$ in (2.6), the relative weighting of λ_1 and λ_2 is not addressed. Thus we perform a parametric analysis over the relative weighting between the parameters λ_1 and λ_2 in the elastic net. For SCAD there are no known calibration methods. Hence for SCAD we perform a parametric study for various λ values and a fixed a_{SCAD} parameter of 3.7 as suggested in [32].

For the PWEN criterion the β weights are set to the same value as the WEN β values. The Δ term in the PWEN criterion is defined as

$$\begin{aligned} \Delta_{i,i} &= \alpha_i \\ \Delta_{i,j} &= \frac{\sqrt{\alpha_i \alpha_j}}{5} \text{ for } i \neq j. \end{aligned}$$

where α are the ℓ_2 weights computed for WEN. Note that Δ in general may not be diagonally dominant and thus the robust optimization interpretation of the PWEN penalty may not hold.

4.4.2 Sharpe Ratio performance

In this section we present performance results for the following 6 criteria:

1. Markowitz (no penalty)
2. Markowitz (WEN penalized)
3. Markowitz (Weighted LASSO penalized) [36]
4. Markowitz (Elastic net penalized) [88]
5. Markowitz (SCAD penalized)

6. $1/N$ equal weighted portfolio

7. Markowitz (PWEN penalized)

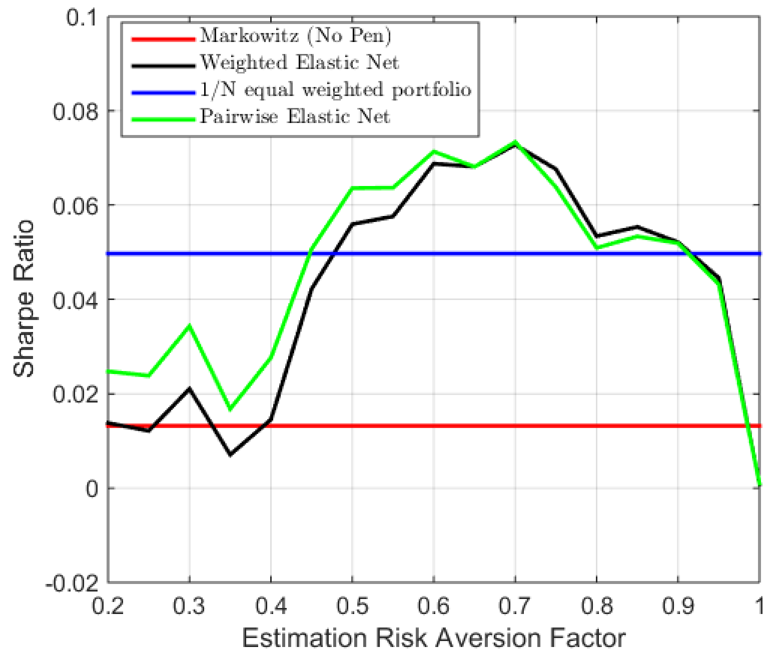

Figure 4.2: Parametric analysis of WEN and PWEN performance using bootstrap calibration

In Figure 4.2 we present the Sharpe ratios of the weighted elastic net penalty as a function of estimation risk aversion factor, i.e. bootstrap percentile. As a comparison the performance of the $1/N$ and unpenalized portfolio are also shown. The figure demonstrates that the PWEN and WEN criteria with bootstrap calibration improves Sharpe ratio performance over the $1/N$ and unpenalized portfolio when the estimation risk aversion factor is between 0.5 and 0.95. Outside of this interval the WEN penalty did not improve performance, which suggests that a moderate amount of estimation risk aversion is optimal for WEN. The PWEN penalized criterion does exhibit some performance gain for risk aversion less than 0.5. This could be attributed to the additional penalization caused by the off-diagonal terms in Δ. For comparison purposes the Sharpe ratio of the weighted LASSO, elastic net and SCAD penalized portfolios are shown in Figures 4.3, 4.4, and 4.5 as a function of their respective penalty scaling parameter. We see that both weighted LASSO and the elastic net do not perform as well as the WEN and PWEN penalties. This could be a consequence of their calibration being derived from a minimum variance perspective. The SCAD penalized portfolio performs comparable to the weighted elastic net penalty if the λ parameter is chosen correctly. However, it is still an open question on how to automate the selection of an optimal λ in the SCAD penalty for portfolio optimization problems.

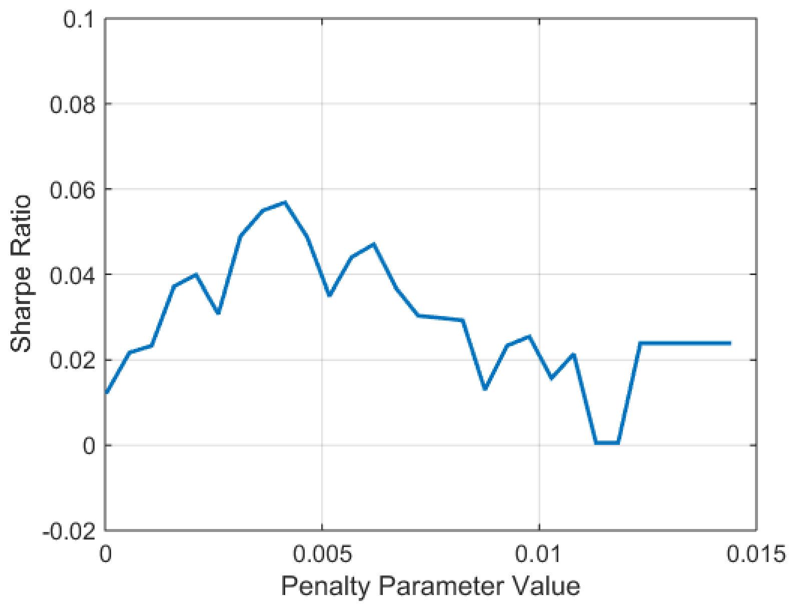

Figure 4.3: Parametric analysis of Weighted LASSO performance

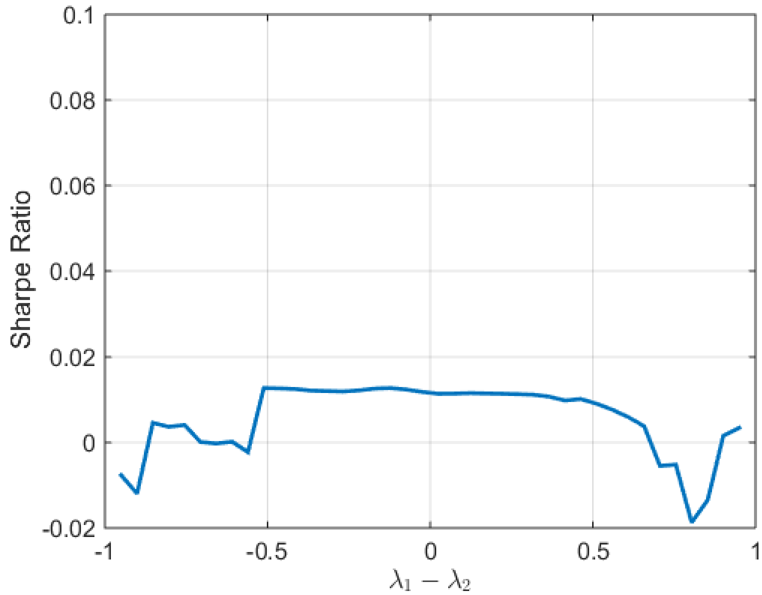

Figure 4.4: Parametric analysis of Elastic Net performance

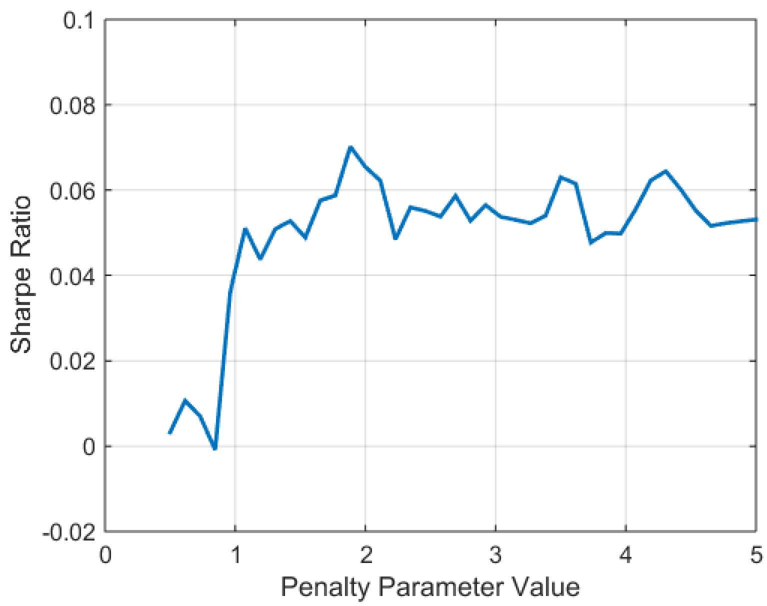

Figure 4.5: SCAD performance as a function of λ parameter.

Chapter 5

Conclusions and Future Work

In this dissertation we examined techniques for enhancing mean-variance portfolio performance in the presence of parameter uncertainty. The approaches proposed in this work draw on recent advances in sparse modeling and optimization techniques. Experimental results using both real world and simulated data demonstrate the merits of using sparse modeling in portfolio design.

In Chapter 3 we introduced two sparse Kalman filtering methods for estimating asset return covariance from high-frequency data. The methods address 3 features found in high frequency data: 1) asynchronous returns, 2) market microstructure noise, and 3) jumps. The first method, a KECM approach, was derived using both Laplace and spike and slab distributed jump models. The second method utilized a MCMC approach to approximate the posterior mean of the covariance estimate where the jumps were modeled using a spike and slab distribution.

Each of these proposed covariance estimation techniques show improved performance versus existing methods when jumps are present and are robust to other stylized facts such as volatility clustering and stochastic microstructure noise variance. When comparing the spike and slab and Laplace jump models using simulated data, the spike and slab approach is more robust to larger jumps and with stochastic microstructure noise variance. The simulated data experiments also show that the MCMC approaches perform better than KECM when large jumps occur.

In Chapter 4 the addition of a weighted elastic net penalty to mean-variance objective function is proposed in order to improve out-of-sample portfolio performance when parameter estimates are uncertain. We have shown that this approach can be motivated by reformulating the mean-variance criterion as a robust optimization problem. With this view we develop a data-driven criterion for calibration of the elastic net weights based on bootstrapping and an investor's aversion to model estimation risk. To compute the portfolio weights efficiently we proposed an adaptive support split-Bregman algorithm for solving our proposed optimization criterion. This technique exploits the sparsity promoting properties of the weighted elastic net penalty to reduce computational requirements.

Our experimental results demonstrate that using the weighted elastic net penalty and calibration approach can result in higher out-of-sample Sharpe ratio than the other norm penalization techniques designed for minimum variance portfolios. In addition, our MATLAB run-time results indicate that the proposed adaptive support split-Bregman algorithm significantly reduces computation time compared with other algorithms such as split-Bregman and FISTA.

There are many possible extensions to the techniques proposed in this dissertation. For example covariance matrix estimation from high frequency data could benefit from low rank + sparse matrix factorization techniques. Although the preliminary results using nuclear norm penalization shown in Appendix 3.5 fall short of KECM performance, other techniques such as weighted nuclear norms or transformed Schatten-1 penalties [92] may improve performance. For the pairwise weighted elastic net it is interesting whether the restrictions on the weighting matrix Δ can be relaxed from the diagonal dominant restriction. This relaxation coupled with an effective calibration procedure could enhance PWEN performance beyond what was reported in this dissertation.

Bibliography

[1] A. Aravkin, B. Bell, J. Burke, and G. Pilonetto. An ℓ_1 Laplace robust Kalman smoother. *IEEE Transactions on Automatic Control*, 56(12), Dec. 2011.

[2] Y. Aït-Sahalia, J. Fan, and D. Xiu. High-Frequency Covariance Estimates With Noisy and Asynchronous Financial Data. *Journal of the American Statistical Association*, 105(492):1504–1517, 2010.

[3] Y. Aït-Sahalia, P. Myklank, and L. Zhang. How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise. *Review of Financial Studies*, 100:1394–1411, 2005.

[4] F. Bandi and J. Russell. Separating microstructure noise from volatility. *Journal of Financial Economics*, 79:655–692, 2006.

[5] O. Barndorff-Nielsen, P. Hansen, A. Lunde, and N. Shephard. Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics*, 162, 2011.

[6] C. Barry. Portfolio Analysis under Uncertain Means,Variances and Covariances. *Journal of Finance*, 29:515–522, May 1974.

[7] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Imaging Sciences*, 2(1):183–202, 2009.

[8] T. Bollerslev. Modeling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH approach. *Review of Economics and Statistics*, 72:498–505, 1990.

[9] K. Boudt, C. Croux, and S. Laurent. Outlyingness weighted covariation. *Journal of Financial Econometrics*, 9(4), 2011.

[10] K. Boudt and J. Zhang. Jump robust two time scale covariance estimation and realized volatility budgets. *Quantitative Finance*, 15(6), 2015.

[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[12] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, 2004.

[13] J. Brodie, I. Daubechies, C. DeMol, Y. Giannone, and I. Loris. Sparse and Stable Markowitz Portfolios. *PNAS*, 106(30):12267–12272, 2009.

[14] J. Campbell, A. Lo, and A. MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, 1996.

[15] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis. *Journal of the ACM (JACM)*, 58(3):11, 2011.

[16] E. Candès and Y. Plan. Matrix completion with noise. *Proc. IEEE*, 98(6):925–936, June 2010.

[17] E. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted ℓ_1 minimzation. *Journal of Fourier Anal. Appl.*, 14:877–905, 2008.

[18] M. Capinski and T. Zastawniak. *Mathematics for Finance: An Introduction to Financial Engineering*. Springer, 2003.

[19] G. Casella and E. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, Aug 1992.

[20] V. Cevher, S. Becker, and M. Schmidt. Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *Signal Processing Magazine,IEEE*, 31(5):32–43, 2014.

[21] W. Chan and J. Maheu. Conditional jump dynamics in stock market returns. *Journal of Business and Economic Statistics*, 20(3):377–389, 2002.

[22] F. Corsi, S. Peluso, and F. Audrino. Missing in Asynchronicity: A Kalman-EM Approach for Multivariate Realized Covariance Estimation. *Journal of Applied Econometrics*, 30(3), 2015.

[23] A. Damodaran. Annual Returns on Stock, T.Bonds and T.Bills: 1928 - Current. http://www.stern.nyu.edu/ adamodar/pc/datasets/histretSP.xls.

[24] V. DeMiguel, L. Garlappi, F. Nogales, and R. Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, pages 798–812, 2009.

[25] V. DeMiguel, L. Garlappi, and R. Uppal. Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2009.

[26] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

[27] B. Efron and C. Morris. Stein's Estimation Rule and its Competitors- An Empirical Bayes Approach. *Journal of American Statistical Association*, 68(341):117–130, March 1973.

[28] B. Efron and C. Morris. Families of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics*, 4:11–21, Jan. 1976.

[29] B. Efron and C. Morris. Stein's Paradox in Statistics. *Scientific American*, 237:119–127, 1977.

[30] B. Efron and R. Tibshirani. Bootstrap methods for standard errors,confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, Feb 1986.

[31] E. Fama and K. French. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33:3–56, 1993.

[32] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[33] J. Fan, Y. Li, and K. Yu. Vast Volatility Matrix Estimation Using High Frequency Data for Portfolio Selection. *Journal of the American Statistical Association*, 107(497):412–428, 2012.

[34] J. Fan and Y. Wang. Multi-scale jump and volatility analysis for high-frequency financial data. *Journal of the American Statistical Association*, 102(480):1349–1362, 2007.

[35] J. Fan, J. Zhang, and K. Yu. Vast Portfolio Selection With Gross-Exposure Constraints. *Journal of the American Statistical Association*, 107(498):592–607, 2012.

[36] B. Fastrich, S. Paterlini, and P. Winker. Constructing optimal sparse portfolios using regularization methods. *Comput. Management Science*, 12(3):417–434, Dec. 2014.

[37] D. Fink. A compendium of conjugate priors. Technical report, Montana State Univeristy, 1997.

[38] S. Frühwirth-Schnatter. Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15:183–202, 1994.

[39] P. Frost and J. Savarino. An Empirical Bayes Approach to Efficient Portfolio Selection. *The Journal of Financial and Quantitative Analysis*, 21(3):293–305, Sept 1986.

[40] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*. CRC Press, 3rd edition, 2014.

[41] Z. Ghahramani and G. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12(4):963–996, 2000.

[42] D. Goldfarb and G. Iyengar. Robust Portfolio Selection Problems. *Mathematics of Operations Research*, 28(1):1–38, 2003.

[43] T. Goldstein and S. Osher. The Split Bregman Method for ℓ_1 Regularized problems. *SIAM Journal on Imaging Sciences*, 2(2):323–343, 2009.

[44] B. Halldorsson and R. Tutuncu. An Interior Point Method for a Class of Saddle Point Problems. *Journal of Optimization Theory and Applications*, 116(3):559–590, March 2003.

[45] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC Press, Boca Raton,FL, 2015.

[46] M. Ho, Z. Sun, and J. Xin. Weighted elastic net penalized mean-variance portfolio design and computation. *SIAM J. Financial Mathematics*, 6:1220–1244, 2015.

[47] R. Jagannathan and T. Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance*, 58:1651–1684, 2003.

[48] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, pages 361–379, 1961.

[49] J. Jobson and B. Korkie. Estimation for Markowitz Efficient Portfolios. *Journal of the American Statistical Association*, 75:544–554, Sept 1980.

[50] P. Jorion. Bayes-Stein Estimation for Portfolio Analysis. *The Journal of Financial and Quantitative Analysis*, 21(3):279–292, Sept 1986.

[51] J. Karpoff. The relation between price changes and trading volume: A survey. *The Journal of Financial and Quantitative Analysis*, 22:109–126, Mar. 1987.

[52] T. Lai and H. Xing. *Statistical Models and Methods for Financial Markets*. Springer, 2008.

[53] O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10:603–621, 2003.

[54] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices . *Journal of Multivariate Analysis*, 88:365–411, 2004.

[55] O. Ledoit and M. Wolf. Honey, I Shrunk the Sample Covariance Matrix. *The Journal of Portfolio Management*, 30(4):110–119, 2004.

[56] E. Lehmann. *Theory of Point Estimation*, volume 1. Springer-Verlag, New York, 1998.

[57] J. Li. Sparse and stable portfolio selection with parameter uncertainty. *Journal of Business and Economic Statistics*, To Appear.

[58] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2009.

[59] R. Litterman and K. Winkelmann. Estimating Covariance Matrices. *Risk Management Series*, 1998. Goldman Sachs.

[60] C. Liu and C. Tang. A quasi-maximum likelihood approach for integrated covariance matrix estimation with high frequency data. *Journal of Econometrics*, 180, 2014.

[61] J. Liu, W. Wong, and A. Kong. Covariance structure of the Gibbs sampler with applications to the comparision of estimators and augmentation schemes. *Biometrika*, 81:27–40, 1994.

[62] A. Lo and A. MacKinlay. An econometric analysis of nonsynchronous trading. *Journal of Econometrics*, 45:181–211, 1990.

[63] A. Lorbert, D. Eis, V. Kostina, D. Blei, and P. Ramadge. Exploiting covariate similarity in sparse regression via the pairwise elastic net. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, volume 9 of *Journal of Machine Learning Research*, pages 477–484, Chia Laguna Resort, Sardinia, Italy, 2010.

[64] D. Lunenberger and Y. Ye. *Linear and Nonlinear Programming.* New York,NY, 2008.

[65] J. Maheu and T. McCurdy. News arrival, jump dynamics, and volatility components for individual stock returns. *Journal of Finance*, 59(2):755–793, 2004.

[66] H. Markowitz. Portfolio selection. *Journal of Finance*, 7:77–91, 1952.

[67] J. Mattingley and S. Boyd. Real-Time Convex Optimization in Signal Processing. *IEEE Signal Processing Magazine*, May 2010.

[68] X. Meng and D. Rubin. Maximum likelihood estimation via the ECM algorithm:A general framework. *Biometrika*, 80:267–278, 1993.

[69] R. Merton. On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, 8:323–361, 1980.

[70] R. Michaud. The Markowitz optimization enigma: Is optimized optimal? *Finan, Anal. J.*, 45:31–42, 1989.

[71] R. Michaud. *Efficient Asset Management: A Practicial Guide to Stock Portfolio Optimization.* Oxford Univ. Press, New York, 1993.

[72] T. Mitchell and J. Beauchamp. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023–1032, Dec 1988.

[73] S. Mohamed, K. Heller, and Z. Ghahramani. Bayesian and ℓ_1 Approaches for Sparse Unsupervised Learning. *Proceedings of the 29th International Conference on Machince Learning*, Edinburgh, Scotland, 2012.

[74] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(Ser B):125–161, 2014.

[75] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 1999.

[76] S. Peluso, F. Corsi, and A. Mira. A Bayesian High-Frequency Estimator of the Multivariate Covariance of Noisy and Asynchronous Returns. *Journal of Financial Econometrics, To Appear.*

[77] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

[78] R. Roll. A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market. *The Journal of Finance*, 39(4):1127–1139, 1984.

[79] M. Seeger. Bayesian Inference and Optimal Design for the Sparse Linear Model. *Journal of Machine Learning Research*, 9:759–813, 2008.

[80] W. Sharpe. A Simplified Model for Portfolio Analysis. *Management Science*, 9(1):277–293, 1963.

[81] R. Shumway and D. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982.

[82] R. Shumway and D. Stoffer. *Time Series Analysis and its Applications with R Examples*. Springer, 2011.

[83] R. Tütüncü and M. Koenig. Robust asset allocation. *Annals of Operations Research*, 132:157–187, 2004.

[84] E. Treister and I. Yavneh. A Multilevel Iterated-Shrinkage Approach to ℓ_1 Penalized Least-Squares Minimization. *IEEE Trans. Signal Proc.*, 60(12):6319–6329, 2012.

[85] S. Wright. *Primal Dual Interior-Point Methods*. SIAM, 1997.

[86] C. Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[87] Y. Yen. *Three essays in financial econometrics*. PhD thesis, University of London, 2012.

[88] Y. Yen and T. Yen. Solving norm constrained portfolio optimization via coordinate-wise descent algorithms. *Computational Statistics and Data Analysis*, 76:737–759, 2014.

[89] W. Zangwill. *Nonlinear Programming: A Unified Approach*. Prentice-Hall, 1969.

[90] L. Zhang. Estimating covariation: Epps effect and microstructure noises. *Journal of Economics*, 160(1):33–47, 2011.

[91] M. Zhang, J. Russel, and R. Tsay. Determinants of bid and ask quotes and implications for the cost of trading. *Journal of Empirical Finance*, 15(4):656–678, Sept. 2008.

[92] S. Zhang, P. Yin, and J. Xin. Transformed schatten-1 iterative thresholding algorithms for matrix rank minimization and applications. *arXiv1506.04444 Prepr.*, pages 1–22, 2015.

[93] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[94] H. Zou and H. Zhang. On the Adaptive Elastic-Net with a Diverging Number of Parameters. *The Annals of Statistics*, 37(4):1733–1751, Aug 2009.

Appendix A

Appendix

A.1 Kalman Smoothing Equations

The Kalman smoother can be used to compute the posterior distribution of $X(t)$ given Y and an estimate of $\Theta = [D, \Gamma, \Sigma_o', J]$. From [81] the posterior distribution is normal and is completely characterized by the following quantities for $m = T$

$$\bar{X}(t|m) = \mathbb{E}(X(t)|y(1:m))$$

$$P(t|m) = cov(X(t), X(t)|y(1:m))$$

$$P(t, t-1|m) = cov(X(t), X(t-1)|y(1:m)).$$

These values can be computed efficiently using a set of well known forward and backward recursions [82] known as the Rauch-Tung-Striebel (RTS) smoother. The forward recursions are

$$\bar{X}(t|t-1) = \bar{X}(t-1|t-1) + D + J(t) \tag{A.1}$$

$$P(t|t-1) = P(t-1|t-1) + \Gamma \tag{A.2}$$

$$G(t) = P(t|t-1)I(t)^T \left(I(t)P(t|t-1)I(t)^T + \Sigma_o^2(t)\right)^{-1} \tag{A.3}$$

$$\bar{X}(t|t) = \bar{X}(t|t-1) + G(t)(y(t) - I(t)\bar{X}(t|t-1)) \tag{A.4}$$

$$P(t|t) = P(t|t-1) - G(t)I(t)P(t|t-1) \tag{A.5}$$

with $\bar{X}(0|0) = \mu$ and $P(0|0) = K$.

The backward equations are given by

$$\begin{aligned} H(t-1) &= P(t-1|t-1)P(t|t-1)^{-1} \\ \bar{X}(t-1|T) &= \bar{X}(t-1|t-1) + H(t-1)(\bar{X}(t|T) - \bar{X}(t|t-1)) \\ P(t-1|T) &= P(t-1|t-1) \\ &\quad + H(t-1)(P(t|T) - P(t|t-1))H(t-1)^T. \end{aligned}$$

A backward recursion for computing $P(t,t-1|T)$ is

$$\begin{aligned} P(t-1,t-2|T) &= P(t-1|t-1)H(t-2)^T \\ &\quad + H(t-1)\left(P(t,t-1|T) - P(t-1|t-1)\right)H(t-2)^T \end{aligned}$$

where

$$P(T,T-1|T) = (I - G(T)I(T))P(T-1|T-1). \tag{A.6}$$

A.2 Derivation of Equation (3.7)

Here we derive the expression for

$$\mathcal{G}(\theta, \Theta^{(j)}) = \mathbb{E}_{p(x|y,\Theta^{(j)})} \log p(X(1:T), y(1:T)|\theta) + \log(p(\theta))$$

given in equation (3.7). First recall the equation for the log-likelihood

$$\begin{aligned}
\log p(x, y|\tilde{\theta}) \quad = \quad & -0.5 \sum_{t=1}^{T} \log(|\Sigma_o(t)|) - \frac{1}{2} \sum_{t=1}^{T} \|y(t) - \tilde{I}(t)\bar{X}(t)\|^2_{\mathrm{diag}(\Sigma_o(t)^{-1}),\ell_2} \\
& - \frac{T-1}{2} \log(|\Gamma|) \\
& - \frac{1}{2} \sum_{t=2}^{T} r(t)^T \Gamma^{-1} r(t) \\
& + const \qquad \text{(A.7)}
\end{aligned}$$

where

$$r(t) = x(t) - x(t-1) - d - j(t).$$

First note that using the relation

$$Y(t) - \tilde{I}(t)X(t) = Y(t) - \tilde{I}(t)(X(t) - \bar{X}(t)) - \tilde{I}(t)\bar{X}(t)$$

we have that

$$\begin{aligned}
\mathbb{E}_{p(x|y,\Theta^{(j)})} \|(y(t) - \tilde{I}(t)X(t))\|^2_{\mathrm{diag}(\Sigma_o(t)^{-1}),\ell_2} \quad = \quad & \|y(t) - \tilde{I}(t)\bar{X}(t)\|^2_{\mathrm{diag}(\Sigma_o(t)^{-1}),\ell_2} + \\
& + \mathrm{tr}(P(t|T)\tilde{I}(t)^T \Sigma_o(t)^{-1} \tilde{I}(t)). \qquad \text{(A.8)}
\end{aligned}$$

Similarly noting that for $R(t) \doteq X(t) - X(t-1) - d - j(t)$

$$R(t) = X(t) - \bar{X}(t) - (X(t-1) - \bar{X}(t-1)) + (\bar{X}(t) - \bar{X}(t-1)) - d - j(t) \qquad \text{(A.9)}$$

we can show that

$$\sum_{t=2}^{T} \mathbb{E}_{p(x|y,\Theta^{(j)})} R(t)^T \Gamma^{-1} R(t) = \operatorname{tr}(\Gamma^{-1}(C - B - B^T + A)). \tag{A.10}$$

using the orthogonality principle. From equations (A.8) and (A.10) we arrive at (3.7).

A.3 Convergence of KECM Algorithms

Convergence of the EM and ECM algorithms in general is considered in [86] and [68] respectively. It is shown in [68] that the ECM algorithm converges to stationary point of the log posterior under the following mild regularity conditions

1. Any sequence $\Theta^{(k)}$ obtained using the ECM algorithm lies in a compact subset of the parameter space, Ω. For our case we need to restrict the parameter space such that $\sigma_o^2 \neq 0$ and Γ is positive definite.
2. $\mathcal{G}(\Theta, \Theta')$ is continuous in both Θ and Θ'.
3. The log posterior $L(\Theta)$ is continuous in Ω and differentiable in the interior of Ω.

A.3.1 Algorithm 1

Since the Laplace prior on J is not differentiable condition 3 is not satisfied and the results in [68] are not directly applicable. However the proofs and solution set in [68] can be modified to handle this irregularity.

Before addressing condition 3 we first verify condition 1. We start by examining the sequence of covariance estimates $\Gamma^{(k)}$.

Lemma A.1. *Assume a noisy asset price is observed at least one time for each asset for* $t > 1$ *and that* $\tilde{I}(t) \neq 0$ *for all* t*. Let* $\Gamma^{(k)}$ *be a sequence of solutions obtained with Algorithm 1, where* $\Gamma^{(0)}$ *is positive definite. Then sequences* $\Gamma^{(k)}$ *and* $\frac{1}{s^{(k)}}$ *are bounded where* $s^{(k)}$ *is the minimum eigenvalue of* $\Gamma^{(k)}$*. In addition the sequence* $\sigma_{o,i}^{2,(k)}$ *is bounded below and above by positive values for all* i*.*

Proof. Since W_o is positive definite we have from equation (3.9) that $s^{(k)}$ is bounded below by a positive constant which implies $\frac{1}{s^{(k)}}$ is bounded. Similarly by equation (3.10) we have $\sigma_{o,i}^{2,(k)}$ is bounded below by a positive constant. To prove that $\Gamma^{(k)}$ is bounded we note that the posterior may be written as

$$\begin{aligned} p(\theta|y) &= C_1 p(y|\theta)p(\theta) \\ &= C_1 p(y(1)|\theta)p(\theta)\prod_{t=2}^{T} p(y(t)|y(1:t-1),\theta) \\ &\leq C_2 p(y(1)|\theta)\prod_{t=2}^{T} p(y(t)|y(1:t-1),\theta) \end{aligned}$$

where C_1 is a constant not dependent on θ and where $C_2 = C_1 \sup_\theta p(\theta)$. Note that $C_2 < \infty$.

For $t > 1$ each of the conditional distributions $p(y(t)|y(1:t-1),\theta)$ is a normal distribution with covariance

$$Q(t) = \tilde{I}(t)P(t|t-1)\tilde{I}(t)^T + \sigma_o^2 I$$

where for notational simplicity we suppress the dependence of $Q(t)$ and $P(t|t-1)$ on k. Since $\sigma_{o,i}^2$ is bounded below by a positive value, it follows that $\frac{1}{|Q(t)|}$ is bounded.

Now suppose that $\Gamma^{(k)}$ is unbounded. Then since

$$P(t|t-1) = P(t-1|t-1) + \Gamma$$

$P(t|t-1)$ is unbounded as k goes to ∞. Since an observation of each asset's price occurs at least once for $t > 1$ it follows that $Q(\tau)$ is unbounded (as $k \to \infty$) for some $\tau > 1$. Then since the smallest eigenvalue of $Q(\tau)$ is bounded below by a positive constant, the determinant of $Q(\tau)$ is unbounded. Thus a subsequence of $p(y(\tau)|y(1:\tau-1), \Theta^{(k)})$ will approach 0. Since $\frac{1}{|Q(t)|}$ is bounded, $p(y(t)|y(1:t-1), \Theta^{(k)})$ will remain bounded above for all t. Then using (A.11) we have

$$\begin{aligned} p(\theta|y) &\leq C_2 p(y(1)|\theta) \prod_{t=2}^{T} p(y(t)|y(1:t-1), \theta) \\ &= C_2 p(y(\tau)|y(1:\tau-1), \theta) \prod_{t \neq \tau}^{T} p(y(t)|y(1:t-1), \theta) \end{aligned}$$

which implies a subsequence of $p(\Theta^{(k)}|y)$ will converge to 0. This contradicts the monotonicity of the ECM algorithm [68]. The proof that the sequence $\sigma_{o,i}^{2,(k)}$ is bounded above for all i is similar. □

Lemma A.2. *Assume the conditions of Lemma A.1. Let $\lambda(t)^{-1,(k)}$ be a sequence of solutions obtained with Algorithm 1 where $\Gamma^{(0)}$ is positive definite. Then there exist finite positive numbers a, b where $a \leq \lambda_i(t)^{(k)} \leq b$ for all t, k and i.*

Proof. By the update equation (3.14) we may set $b = \frac{\alpha_\lambda + 2}{\beta_\lambda}$ which is positive and finite. By way of contradiction assume the lower bound does not hold. Then for some i and t there exists a subsequence $\lambda_i(t)^{(k_n)}$ such that $\lim_{n\to\infty} \lambda_i(t)^{-1,(k_n)} = \infty$. Since each $\lambda_i(t)^{-1}$ is the mode of an inverse gamma distribution it follows that the posterior scale parameter,$(\beta_\lambda + |j^{(k_n)}|)$ goes to infinity . This implies that $p(\lambda_i(t)^{-1,(k_n)}, j_i(t)^{(k_n)}) \to 0$. Since each prior density function is bounded as $\lambda_i(t) \to 0$ this implies that $p(\theta)$ goes to zero, contradicting the monotonicity of the ECM algorithm. Thus there exists an $a > 0$ such that $\lambda_i(t)^{(k)} > a$ for all t, k and i. □

Now we prove that the sequences $J^{(k)}$ and $D^{(k)}$ are also well behaved.

Lemma A.3. *Assume the conditions of Lemma A.1. Let $J^{(k)}$ and $D^{(k)}$ be sequences of solutions obtained with Algorithm 1 where $\Gamma^{(0)}$ is positive definite. Then sequences $J^{(k)}$ and*

$D^{(k)}$ *are bounded.*

Proof. From Lemma A.1 the likelihood $p(y|\theta)$ is bounded above. Recall from the previous lemma that there exists an $a > 0$ such that for all k, $\lambda_i(t)^{(k)} \geq a$. Since the prior density function is bounded above for each parameter it follows that $\lim_{j\to\infty} p(\theta) = 0$. This implies $J^{(k)}$ is bounded by the monotonicity of the ECM algorithm. Since $\lim_{d\to\infty} p(\theta) = 0$ it also follows that $D^{(k)}$ is bounded. $\square$

The above lemmas imply the following corollary.

Corollary 1. *The sequence $\Theta^{(k)}$ is bounded and all limit points are feasible (e.g. variance non-zero, positive definite covariance).*

Now we derive some additional properties of the limit points of $\Theta^{(k)}$. To do this we shall refer to Zangwill's convergence theorem [89]. To use Zangwill's theorem, we first define $\mathcal{A}$ to be a point to set mapping defined by the ECM algorithm i.e. $\Theta^{(k+1)} \in \mathcal{A}(\Theta^{(k)})$. Let us define a solution set, $\mathcal{S}$, as the set of θ such that

$$\theta_1 = \arg\max_v \mathcal{G}\left(\left[v, \theta_2, \theta_3, \theta_4, \theta_5\right], \theta\right)$$

$$\theta_2 = \arg\max_v \mathcal{G}\left(\left[\theta_1, v, \theta_3, \theta_4, \theta_5\right], \theta\right)$$

$$\theta_3 = \arg\max_v \mathcal{G}\left(\left[\theta_1, \theta_2, v, \theta_4, \theta_5\right], \theta\right)$$

$$\theta_4 = \arg\max_v \mathcal{G}\left(\left[\theta_1, \theta_2, \theta_3, v, \theta_5\right], \theta\right)$$

$$\theta_5 = \arg\max_v \mathcal{G}\left(\left[\theta_1, \theta_2, \theta_3, \theta_4, v\right], \theta\right).$$

By definition $\theta \in \mathcal{A}(\theta)$ for all $\theta \in \mathcal{S}$. This along with the monotonicity of the ECM algorithm implies that $L(\theta)$ is an ascent function, i.e.

$$L(\theta') > L(\theta) \text{ for all } \theta \notin \mathcal{S}, \theta' \in \mathcal{A}(\theta)$$
$$L(\theta') \geq L(\theta) \text{ for all } \theta \in \mathcal{S}, \theta' \in \mathcal{A}(\theta).$$

Since $\mathcal{G}(\theta, \theta')$ is continuous in both θ and θ' we have that $\mathcal{A}$ is a closed mapping. Thus we have the following theorem.

Theorem A.1. *All limit points of $\Theta^{(k)}$ belong to $\mathcal{S}$.*

Proof. This is a direct consequence of Zangwill's convergence theorem [89] (also known as the Global convergence theorem [64]). To invoke the theorem we must meet the following conditions

- $\Theta^{(k)}$ belongs to a compact subset of the feasible solutions
- $\mathcal{A}$ is closed
- There exists a continuous ascent function

All three of these conditions were shown above, thus the theorem follows from Zangwill's convergence theorem. □

Now we show that if $\theta' \in \mathcal{S}$ then θ' is in some sense a "stationary" point of the log posterior $L(\theta) = \log p(\theta|y)$.

Theorem A.2. *Let $\theta' \in \mathcal{S}$. Then*

$$\nabla_{\theta_i} L(\theta)_{|\theta=\theta'} = 0 \text{ for } i \in 1, 2, 3, 5$$

and

$$0 \in \partial_{\theta_4} L(\theta)_{|\theta=\theta'}.$$

Proof. To show this we first note that $L(\theta)$ can be written as [68]

$$L(\theta|y) = \mathcal{G}(\theta,\theta') - H(\theta,\theta')$$

where

$$H(\theta,\theta') = \mathbb{E}_{p(x|y,\theta')} \log p(X|y,\theta).$$

From the information inequality we have that $H(\theta',\theta') \geq H(\theta,\theta')$ for all feasible θ. Since $H(\theta,\theta')$ is differentiable with respect to θ it follows that

$$\nabla_\theta H(\theta,\theta')_{|\theta=\theta'} = 0.$$

Since $\nabla_{\theta_i}\mathcal{G}(\theta,\theta')_{|\theta=\theta'} = 0 \;\text{ for }\; i \in 1,2,3,5$ it follows that

$$\nabla_{\theta_i} L(\theta)_{|\theta=\theta'} = 0 \text{ for } i \in 1,2,3$$

Also since $\mathcal{G}(\theta,\theta')$ and $H(\theta,\theta')$ are convex in j, and $\theta' \in \mathcal{S}$, it follows that

$$0 \in \partial_{\theta_4}\mathcal{G}(\theta,\theta')$$

which implies

$$0 \in \partial_{\theta_4} L(\theta,\theta').$$

□

A.3.2 Algorithm 3

Analogous results to Corollary 1 and Theorem A.1 may proven for Algorithm 3 using same arguments as Algorithm 1. The following result is analogous to Theorem A.2.

Theorem A.3. *Let $\theta' \in \mathcal{S}$ where $\mathcal{S}$ is the set of fixed points of the Algorithm 3. Then*

$$\nabla_{\theta_i} L(\theta)_{|\theta=\theta'} = 0 \text{ for } i \in 1,2,3,5,6.$$

The proof of this result is the same as Theorem A.2.

A.4 MCMC Details

In this section we state the conditional distributions needed to implement the Gibbs sampling approach in Section 3.3.1.

A.4.1 Conditional Price Distribution

Let $\mathcal{N}(x,\mu,R)$ be the normal PDF in x with mean μ and covariance R. For the Gibbs sampling approach we need to determine the conditional distribution of $X(t)$ given Φ_{-1},Γ,Y, and $X(s) s \neq t$. Let $Y_{tot}(t)$ be the total price vector obtained from observed prices $Y(t)$ and the current sample of the unobserved prices $Y_{miss}(t)$. We first note that for $t>1, t<T$

$$\begin{aligned}
p(x(t)|x(s),\phi_{-1},\gamma,y_{tot};\forall s\neq t) &= p(x(t)|x(t-1),x(t+1),\phi_{-1,-0},\gamma,y_{tot}(t)) \\
&\propto p(x(t+1)|x(t),\phi_{-1,-0},\gamma) \\
&\quad\; p(y_{tot}(t)|x(t),\phi_{-1,-0},\gamma) \\
&\quad\; p(x(t)|x(t-1),\phi_{-1,-0},\gamma).
\end{aligned}$$

By properties of normal distributions

$$p(x(t+1)|x(t),\phi_{-1,-0},\gamma) = \mathcal{N}(x(t+1),x(t)+j(t+1)+d,\Gamma) \tag{A.11}$$

and

$$p(y_{tot}(t)|x(t), \phi_{-1}, \gamma) = \mathcal{N}(y_{tot}(t), x(t), \sigma_o^2 I). \tag{A.12}$$

With this recall the following multiplication property of normal PDFs

$$\mathcal{N}(x, \mu_1, R_1)\mathcal{N}(x, \mu_2, R_2) \propto \mathcal{N}(x, \mu_3, R_3)$$

where

$$R_3 = (R_1^{-1} + R_2^{-1})^{-1}$$

and

$$\mu_3 = R_3 R_1^{-1} \mu_1 + R_3 R_2^{-1} \mu_2.$$

Using the multiplication property above

$$p(x(t)|x(t-1), y_{tot}(t)) = \mathcal{N}(x(t), q, Q)$$

where

$$Q = (\Gamma^{-1} + \sigma_j^{-2} I)^{-1}$$

$$q = Q\Gamma^{-1}(x(t-1) + J(t) + D) + \sigma_j^{-2} Q y_{tot}(t).$$

Applying the multiplication property again gives

$$p(x(t)|x(t-1), y_{tot}(t), x(t+1)) = \mathcal{N}(x(t), q', Q')$$

where

$$Q' = (\Gamma^{-1} + Q^{-1})^{-1}$$

and

$$q' = Q'Q^{-1}q + Q'\Gamma^{-1}(x(t+1) - D - J(t+1).$$

The conditional distributions for $t = 1$ and $t = T$ can be derived similarly.

Another approach to sampling from the conditional distribution $p(X|Y, \Phi)$ is the Forward Filtering Backward Simulation (FFBS) approach [38]. The FFBS algorithm allows one to sample directly from the conditional joint distribution of $X(1:T)$, but the required backward simulation can be computationally intensive as one must compute T Cholesky decompositions. In the approach outlined above one only needs to compute 3 Cholesky decompositions ($t = 1$,$t = T$ and once for $1 < t < T$).

A.4.2 Conditional Jump Distribution

Let $\mathcal{N}(x, \mu, \tau)$ be the normal PDF in x with mean μ and variance τ. Recall that the prior distribution of the jumps is the spike and slab prior

$$p(j) = f(j) = \zeta\delta_0(j) + (1-\zeta)\mathcal{N}(j, 0, \sigma_j^2)$$

and that in the prior distribution the jumps are independent and identically distributed. When conditioned on σ_j^2, ζ, X, D and Γ, $J_m(t)$ and $J_n(s)$ remain independent for $s \neq t$ but $J_m(t)$ and $J_n(t)$ become dependent. The conditional distribution $p(j(t)|\phi_{-2}, \gamma)$ can be written as

$$p(j(t)|\phi_{-2}, \gamma) = \frac{c\exp\left(\frac{-(j(t)-v(t))^T\Gamma^{-1}(j(t)-v(t))}{2}\right)}{\sqrt{(2\pi)^N|\Gamma|}} \prod_{i=1}^{N} f(j_i(t)) \tag{A.13}$$

where $v(t) = X(t) - X(t-1) - D$ and where $c > 0$ is independent of $j(t)$. Sampling directly from this distribution is difficult due to the combinatorial nature of the prior. Therefore we sample sequentially each component of $j_i(t)$ conditioned on $j_{-i}(t)$.

To derive the posterior distribution of $j_i(t)$ we note that from properties of the multivariate normal distribution

$$p(j_i(t)|j_{-i}(t), \gamma, \phi_{-2}) \propto \mathcal{N}(j_i(t), a(i), b^2(i)) f(j_i(t))$$

where

$$a(i) = v_i(t) + \Gamma_{i,-i}\Gamma^{-1}_{-i,-i}(j_{-i}(t) - v_{-i}(t))$$

and

$$b^2(i) = \Gamma_{i,i} - \Gamma_{i,-i}\Gamma^{-1}_{-i,-i}\Gamma_{-i,i}.$$

Next we determine $Pr(Z_i(t) = z|\phi_{-2}, \gamma, j_{-i}(t))$ for $z = 0, 1$. Recall the following identity for normal PDFs

$$\mathcal{N}(x, u_1, \tau_1^2)\mathcal{N}(x, u_2, \tau_1^2) = \mathcal{N}(u_1, u_2, \tau_1^2 + \tau_2^2)\mathcal{N}(x, u, \tau^2)$$

$$u = \frac{\tau_1^{-2}u_1 + \tau_2^{-2}u_2}{\tau_1^{-2} + \tau_2^{-2}}$$

$$\tau^2 = \frac{\tau_1^2\tau_2^2}{\tau_1^2 + \tau_2^2}.$$

Using the relationship above we have

$$Pr(Z_i(t) = z|\phi_{-2}, \gamma, j_{-i}(t)) \propto \begin{cases} \zeta\mathcal{N}(0, a(i), b^2(i)) \text{ if } z = 0 \\ (1-\zeta)\mathcal{N}(0, a(i), b(i) + \sigma_j^2) \text{ if } z = 1 \end{cases} \tag{A.14}$$

We now draw $Z_i(t)$ from this distribution. If $Z_i(t) = 0$, $J_i(t)$ is set to zero, otherwise we draw $J_i(t)$ from the distribution

$$p(j_i(t)|z_i(t) = 1, \phi_{-2}, \gamma, j_{-i}(t)) \tag{A.15}$$

which from the above relationship is a normal distribution with mean

$$\frac{a(i)}{1 + b^2(i)\sigma_j^{-2}}$$

and variance

$$\frac{b^2(i)\sigma_j^2}{b^2(i)+\sigma_j^2}.$$

Conditional Posterior Mode of j_i in KECM spike and slab model

Note that the conditional maximization steps for J used in KECM algorithm for spike and slab models can be derived in a similar manner as above. To see this note that (3.15) is up to a constant the logarithm of (A.13) where v is replaced with Δ. Thus one can compute the modes of Z_i and J_i from the conditional distributions defined above in (A.14) and (A.15).

A.4.3 Other conditional distributions

The remaining conditional distributions for the other parameters are easily obtained due to conjugate prior relationships [37].