The Dissertation Committee for David Walker Puelz
certifies that this is the approved version of the following dissertation:

Regularization in Econometrics and Finance

Committee:

Carlos Carvalho, Supervisor

P. Richard Hahn

James Scott

Sheridan Titman

Sinead Williamson

Regularization in Econometrics and Finance

by

David Walker Puelz

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2018

Dedicated to my wife Hanna.

Acknowledgments

First, thank you to my advisors. My experience at the University of Texas was enriched by all of you. Carlos and James – you all have provided me tremendous mentorship. I hope to pass it along to my future students. Richard, thank you for always being honest and skeptical as well as a great mentor. All of my work was made better through our collaboration. Sheridan, I appreciate your perspective in all of my finance related projects. Sinead, I thoroughly enjoyed our engaging conversations about utility-based selection.

Second, thank you to my Ph.D. colleagues. Long, you were with me since the beginning, and I value our friendship just as much as our deep conversations about research. Jared, your optimism is contagious and important for those days when the code doesn't work and the math doesn't make sense.

Finally, thank you to my wife Hanna, my parents Amy and Bob, and my twin brother Charles. There were many days of joy and progress but also days of frustration. I could not have completed this degree without your love and support through all of it.

Regularization in Econometrics and Finance

David Walker Puelz
The University of Texas at Austin, 2018

Supervisor: Carlos Carvalho

This dissertation develops regularization methods for use in finance and econometrics problems. The key methodology introduced is utility-based selection (UBS) – a procedure for inducing sparsity in statistical models and practical problems requiring the need for simple and parsimonious decisions.

The introduction section describes statistical model selection in light of the "big data hype" and desire to fit rich and complex models. Key emphasis is placed on the fundamental bias-variance tradeoff in statistics. The remaining portions of the introduction tie these notions into the components and procedure of UBS. This latter half frames model selection as a decision and develops the procedure using decision-theoretic principles.

The second chapter applies UBS to portfolio optimization. A dynamic portfolio construction framework is presented, and the asset returns are modeled using a Bayesian dynamic linear model. The focus here is constructing simple, or sparse, portfolios of passive funds. We consider a set of the most liquid exchange traded funds for our empirical analysis.

The third chapter discusses variable selection in seemingly unrelated regression models (SURs). UBS is applied in this context where an analyst wants to find, among p available predictors, what subset are most relevant for describing variation in q different responses. The selection procedure takes into account uncertainty in both the responses and predictors. It is applied to a popular problem in asset pricing – discovering which factors (predictors) are relevant for pricing the cross section of asset returns (responses). We also discuss future work in monotonic function estimation and how UBS is applied in this context.

The fourth chapter considers regularization in treatment effect estimation using linear regression. It introduces "regularization-induced confounding" (RIC), a pitfall of employing naive regularization techniques for estimating a treatment effect from observational data. A new model parameterization is presented that mitigates RIC. Additionally, we discuss recent work that considers uncertainty characterization when model errors may vary by clusters of data. These developments employ empirical-Bayes and bootstrapping techniques.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

You can either have maximum explainability or maximum predictability.

Paraphrase of Leo Breiman

Information is accumulating rapidly and at an increasing rate. In tandem, statistical models are gaining complexity as theory advances and computational power becomes cheaper. The ability to store more data and fit rich models will only get easier with time, and statistical methods need to keep pace with this progress. The developments advocated for in this thesis cut across the obvious need for statistical model complexity and richness. Without a doubt, there is an important place for advanced and nuanced modeling techniques to describe complex data, and we have seen an explosion of these methods from deep learning and ensemble methods in statistical learning to hierarchical modeling in Bayesian analysis. However, there exists a simultaneous need for *simplicity* and *parsimony* in statistical inference to make big data and complex models digestible for decision makers. Developing novel approaches for parsimonious modeling while taking statistical uncertainty of all forms into account will be the goal of this dissertation.

Beyond the call for simplicity in the age of "big data," dealing with fewer variables or predictors in a data set or a manifestly simpler model is intuitively valuable. Humans work well when thinking about a small number of objects. Information can be understood and processed at a deeper level, and higher quality decisions based on a small set of variables or objects can be made. In this data reduction setting, bigger is actually not better! A key feature of this dissertation is the focus on applied problems in finance and econometrics that require some notion of simplicity in order to be solved. Analysis of three questions comprise the subsequent chapters.

1. Among many possible funds for investment, which are the best to hold in a portfolio and in what proportions?

2. Which tradable factors represent fundamental dimensions of the financial market?

3. Can we harness the power of regularization in a high dimensional causal inference setting? If so, how does it affect estimation and how can we alleviate bias?

Each of the three questions spanning finance, asset pricing, and econometrics shares the common theme of requiring some notion of parsimony or selection of a subset of objects from a potentially massive set objects. In investing, the objects are investable assets or funds. In asset pricing, the objects are tradable factors proposed by theoreticians in finance that are long-short

portfolios of stocks. In causal inference with observational data, the objects are measured covariates (of which there may be many) that represent characteristics of observations that may need to be controlled for in a regression.

Achieving parsimony in modeling is viewed through the lens of statistical regularization. In its most simple form, regularization is a way to navigate the bias-variance tradeoff of an estimator, or equivalently, a tradeoff between estimator complexity and predictive accuracy. An estimator is any rule or function used to estimate a quantity of interest based on observed data. As a simple example, we can think of an estimator as describing how a response variable Y is generated probabilistically given a set of observed covariates X. Examples of estimators in this setting may be a general nonparametric function f that maps X to the conditional expectation of the response $\mathbb{E}[Y \mid \mathrm{X}] = f(\mathrm{X})$ or the coefficients β in a linear regression which parameterizes the conditional expectation as a linear function of the covariates: $\mathbb{E}[Y \mid \mathrm{X}] = \mathrm{X}\beta$. In both nonparametric and parametric modeling scenarios the notion is the same – a complex estimator will be difficult to interpret by a data analyst and may not predict well as new data is brought into the fold; a problem known as overfitting. A simple estimator may be easily digestible for a data analyst but will similarly fail to predict well "out-of-sample" because it is too biased toward simple predictions. Statistical regularization helps build an estimator that interpolates between these two extremes of complexity. The *best* estimator will often be located somewhere in the middle of the simplest and the most complex.

The quote from Leo Breiman – *You can either have maximum explainability or maximum predictability* – narrowly describes bias' relationship to predictive accuracy. What if we abstracted away from bias, variance, and its relation to predictive accuracy? Are there models where bias improves interpretability and little is lost in how well the model explains future data? How can we formally study that relationship? This thesis offers answers to these questions by developing a new framework for model comparison. We are interested in scrutinizing bias' influence on *general* and *customizable* functions of future data where uncertainty naturally arises from the predictive distribution.

1.1 Preliminaries

The following section presents an overview of foundational ideas pertinent to the work in this dissertation. The first part discusses model selection and the bias-variance tradeoff, the second comments on generalizing these notions and incorporating decision theory into model selection, and the third outlines a general "utility-based" selection procedure applied in the second and third chapters.

1.1.1 The bias-variance decomposition

The tradeoff between the bias and variance of a statistical estimator can be most clearly seen by considering the formula for the mean squared prediction error (MSE). Suppose our task is to estimate a function $f(x)$ evaluated

at a point x. Define $\hat{f}(x)$ to be our noisy estimate. The MSE is:

$$\text{MSE}(\hat{f}, f) = \mathbb{E}\left[\left(f(x) - \hat{f}(x)\right)^2\right].$$

Routine calculation shows that it may be written as:

$$\text{MSE}(\hat{f}, f) = b^2 + v,$$

where $b = \mathbb{E}[\hat{f}(x)] - f(x)$ and $v = \text{var}[\hat{f}(x)]$, hence the breakdown of a squared "bias" term given by b^2 and "variance" term given by v.

We can see this in action by considering a data set of 500 (x, y) pairs. These are shown by the gray dots plotted in both figures in (1.1). Our task is to estimate a function f that will describe the relationship between x and y and will hopefully work well at describing the relationship as new data is observed. An estimate of y at a particular x may be written at $\hat{y} = \hat{f}(x)$, where $\hat{f}$ is the function estimate we aim to infer.

Figure (1.1) shows two extremes of potential estimates for f in black. The left estimate is the most "simple" – we estimate all y's as exactly the same regardless of the value of x. The estimate for f in this case is taken to be a constant function that is the mean of y over all observed x values. The right is a more flexible and thus more "complicated" estimate for f. Instead of taking one grand average over all observed x's, this estimate takes a local average over the five nearest data points to x. Specifically, the estimate is $\hat{f}(x) = \sum_{j \in \mathcal{N}} y_j/5$ where $\mathcal{N}$ are the collection of five observations that are nearest to x in the domain.

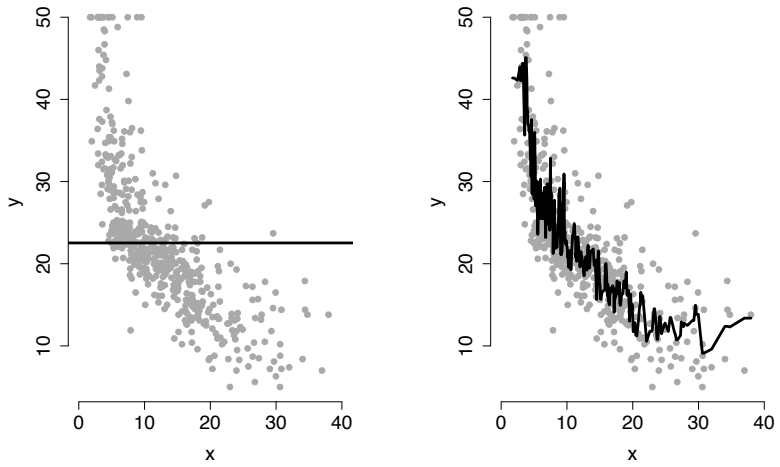


Figure 1.1: The left graph shows the estimated $\hat{f}(x)$ as a constant function shown in black, and the right graph shows an estimated $\hat{f}(x)$ by taking an average of the five nearest data points around x, also shown in black.

The tradeoff between simplicity and complexity is clear. The constant function is simple and easily digestible – it says to predict everything as a single number! The locally averaged function based on nearest neighbor data is more complex. It relies on x information in a surrounding neighborhood of the function evaluation. One can also see "bias" and "variance" in these two estimates. The constant function is severely biased, and its predictions have zero variance by definition. Conversely, the observable "wiggles" in the locally averaged function correspond to variance in its predictions. Since it also uses local information, its predictions are less biased.

Figure (1.2) expresses this tradeoff by evaluating the mean squared prediction error for a variety of models spanning from the constant model (far

left) to locally averaged models whose neighborhood shrinks to include only a couple of data points (far right). We consider a local averaging procedure called "k nearest neighbors." The constant function corresponds to $k = 500$ (averaging over all data points for each prediction). Progressively complex, higher variance estimates correspond to smaller values of k. Complexity increases from the left to right in the figure where smaller k correspond to higher variance estimates. The "u-shape" of this figure demonstrates the empirical tradeoff between bias and variance. High bias, low variance estimates (left) and low bias, high variance estimates (right) will perform poorly at prediction – the sweet spot where the MSE is minimized exists somewhere in the middle.

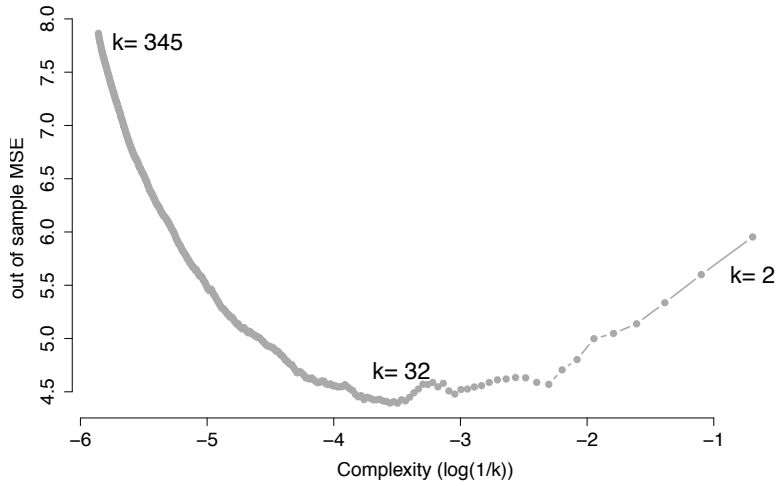


Figure 1.2: Mean squared error (calculated on testing data) versus model complexity. In the nearest neighbor sense, model complexity is related to the inverse of the number of data points we choose to include in the local average.

1.1.2 Beyond the bias-variance decomposition and toward model selection and decision theory

The tradeoff presented in the previous section is a foundational starting point for the work in this dissertation. It relates model complexity to a clear objective of the statistical analyst: Prediction. Over the past several decades, there have been many related selection approaches in Bayesian and classical statistics (shrinkage priors as a part of model specification, stochastic search algorithms, forward and backward stepwise selection, penalized likelihood methods, etc.) that extend the notion of selection based on other objectives; whether it be a likelihood-based fit (AIC and BIC), or maximizing a penalized objective function (the "LASSO" as described in Tibshirani [1996]). Which selection approach is best? What if I am uncertain about the observed variables in my data set, how they were collected, or how accurately they represent the true covariates? Is there a way to generalize these selection procedures while taking all forms of statistical uncertainty, including parameter and predictor uncertainty, into account? These questions will guide the methodological developments to follow.

This dissertation will frame model selection as a decision process and apply it to problems in finance and econometrics. Since the analyst is choosing among many possible models of ranging complexity and performance based on their objective – we view this as a *decision making procedure* that should be formalized. The originating question for the ideas presented in this dissertation is:

Can we design a decision-theoretic framework for model selection where the objective is a utility function?

The utility function is the first important piece of our framework. It is a function defined over the choice set of models *and* future data, and using it, we cast the original tradeoff as one between complexity and *utility*. Since future data is an input of the utility function – it is itself a random variable! The complexity-utility tradeoff can then be viewed in the backdrop of uncertainty in future data and model parameters. This is where the second key feature, statistical uncertainty characterization, comes into play. We will rely on Bayesian methods of inference for uncertainty characterization. This step involves learning about the joint distribution of the future data $\tilde{Y}$ and model parameters Θ conditioned on past data $\mathbf{Y}$:

$$p(\tilde{Y}, \Theta \mid \mathbf{Y}).$$

In a traditional Bayesian setting, inference and model selection are intertwined; with carefully designed shrinkage priors, inference is not viewed as a "step" in a selection procedure but rather *the procedure*. Bayesians focus on understanding the parameter margin of this distribution: $p(\Theta \mid \mathbf{Y})$, known as the posterior distribution and proportional to the data likelihood and prior distribution: $p(\Theta \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \Theta)p(\Theta)$. A selection procedure will infer the posterior distribution of the parameters under shrinkage priors $p(\Theta)$ that bias estimates of Θ towards interpretable quantities (such as the zero vector if Θ is a regression coefficient). Since the parameters Θ serve in defining a map between

variables and responses in a statistical model, inferring sparsified estimates of Θ tell the analyst which variables may be important to which responses; see Liang et al. [2008a] and Bayarri et al. [2012a] for recent surveys of the field. This is where a Bayesian selection procedure concludes.

In contrast, we attempt to disentangle inference and selection. The decision-theoretic methodology presented here utilizes Bayesian inference as a means for uncertainty characterization, but it is only one piece of the selection methodology. The utility function assumes the role of encoding the analyst's desire for sparsity, and it is used in tandem with uncertainty characterization to provide a complete analysis of the model space. Hahn and Carvalho [2015] present these ideas in the context of linear regression models. Their paper introduces a *decoupling* of shrinkage (uncertainty characterization) and selection (utility specification and optimization) in regression models, but nonetheless stresses that both components are necessary for a selection procedure. In the following section, we will describe our precise procedure which is closely related; either called "utility-based selection" or "utility-based posterior summarization."

1.1.3 Utility-based selection

The outline of the procedure is presented in this section. Let d be a model decision (often a vector or matrix and synonymous with our description of an estimator), Θ be a vector of model parameters, λ be a complexity parameter governing the degree to which complexity is present in the utility

function, and $\tilde{Y}$ be future data. The components of the procedure are:

1. Loss function $\mathcal{L}(d, \tilde{Y})$ – measures utility of decision d.
2. Complexity function $\Phi(\lambda, d)$ – measures sparsity of decision d.
3. Statistical model $\tilde{Y} \sim \Pi(\Theta)$ – characterizes uncertainty.

Since we have a utility and probability measure, we are able to undertake an expected utility exercise. Specifically, we optimize the expected utility (calculated by integrating over the probability measure) for a range of penalty parameters λ. We then view these optimal decisions in light of the statistical uncertainty from the joint predictive distribution of future data and parameters $(\tilde{Y}, \Theta)$. The steps of the procedure spelled out are:

- Optimize $\mathbb{E}\left[\mathcal{L}(d, \tilde{Y}) + \Phi(\lambda, d)\right]$, where the expectation is taken over $p(\tilde{Y}, \Theta \mid \mathbf{Y})$.
- Extract optimal decisions d^*_λ from the expected loss minimization, and calculate loss at each decision $\mathcal{L}(d^*_\lambda, \tilde{Y})$.
- Analyze the loss *distributions* $\mathcal{L}(d^*_\lambda, \tilde{Y})$ for use in the final model selection.

The first step is the familiar decision-theoretic approach – optimize expected loss. The complexity enters into the objective additively, and its role is to encourage optimal decisions d that are sparse, or simple. In this way, it can be

thought of as a penalty function whose influence in the objective is controlled by the scalar λ. For example, if $\Phi(\lambda, d) = \lambda \|d\|_1$, a very large λ would result in a d^*_λ that is sparse because the penalty becomes the "dominate component" in the objective function. Conversely, a small λ would result in a dense d^*_λ because the penalty vanishes from the objective. This method may also be called utility-based posterior summarization because the set of optimal sparse decisions $\{d^*_\lambda\}$ can be thought of as point summaries of a potentially complex posterior distribution. The "summarization mechanism" is the utility function, and the posterior is filtered through the preferences set out by the statistical analyst during the integration.

The second and third steps revolve around a key quantity – the loss evaluated at optimal decision d^*_λ given by $\mathcal{L}(d^*_\lambda, \tilde{Y})$. Crucially, this loss is a random variable whose uncertainty is induced by the distribution of future data $\tilde{Y}$! The posterior reappears here by defining the posterior predictive distribution

$$p(\tilde{Y} \mid \mathbf{Y}) = \int p(\tilde{Y} \mid \Theta, \mathbf{Y}) p(\Theta \mid \mathbf{Y}) d\Theta.$$

Examination of the random variables $\{\mathcal{L}(d^*_\lambda, \tilde{Y})\}$ for a range of λ's constitutes the final step of the procedure. While d^*_λ provides the optimal point summary, the loss reintroduces uncertainty in the proper way by considering the analyst's utility function.

There is flexibility in this final step. One approach we use to analyze the loss random variables is to consider benchmarking versus a target decision:

d^*_{target}. In this case, we consider a new random variable ρ that is defined as:

$$\rho(d^*_\lambda, d^*_{\text{target}}, \tilde{Y}) = \mathcal{L}(d^*_\lambda, \tilde{Y}) - \mathcal{L}(d^*_{\text{target}}, \tilde{Y}).$$

This is the difference in loss between the sparse and target decisions. By definition, a large ρ corresponds to a large sparse loss relative to the target loss; so we refer to ρ as the "regret random variable." Interestingly, its distribution may be quite different from the loss distributions on their own. With ρ in hand, we can, for example, integrate to find the amount of probability mass below zero. This value represents the chance that regret is less than zero, i.e.: The chance that we are *satisfied* with the sparse decision relative to the target decision. Thresholding this quantity provides a natural way to select a sparse decision from all possibilities indexed by λ. The target decision may be chosen as an unattainable "dense" decision that includes all variables or any other decision the analyst wants to benchmark against. We consider several target decisions in the applied problems presented below.

1.1.4 A final comment on bias and prediction

This document frequently uses the terms *bias* and *prediction*. Both are venerable concepts in statistics. In a frequentist context, bias has a negative connotation. If an estimator's sampling distribution is not centered on the true value, this is perceived as undesirable. Utility-based selection not only produces biased estimators, but they are often inconsistent as well. As we will argue in the chapters to follow, this thesis strongly differentiates itself from

the negative aspects of bias and inconsistency. Bias is embraced for its ability to encourage simplicity in estimators since model interpretability is a key goal.

Secondly, although utility-based selection is packaged as distinct from methods solely based on predictive ability (like the MSE), **it does not render prediction irrelevant.** In fact, **prediction is central to the methodology**. Without careful construction of the predictive distribution, we would be unable to understand the predictive uncertainty in the sparse estimators (decisions) we are choosing between. Prediction is not bypassed in our selection procedure. Instead, it is a central component through careful characterization of $p(\tilde{Y} \mid \mathbf{Y})$.

Lastly, machine learning researchers may initially be underwhelmed with the choice of models in our applications because they are more structured than methods involving deep learning, manifold learning, and nonlinear dimensionality reduction. However, the structure we impose is imperative to answer our applied questions in econometrics and finance. We will see that "low signal high noise" is the default environment in these problems. Thus, structured models are necessary and may even outperform more complex nonlinear techniques in prediction. Although beyond the scope of this thesis, future work will involve considering more complex ML techniques and embedding them within our new utility-based methodology.

The remaining section discusses the layout of the dissertation.

1.2 Structure of the thesis

This dissertation is a compilation of work during my time at the University of Texas. The second and third chapters apply utility-based selection procedure to two problems in finance and econometrics and represent the bulk of the dissertation. The fourth chapter discusses statistical regularization in treatment effect estimation and future work. Literature reviews related to the problem at hand are presented in each chapter.

The content of this thesis closely follows several papers:

- David Puelz, P. Richard Hahn, and Carlos M. Carvalho. Regret-based selection for sparse dynamic portfolios. 2018.
- David Puelz, P. Richard Hahn, and Carlos M. Carvalho. Variable selection in seemingly unrelated regressions with random predictors. *Bayesian Analysis*, 12(4):969–989, 2017.
- P. Richard Hahn, Carlos M. Carvalho, David Puelz, and Jingyu He. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 13(1):163–182, 2018a.
- David Puelz, P. Richard Hahn, and Carlos M. Carvalho. Optimal ETF selection for passive investing. 2016.
- Carlos M. Carvalho, Jared D. Fisher, and David Puelz. Monotonic effects of characteristics on returns. *preprint*, 2018.

We clearly indicate where content from these papers appear in the following chapters.

Chapter 2

Sparse Portfolio Construction

Analysis and text in this chapter closely follows Puelz et al. [2016] and Puelz et al. [2018].

2.1 Introduction

Practical investing requires balancing portfolio optimality and simplicity. In other words, investors desire well-performing portfolios that are *easy* to manage, and this preference is driven by many factors. Managing large asset positions and transacting frequently is expensive and time-consuming, and these complications arise from both trading costs and number of assets available for investment. For the individual investor, these challenges are strikingly amplified. Their choice set for investment opportunities is massive and includes exchange-traded funds (ETFs), mutual funds, and thousands of individual stocks. This raises the question: *How does one invest optimally while keeping the simplicity (sparsity) of a portfolio in mind?* Further challenges arise when sparse portfolio selection is placed in a dynamic setting. The investor will want to update her portfolio over time as future asset returns are realized while maintaining her desire for simplicity.

The focus of this chapter are loss functions that balance utility and sparsity myopically for each time t:

$$\mathbf{L}_{\lambda_t}(w_t, \tilde{R}_t) = \mathcal{L}(w_t, \tilde{R}_t) + \Phi(\lambda_t, w_t), \tag{2.1}$$

where $\tilde{R}_t$ is a vector of future asset returns, $\mathcal{L}$ is the negative utility of an investor, w_t is a vector of portfolio weights, Φ is a function that encourages sparsity in w_t, and λ_t is a penalty parameter (or more generally, a penalty vector of equal length as w_t) governing the degree to which the complexity function Φ is influential in the overall loss function. Special attention must be paid to λ_t, the parameter that governs the utility-sparsity tradeoff. If it is known a priori, the investor's optimal portfolio may be found for each time by routine minimization of the expectation of (2.1). By contrast, this chapter considers the more challenging case where this parameter is unknown and may be thought of as part of the investor's decision.

The interplay between dynamics, utility and portfolio simplicity in the investor's portfolio decision are viewed through the lens of regret. Assuming the existence of a desirable target portfolio, we define regret as the difference in loss (or negative utility) between the simple portfolio and the target; our investor would like to hold a simple portfolio that is "almost as good as" a target. This chapter distills a potentially intractable dynamic selection procedure into one which requires specification of only a single threshold of regret. At the outset, the investor need only answer the question: *With what degree of certainty do I want my simple portfolio to be no worse than the target port-*

folio? Put differently: *What maximum probability of regretting my portfolio decision am I comfortable with?*

Once the regret threshold is specified, the investor's preference for portfolio simplicity will automatically adjust over time to accommodate this threshold. In other words, the penalty parameter λ_t continuously adjusts to satisfy the investor's regret tolerance. In one period, her portfolio may only need to be invested in a small number of assets to satisfy her regret threshold. However, that same portfolio may be far off from the target in the next period, requiring her to invest in more assets to accommodate the same level of regret. This thought experiment illustrates that our procedure, although requiring a static regret tolerance to be specified at the outset, results in investor preferences for sparsity that are *dynamic*.

The regret-based selection approach presented in this chapter is related to the *decoupling shrinkage and selection (DSS)* procedure from Hahn and Carvalho [2015] and Puelz et al. [2017]. In both papers, loss functions with explicit penalties for sparsity are used to summarize complex posterior distributions. Posterior uncertainty is then used as a guide for variable selection in static settings. This chapter expands on these notions by developing a regret-based metric for selection and placing it within a dynamic framework. The important innovations presented herein are (*i*) The use of investor's *regret* for selection and (*ii*) The development of a principled way to choose a *dynamic* penalty parameter λ_t (and thus select a portfolio) for a time-varying investment problem.

A key finding of this chapter is that sparse portfolios and their more complex (or dense) counterparts are often very similar from an *ex ante* regret perspective. More surprisingly, this similarity often persists *ex post*. This gives credence to a common piece of investing advice: "Don't worry about investing in a variety of funds; just buy the market."

2.1.1 Previous research

The seminal work of Markowitz [1952] provides the foundation of utility design for portfolio optimization related to this chapter. One area of relevant research highlighting Bayesian approaches to this problem may be found in Zhou et al. [2014]. In this paper, the authors consider portfolio construction with sparse dynamic latent factor models for asset returns. They show that dynamic sparsity improves forecasting and portfolio performance. However, sparsity in their context is induced at the factor loading level, not the portfolio decision. In contrast, our methodology seeks to sparsify the portfolio decision directly for *any* generic dynamic model.

Additional areas of research focus on the portfolio selection problem, particularly in stock investing and index tracking. Polson and Tew [1999] consider the S&P 500 index and develop a Bayesian approach for large-scale stock selection and portfolio optimization from the index's constituents. Other insightful Bayesian approaches to optimal portfolio choice include Johannes et al. [2014], Irie and West [2016], Zhao et al. [2016], Gron et al. [2012], Jacquier and Polson [2010b], Puelz et al. [2015] and Pettenuzzo and Ravazzolo [2015].

Methodological papers exploring high dimensional dynamic models relevant to this work are Carvalho et al. [2007] and Wang et al. [2011b].

2.1.2 Regularized portfolio optimization and the obsession with OOS Sharpe ratio

There is a sprawling literature in management science and operations research that focuses on portfolio construction and out of sample (OOS) performance. This performance is typically measured by the *out of sample Sharpe ratio*, which is defined as the ratio of OOS portfolio return and standard deviation of returns over a specified "testing" interval, i.e. 12 months. Mean-variance portfolio optimization is a cornerstone idea in this academic contingent as well [Markowitz, 1952]. Since this procedure relies on estimates of the first two moments of return, the optimal weights depend heavily on the modeling choice for these moments. It is well documented that sample estimates for the mean and variance produce portfolios with extreme weights that perform poorly out of sample [Jobson and Korkie, 1980], [Best and Grauer, 1991], [Broadie, 1993], [Britten-Jones, 1999], [DeMiguel et al., 2009b], [Frankfurter et al., 1971], [Dickinson, 1974], and [Frost and Savarino, 1988]. Estimation errors are the culprit, and it is widely acknowledged that errors in the expected return are much larger than errors in the covariance matrix, see for example Merton [1980]. As a result, researchers have focused on ways to stabilize, through regularization, the entire optimization process.

Regularization methods for portfolio optimization follow two schools of

thought. The first attacks estimation errors directly by attempting to regularize statistical estimates. This is often done in a Bayesian context by shrinking the mean estimates to a "grand mean" [James and Stein, 1961]. Similar applications of shrinkage used for regularization include Jorion [1985] and Jorion [1986] who set the mean return of the minimum variance portfolio to be the grand mean. In addition to the mean, others have focused on shrinking the covariance matrix in the portfolio selection problem, including Ledoit and Wolf [2003b], Ledoit and Wolf [2003a], and Garlappi et al. [2007]. In Garlappi et al. [2007], they consider parameter and model uncertainty to improve portfolio selection and out of sample Sharpe ratio. The second stream of literature focuses on regularizing the portfolio weights directly. This literature is distinct in that regularization is delegated to the portfolio optimization step in contrast to inference of the optimization inputs. This is a popular approach since mean-variance optimization can be formulated in terms of a Least Absolute Shrinkage and Selection Operator (LASSO) objective function from Tibshirani [1996] with fast computation of the optimal solution. In this setting, the weights in the objective function are penalized with an l-1 norm. Among the first to investigate weight regularization in the context of portfolio optimization is Brodie et al. [2009]. They document improved out of sample performance as measured by the Sharpe ratio and consistently beat a naive strategy of investing an equal amount in each asset. DeMiguel et al. [2009a] is a separate study showing similar results. These more recent studies have clarified and expanded on earlier work by Jagannathan and Ma [2002] who showed that

constraining the weights is equivalent to shrinking sample estimation of the covariance matrix. Fan et al. [2012] show that l-1 regularization limits gross portfolio exposure, leads to no accumulation of estimation errors, and results in outperformance compared to standard mean-variance portfolios. Recent work on regularization for portfolio selection includes Yen [2013], Yen and Yen [2014], Carrasco and Noumon [2011], Fernandes et al. [2012], Fastrich et al. [2013b] with applications to index tracking in Fastrich et al. [2014], Takeda et al. [2013], Wu and Yang [2014], and Wu et al. [2014].

There are many overlapping issues in this literature, and the rational for parameter shrinkage and weight regularization varies widely. Some motivate it statistically for overcoming estimation error and overfitting while others simply acknowledge that certain forms of regularization improve out of sample performance. The operations research approach to portfolio optimization faces a dilemma by straddling multiple fields. Careful statistical modeling of optimization inputs is *very* important; a bad model will yield bad decisions. As they often say in the computer science field when discussing appropriately designing an algorithm that takes inputs and gives an output: "garbage in, garbage out." Equal care must be given to utility design and optimization. Inaccurate representation of investor preferences will result in unreasonable optimal weight vectors and portfolio strategies. In general, the operations research approach to portfolio selection provides uninspiring and dated solutions to both modeling and optimization. Even worse, their metric for success (the out of sample Sharpe ratio – OOS SR) is an obsession that has led the

papers discussed above to undertake horserace comparisons at the expense of thoughtful research. In light of considering a decision-theoretic approach to portfolio selection, this obsession begs the question: If the OOS SR is what we ultimately care about, why not make that the utility and optimize it directly? This chapter seeks to strongly differentiate itself from the operations research approach to portfolio selection.

Under a new decision-theoretic lens, this chapter frames weight regularization in a simple portfolio selection methodology where we view it as a necessary part of the investor's utility. Our loss function regularizes dynamic weights to mirror the investor's joint preference for portfolio *simplicity* and high risk-adjusted return. Also, while previous studies conflate regularization and statistical uncertainty, our approach adapting the work of Hahn and Carvalho [2015] explicitly defines uncertainty's role in portfolio selection. We use well known Bayesian dynamic linear models to estimate time varying mean and variance and choose the amount of regularization using the predictive uncertainty in the investor's utility. By following the *DSS* procedure from Hahn and Carvalho [2015], we decouple (and deemphasize) modeling the mean and variance from the key problem of optimal portfolio choice. Our contribution to the literature is an approach that clearly separates challenging inference from the investor's preference for simplicity.

2.2 Overview

The focus will be loss functions of the following form:

$$\mathbf{L}_{\lambda_t}(w_t, \tilde{R}_t) = \mathcal{L}(w_t, \tilde{R}_t) + \Phi(\lambda_t, w_t) \tag{2.2}$$

where $\mathcal{L}$ is the negative utility of an investor, $\tilde{R}_t$ is a vector of N future asset returns, w_t is a vector of portfolio weights, and λ_t is a penalty parameter governing the degree to which the complexity function Φ is influential in the overall loss function. Let the future asset returns be generated from a model parameterized by Θ_t so that $\tilde{R}_t \sim \Pi(\Theta_t)$ and Π is a general probability distribution.

The time-varying preferences in (2.2) take into account an investor's negative utility as well as her desire for portfolio simplicity. Optimization of (2.2) in practice poses an interesting challenge since there is uncertainty in model parameters Θ_t and future asset returns $\tilde{R}_t$. Also, the penalty parameter λ_t is not known by the investor a priori, making her risk preferences ambiguous in portfolio complexity. A further obvious complication is that all of these unknowns are varying in time.

We propose a three-step approach to constructing a sequence of sparse dynamic portfolios. This procedure will be based on an investor's regret from investing in an alternative (target) portfolio defined by w_t^*. The three general steps are:

1. *Model specification:* Model the future asset returns $\tilde{R}_t \sim \Pi(\Theta_t)$.

2. *Loss specification:* Specify $\mathcal{L}$ and Φ in Loss (2.2). Then, the expected loss given by $\mathbf{L}_{\lambda_t}(w_t) = \mathbb{E}[\mathbf{L}_{\lambda_t}(w_t, \tilde{R}_t)]$ may be minimized for a sequence of $\lambda_t \, \forall t$. Define the collection of optimal portfolios in the cross-section as $\{w^*_{\lambda_t}\}$.

3. *Regret-based summarization:* Compare regret-based summaries of the optimal portfolios versus a target portfolio w^*_t by thresholding quantiles of a *regret* probability distribution, where regret as a random quantity is given by $\rho(w^*_{\lambda_t}, w^*_t, \tilde{R}_t)$. This random variable is a function of a sparse portfolio decision $w^*_{\lambda_t}$, the target portfolio w^*_t, and future asset returns.

The expectation and probability are both taken over the joint distribution of unknowns $(\tilde{R}_t, \Theta_t \mid \mathbf{R}_{t-1})$ where $\mathbf{R}_{t-1}$ is observed asset return data. Flexibly, the regret function $\rho(w^*_{\lambda_t}, w^*_t, \tilde{R}_t)$ can be any metric that is a function of the portfolio weights and unknowns and may be constructed using any target portfolio w^*_t. In this chapter, we consider the difference in loss between the sparse portfolio decision and the target portfolio $\rho(w^*_{\lambda_t}, w^*_t, \tilde{R}_t) = \mathcal{L}(w^*_{\lambda_t}, \tilde{R}_t) - \mathcal{L}(w^*_t, \tilde{R}_t)$ as our measure of regret in keeping with the usual decision theoretic definition.

We can now see how portfolio sparsity appears in the dynamic setting. The dynamic model given by $\tilde{R}_t \sim \Pi(\Theta_t)$ interacts with the portfolio decision w_t via the expected loss minimization in step 2. Iterating step 3 over time gives a sequence of sparse portfolio summaries $\{w^*_{\lambda^*_t}\}$ where λ^*_t provides an index for the selected sparse decision. Ultimately, these sparse portfolio summaries

select which subsets of assets are relevant for our "simple portfolio".

The details of this procedure will be fleshed out in the following subsection.

2.2.1 Details of Regret-based summarization

In following section, we discuss the specifics of the regret-based summarization procedure. We focus on expanding on step 3 of the methodology which represents the main innovation of this chapter. Model specification and fitting as well as loss specification comprising steps 1 and 2 are only highlighted; a detailed formulation of these first two steps is presented in the Application section.

Suppose that we have inferred a model for future asset returns given observed data $\mathbf{R}_{t-1}$ that is parameterized by the vector Θ_t: $\tilde{R}_t \sim \Pi(\Theta_t \mid \mathbf{R}_{t-1})$. Let the resulting posterior distribution of parameters and future asset returns be $p(\Theta_t, \tilde{R}_t \mid \mathbf{R}_{t-1})$. For example, Π may be parameterized by dynamic mean and variance parameters $\Theta_t = (\mu_t, \Sigma_t)$. Further, suppose we have specified the utility and complexity components of the investor's loss function: $\mathbf{L}_{\lambda_t}(w_t, \tilde{R}_t) = \mathcal{L}(w_t, \tilde{R}_t) + \Phi(\lambda_t, w_t)$.

For each investing period t, we obtain a sequence of portfolio decisions indexed by λ_t by optimizing the expected loss function $\mathbb{E}[\mathbf{L}_{\lambda_t}(w_t, \tilde{R}_t)]$. Regret-based summarization is an approach to select the appropriate optimal decision from the collection (i.e., select λ_t) for each time t, and this choice may be visualized by using sparsity-regret tradeoff plots.

Revisiting the regret, samples of the $\tilde{R}_t$ margin from posterior distribution $(\tilde{R}_t, \Theta_t \mid \mathbf{R}_{t-1})$ define the distribution for the regret random variable $\rho(w^*_{\lambda_t}, w^*_t, \tilde{R}_t)$ given by a difference in loss:

$$\rho(w^*_{\lambda_t}, w^*_t, \tilde{R}_t) = \mathcal{L}(w^*_{\lambda_t}, \tilde{R}_t) - \mathcal{L}(w^*_t, \tilde{R}_t), \tag{2.3}$$

where $w^*_{\lambda_t}$ are the optimal sparse portfolio weights for penalty parameter λ_t and w^*_t are the weights for a target portfolio – any portfolio decision the investor desires to benchmark against. Regret (2.3) is a random variable whose uncertainty is induced from the joint posterior distribution $(\tilde{R}_t, \Theta_t \mid \mathbf{R}_{t-1})$ from step 1 of the procedure.

We use the regret random variable as a tool for sparse portfolio selection. Each portfolio decision indexed by λ_t is assigned a number:

$$\pi_{\lambda_t} = \mathbb{P}[\rho(w^*_{\lambda_t}, w^*_t, \tilde{R}_t) < 0] \tag{2.4}$$

which is the probability that the sparse portfolio is no worse than the dense (target) portfolio. In other words, π_{λ_t} is the probability that I will not "regret" investing in the sparse λ_t-portfolio over the target portfolio. This may also be called the *satisfaction probability* for the sparse λ_t portfolio decision.

In Figure (2.1), we provide an illustration of the connection bewteen the loss and regret random variables. This figure is constructed using returns on 25 passive indices and a next period "log cumulative wealth" utility function. This is done for a snapshot in time with a focus on one sparse decision that is invested in 4 out of the 25 indices. The investor is considering this decision

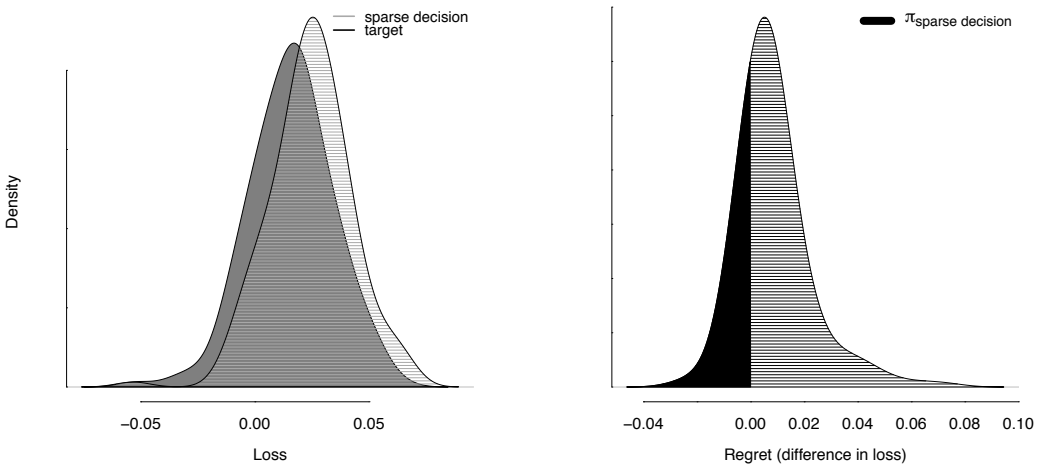


Figure 2.1: Loss (left) and regret (right) for an example using returns on 25 passive indices. The loss is defined by the log cumulative wealth. The sparse decision is a portfolio invested in 4 indices and is represented by the light shaded gray region. The target decision is a portfolio optimized over all 25 indices and is represented by the shaded black region. The regret distributions shown on the right represent the random variables constructed by subtracting the sparse decision loss from the target loss. Additionally, the black shaded region on the right shows $\pi_{\text{sparse decision}}$: The probability that the sparse decision is no worse than the target decision.

versus her target – a portfolio optimized over all 25 indices. The left figure displays the loss distributions of the sparse decision and target. The probability mass of the sparse loss is gathered at larger values compared with the target loss. It is "more costly" (higher loss potential) to neglect diversification benefits and invest in fewer assets.

The right plot in Figure (2.1) displays the regret distribution for the sparse decision. This is constructed by taking the difference between the sparse

and target losses, as given in Equation (2.3), defining the regret random variable. With the regret distribution in hand, we can compute the probability that the sparse decision is no worse than the target portfolio given by Equation (2.4) – this may be referred to as the "satisfaction probability" for the sparse decision. This is shown by the black shaded area on the right in Figure (2.1). The larger this probability, the "closer" the sparse decision's loss is to the target loss. By making a decision that satisfies a lower bound on this probability called κ, we are able to control our chance of being the same or better than a target portfolio. A lower bound (κ) on the probability of satisfaction (no regret) implies an upper bound ($1 - \kappa$) on the probability of regret.

The investor's portfolio decision boils down to answering the question first posed in the introduction: *With what degree of certainty do I want my simple portfolio to be no worse than the target portfolio?* As the investor moves through time, the loss and regret distributions will evolve and so will the probability associated with the sparse λ_t portfolio decisions. A dynamic sparse portfolio decision extends this probability thresholding approach to a time varying framework. The investor chooses the a portfolio decision satisfying $\pi_{\lambda_t} > \kappa \ \ \forall t$, ensuring she holds a portfolio that satisfies her regret tolerance in every investing period.

In Figure (2.2), we show an example sequence of regret distributions (right vertical axis and gray bars) indexed by λ_t as well as the satisfaction probability π_{λ_t} (left vertical axis and open circles). Specifically, we show the regret distributions for 300 sparse decisions under the log cumulative wealth

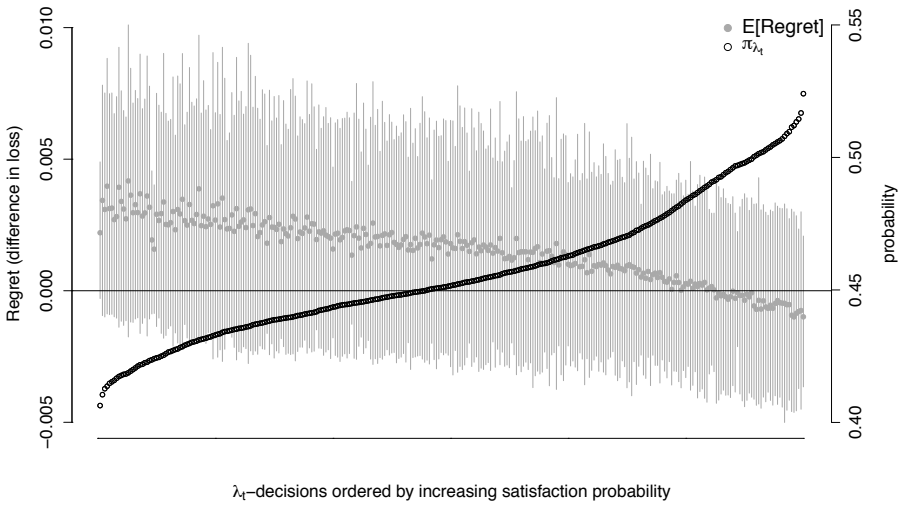


Figure 2.2: Regret distributions (left vertical axis) and π_{λ_t} (right vertical axis) for increasing π_{λ_t} values from left to right on the horizontal axis. Displayed are 300 sparse portfolio decisions indexed by λ_t. As the satisfaction probability (π_{λ_t}) increases, the mean regret represented by the gray dots will will typically trend downwards. Gray bands represent 20% centered posterior credible intervals for the regret.

utility and versus a target portfolio optimized over all available assets. The highest regret decisions (by satisfaction probability) are on the left, and they become less regretful as one moves to the right on the horizontal axis. These sparse portfolio decision are all fairly close in terms of loss to the target decision. Therefore, the corresponding regret distributions hover around zero, and the π_{λ_t}'s hover around 0.5. Exploratory plots like Figure (2.2) aid in choosing a proper value for the κ threshold. In this case $\pi_{\lambda_t} > 0.5$ is quite high.

Once the time invariant threshold κ is specified, a dynamic selection strategy is easily implementable. At each time t, we are presented with a

set of decisions such as those displayed in Figure (2.2) and choose the sparse portfolio decision such that its $\pi_{\lambda_t} > \kappa$. If there are several sparse decisions satisfying the threshold, we may choose the one whose π_{λ_t} is closest to κ. Of course, there is flexibility in the final step of selecting among admissible sparse decisions. For example, one may select a sparse decision at time t that is "close" (in terms of a norm or assets held) to the previous sparse decision at time $t-1$ to reduce transaction costs related to the buying and selling of assets. These features will be discussed in the Application section.

2.3 Application

This section is divided into three parts. First, we describe a dynamic linear model used to infer time-varying moments of asset returns and fulfill step 1 of the methodology. Second, we specify loss and resulting regret functions used for selection (steps 2 and 3). Third, we demonstrate the methodology using a set of 25 passive funds. In the demonstration sections, we consider a simple example and a more practical case study.

2.3.1 Model specification and data

The general model we infer parameterizes the distribution of future asset returns with a mean and covariance indexed by time: $\tilde{R}_t \sim \Pi(\mu_t, \Sigma_t)$. An important feature of our proposed methodology is that any model providing estimations of these time varying moments may be used.

To demonstrate our methodology, we estimate a dynamic linear model

(DLM) motivated by Fama and French [2015] who detail five "risk factors" with future returns $\tilde{R}_t^F$ as relevant to asset pricing. Specifically, we model the joint distribution of future asset returns and factor returns $p(\tilde{R}_t, \tilde{R}_t^F)$ compositionally by modeling $p(\tilde{R}_t \mid \tilde{R}_t^F)$ and $p(\tilde{R}_t^F)$. Following the dynamic linear model setup from Harrison and West [1999], the future return of asset i is a linear combination of future factor returns $(\tilde{R}_t^i \mid \tilde{R}_t^F)$:

$$\begin{aligned}
\tilde{R}_t^i &= (\beta_t^i)^T \tilde{R}_t^F + \epsilon_t^i, \qquad \epsilon_t^i \sim N(0, 1/\phi_t^i), \\
\beta_t^i &= \beta_{t-1}^i + w_t^i, \qquad w_t^i \sim \mathrm{T}_{n_{t-1}^i}(0, W_t^i),
\end{aligned}$$

$$\begin{aligned}
\beta_0^i \mid D_0 &\sim \mathrm{T}_{n_0^i}(m_0^i, C_0^i), \\
\phi_0^i \mid D_0 &\sim \mathrm{Ga}(n_0^i/2, d_0^i/2),
\end{aligned} \tag{2.5}$$

$$\begin{aligned}
\beta_t^i \mid D_{t-1} &\sim \mathrm{T}_{n_{t-1}^i}(m_{t-1}^i, R_t^i), \qquad R_t^i = C_{t-1}^i/\delta_\beta, \\
\phi_t^i \mid D_{t-1} &\sim \mathrm{Ga}(\delta_\epsilon n_{t-1}^i/2, \delta_\epsilon d_{t-1}^i/2),
\end{aligned}$$

where $W_t^i = \frac{1-\delta_\beta}{\delta_\beta} C_{t-1}^i$ and D_t is all information up to time t. This model permits the coefficients on the factors as well as the observation and state level variances to vary in time. Pre-specified discount factors δ_ϵ and $\delta_\beta \in (0.8, 1)$ accomplish this goal for the observation and state level variances, respectively. Also, note that C_t^i (the posterior variance of the state equation for β_t^i) is updated through moments of the prior $\beta_t^i \mid D_{t-1}$ and the one-step ahead forecast distribution $\tilde{R}_t^i \mid D_{t-1}$. Theorem 4.1 in Harrison and West [1999] provides the general updating equations for the univariate DLM. Table 10.4 in

the book summarizes the recurrence relationships in the special case of variance discounting, providing the moments of the posteriors of the parameters $\{m_t^i, C_t^i, n_t^i, d_t^i\}$ for all t and each asset i.

We model the five factor future returns $\tilde{R}_t^F$ with a full residual covariance matrix using the following matrix normal dynamic linear model:

$$\begin{aligned}
\tilde{R}_t^F &= \mu_t^F + \nu_t \qquad \nu_t \sim \mathrm{N}(0, \Sigma_t^F), \\
\mu_t^F &= \mu_{t-1}^F + \Omega_t \qquad \Omega_t \sim \mathrm{N}(0, W_t, \Sigma_t^F),
\end{aligned}$$

$$(\mu_0^F, \Sigma_0^F \mid D_0) \sim \mathrm{NW}_{n_0}^{-1}(m_0, C_0, S_0), \tag{2.6}$$

$$\begin{aligned}
&(\mu_t^F, \Sigma_t^F \mid D_{t-1}) \sim \mathrm{NW}_{\delta_F n_{t-1}}^{-1}(m_{t-1}, R_t, S_{t-1}), \\
&R_t = C_{t-1}/\delta_c,
\end{aligned}$$

where $W_t = \frac{1-\delta_c}{\delta_c} C_{t-1}$. Analogous to Model (2.5), the discount factors δ_F and δ_c in Model (2.6) serve the same purpose of permitting time variation in the observation and state level variances, respectively. An added benefit of (2.6) is that Σ_t^F is a full residual covariance matrix.

Elaborating on the intuition behind Models (2.5) and (2.6) and guided by Harrison and West [1999], the purpose of variance discounting is to provide a natural way to evolve the variance from the posterior to the prior while maintaining conjugacy for sequential updating. For example, consider the posterior of the precision in Model (2.5):

$$\phi_{t-1}^i \mid D_{t-1} \sim \mathrm{Ga}(n_{t-1}^i/2, d_{t-1}^i/2). \tag{2.7}$$

To construct $p(\phi_t^i \mid D_{t-1})$, we wish to maintain a Gamma form so it is conjugate with the likelihood function for $\tilde{R}_t^i$ given by the one-step ahead forecast distribution. One reasonable approach is to preserve the mean of Distribution (2.7), but inflate the variance by discounting the degrees of freedom parameter $n_{t-1} \to \delta_\epsilon n_{t-1}$. The prior distribution then becomes:

$$\phi_t^i \mid D_{t-1} \sim \text{Ga}(\delta_\epsilon n_{t-1}^i/2, \delta_\epsilon d_{t-1}^i/2). \tag{2.8}$$

Moving from Distribution (2.7) to (2.8) increases the dispersion of the prior to represent a "loss of information" characteristic of moving forward to time t with a lack of complete information in D_{t-1}. The remaining discount factors δ_β, δ_C, and δ_F in Models (2.5) and (2.6) serve the analogous purpose of *variance inflation* for their respective stochastic processes.

The limiting behavior of variance-discounted learning corresponds to exponentially weighting historical data with decreasingly smaller weights given to values further in the past (see sections 10.8.2-3 in Harrison and West [1999]). Larger discount factors correspond to slower decaying weights and suggest the time series of parameters is slower-fluctuating. Smaller discount factors intrinsically mean we have less data with which to estimate the parameters because the sequence is believed to be more rapidly fluctuating. Thus, the choice of discount factors amounts to choosing decaying weights for previous data that are relevant for predicting the parameters today.

Models (2.5) and (2.6) together constitute a time-varying model for the joint distribution of future asset and factor returns: $p(\tilde{R}_t, \tilde{R}_t^F) = p(\tilde{R}_t \mid$

$\tilde{R}_t^F)p(\tilde{R}_t^F)$. As detailed in Harrison and West [1999], they are Bayesian models that have closed form posterior distributions of all parameters at each time t, and the absence of MCMC is convenient for fast updating and uncertainty characterization – a necessary ingredient for our regret-based portfolio selection procedure. Under these models, we obtain the following mean and covariance structure:

$$\begin{aligned} \mu_t &= \beta_t^T \mu_t^F, \\ \Sigma_t &= \beta_t \Sigma_t^F \beta_t^T + \Psi_t, \end{aligned} \tag{2.9}$$

where column i of β_t are the coefficients on the factors for asset i, β_t^i. Also, Ψ_t is a diagonal matrix with ith element $\Psi_{tii} = 1/\phi_t^i$. The parameters $\Theta_t = (\mu_t, \Sigma_t)$ are inputs to step 2 of the procedure.

2.3.1.1 Data and choice of discount factors

We use data on the 25 most highly traded (i.e., most liquid) equity funds from ETFdb.com as our investable assets. This is monthly data from the Center for Research in Security Prices (CRSP) database from February 1992 through May 2016 [CRSP, 1992-2016].

The fund names, tickers, and sample statistics are displayed in Table (2.1). The returns on the Fama-French five factors are obtained from the publicly available data library on Ken French's website.[1] We start with 10 years of data to train the model and begin forming portfolios in February 2002.

[1] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/

Fund name	**Ticker**	**Return (%)**	**St. Dev. (%)**
SPDR Dow Jones Industrial Average	DIA	8.06	14.15
SPDR S&P 500 ETF	SPY	7.46	14.34
SPDR Industrial Select Sector	XLI	9.19	16.49
Guggenheim S&P 500 Equal Weight ETF	RSP	9.37	15.92
iShares MSCI Emerging Markets	EEM	6.06	22.78
iShares MSCI EAFE	EFA	3.94	16.43
iShares MSCI Germany	EWG	6.82	22.23
iShares MSCI Japan	EWJ	0.22	19.28
iShares MSCI United Kingdom	EWU	4.63	15.88
iShares MSCI South Korea Capped	EWY	9.46	36.78
iShares MSCI Eurozone	EZU	6.00	19.85
iShares S&P 500 Value	IVE	7.36	14.95
iShares Core S&P 500	IVV	7.48	14.34
iShares Russell 1000	IWB	8.22	14.00
iShares Russell 1000 Value	IWD	8.10	14.26
iShares Russell 1000 Growth	IWF	7.63	14.56
iShares Russell 2000	IWM	8.00	18.76
iShares Russell 2000 Value	IWN	7.78	18.67
iShares Russell 2000 Growth	IWO	6.63	22.14
iShares Russell Mid-Cap Growth	IWP	8.52	19.98
iShares Russell Mid-Cap	IWR	9.52	15.96
iShares Russell 3000	IWV	7.52	14.61
iShares US Real Estate	IYR	9.32	19.12
iShares US Technology	IYW	11.64	26.09
iShares S&P 100	OEF	7.32	14.61

Table 2.1: List of exchange-traded funds (ETFs) used for the empirical study. Also displayed are the ticker symbols and realized return and standard deviation (annualized) over their sample period.

Step 1 in our procedure is specifying and inferring a model for asset returns. For our empirical analysis, we use Models (2.5) and (2.6). The discount factors for the factor coefficient and factor mean processes are set to $\delta_c = \delta_\beta = 0.9925$, and we consider time varying residual variances $\delta_F = \delta_\epsilon = 0.97$. Evidence of time varying residual variance is well-documented in the finance literature [Ng, 1991]. The discount factors are chosen to incorporate

an adequate amount of data in the exponentially weighted moving window. When $\delta = 0.9925$ (0.97), data eight (three) years in the past receives half the weight of data today. We require slower decay for the factor coefficients and mean processes because more data is needed to learn these parameters than residual volatility.

2.3.2 Loss and regret specification

We consider a loss function defined by the negative log cumulative return of a portfolio decision for N assets. Recalling general form of the loss function: $\mathbf{L}_{\lambda_t}(w_t, \tilde{R}_t) = \mathcal{L}(w_t, \tilde{R}_t) + \Phi(\lambda_t, w_t)$, define:

$$\mathcal{L}(w_t, \tilde{R}_t) = -\log\left(1 + \sum_{k=1}^{N} w_t^k \tilde{R}_t^k\right), \tag{2.10}$$

The utility in Loss (2.10) may be viewed as a version of the Kelly portfolio criterion [Kelly Jr, 1956] where the investor's preferences involve the portfolio growth rate. The complexity function $\Phi(\lambda_t, w_t)$ is separately specified in each of the two examples to follow.

Portfolio decisions w_t may now be evaluated using the negative log cumulative return preferences given by $\mathcal{L}(w_t, \tilde{R}_t)$. However, in order to find these portfolio decisions, we must first optimize the expectation of Loss (2.10). We do this in two steps: First, we approximate the loss using a second order Taylor expansion, and second, we take the expectation over all unknowns and optimize for each λ_t.

Following the work of Rising and Wyner [2012], we consider a conve-

nient second order Taylor approximation $\mathcal{L}(w_t, \tilde{R}_t) \approx \mathring{\mathcal{L}}(w_t, \tilde{R}_t)$ of the original Loss (2.10) expanded about $w_0 = \vec{0}$:

$$\mathring{\mathcal{L}}(w_t, \tilde{R}_t) = \frac{1}{2} \sum_{k=1}^{N} \sum_{j=1}^{N} w_t^k w_t^j \tilde{R}_t^k \tilde{R}_t^j - \sum_{k=1}^{N} w_t^k \tilde{R}_t^k, \tag{2.11}$$

where we write the approximate loss including the penalty function as $\mathring{\mathbf{L}}_{\lambda_t}(w_t, \tilde{R}_t) = \mathring{\mathcal{L}}(w_t, \tilde{R}_t) + \Phi(\lambda_t, w_t)$. The approximate expected loss $\mathbb{E}[\mathring{\mathbf{L}}_{\lambda_t}(w_t, \tilde{R}_t)]$ is written as:

$$\mathbb{E}[\mathring{\mathbf{L}}_{\lambda_t}(w_t, \tilde{R}_t)] = \mathbb{E}\left[\frac{1}{2} \sum_{k=1}^{N} \sum_{j=1}^{N} w_t^k w_t^j \tilde{R}_t^k \tilde{R}_t^j - \sum_{k=1}^{N} w_t^k \tilde{R}_t^k + \Phi(\lambda_t, w_t)\right]. \tag{2.12}$$

With the posterior distribution $(\Theta_t, \tilde{R}_t \mid \mathbf{R}_{t-1})$ in hand from step 1, we can take the expectation. We integrate over $(\tilde{R}_t \mid \Theta_t, \mathbf{R}_{t-1})$ followed by $(\Theta_t \mid \mathbf{R}_{t-1})$ to obtain the integrated approximate loss function:

$$\begin{aligned} \mathring{\mathbf{L}}_{\lambda_t}(w_t) = \mathbb{E}[\mathring{\mathbf{L}}_{\lambda_t}(w_t, \tilde{R}_t)] &= \mathbb{E}_{\Theta_t}\left[\mathbb{E}_{\tilde{R}_t|\Theta_t}\left[\mathring{\mathbf{L}}_{\lambda_t}(w_t, \tilde{R}_t)\right]\right] \\ &= \mathbb{E}_{\Theta_t}\left[\frac{1}{2} w_t^T \Sigma_t^{\text{nc}} w_t - w_t^T \mu_t + \Phi(\lambda_t, w_t)\right] \\ &= \frac{1}{2} w_t^T \overline{\Sigma}_t^{\text{nc}} w_t - w_t^T \overline{\mu}_t + \Phi(\lambda_t, w_t), \end{aligned} \tag{2.13}$$

where the overlines denote posterior means of the mean μ_t and non-central second moment Σ_t^{nc}. The non-central second moment is calculated from the variance as $\Sigma_t^{\text{nc}} = \Sigma_t + \mu_t \mu_t^T$. Loss function (2.13) may be minimized for a range of λ_t values at each time t.

In the subsections to follow, we will present two analyses using this model and data. First, we discuss a simple unsupervised example to demonstrate the regret-based selection procedure. Second, we present an in depth practical case study.

2.3.3 Example: Portfolio decisions with limited gross exposure

In this section, we present a simple example demonstrating the main components of regret-based selection. We complete Loss function (2.2) by specifying the complexity function as the ℓ_1 norm of the weight vector: $\Phi(\lambda_t, w_t) = \lambda_t \|w_t\|_1$. The complexity function measures *gross exposure* of a decision by summing the absolute value of each position: $\|w_t\|_1 = \Sigma_i |w_t^i|$. Decisions with larger absolute value components will evaluate to larger ℓ_1 norms. The penalty parameter λ_t corresponds directly to a single portfolio decision by amplifying the penalty in the loss function.

The approximate Loss function (2.13) is now convex and may be written as:

$$\mathring{\mathbf{L}}_{\lambda_t}(w_t) = \frac{1}{2} w_t^T \overline{\Sigma}_t^{\text{nc}} w_t - w_t^T \overline{\mu}_t + \lambda_t \|w_t\|_1 . \tag{2.14}$$

Loss function (2.14) is readily optimized by a variety of software packages – please see the Appendix for details. Given its computational convenience, it possesses a couple important features worth noting.

First, Loss (2.14) requires no enumeration of decisions; it can be minimized quickly for a range of λ_t. In this way, it is an "unsupervised" approach to the sparse portfolio selection problem. Second, λ_t now has explicit meaning beyond indexing the decisions. Since it is multiplying the complexity function, larger (smaller) λ_t will generally correspond to sparser (denser) portfolio decisions. Conveniently, this displays the regret-based procedure's usefulness

in selecting tuning parameters in penalized optimization problems with time-varying inputs. The dynamic nature of asset return data renders traditional cross validation approaches using i.i.d. sampled testing and training splits inappropriate.

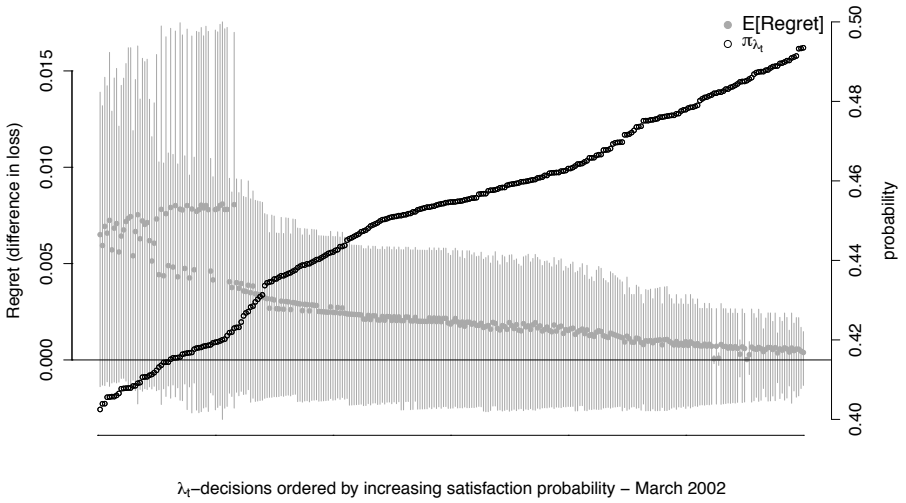


Figure 2.3: Regret distributions (left vertical axis) and π_{λ_t} (right vertical axis) for increasing π_{λ_t} values from left to right on the horizontal axis. Displayed are 300 of the sparse portfolio decisions indexed by λ_t for March 2002. As the satisfaction probability (π_{λ_t}) increases, the mean regret represented by the gray dots will will typically trend downwards. Gray bands represent 20% centered posterior credible intervals for the regret.

We optimize Loss function (2.14) for a range of λ_t and for each time t. Specifically, we consider 500 λ_t values spanning the sparsest "one asset decision" to the $\lambda_t = 0$ decision. The latter decision is referred to as the unpenalized *Kelly optimal* portfolio and puts nonzero weight in all available assets. This is used as the target decision since there is no limit on gross

exposure and will also be called the *dense portfolio*. We also normalize all decisions to sum to one, and allow long *and* short positions in assets. We normalize all decisions so that the investor is neither borrowing nor investing in the risk free (cash) rate. Instead, she is fully invested in a risky asset portfolio. In other words, denoting w_{cash} and w_{risky} as the percentage of wealth in cash and risky assets (the ETFs) respectively, $w_{\text{cash}} + w_{\text{risky}} = 1$ will always hold, we consider the case when $w_{\text{risky}} = 1$.

At each point in time, the investor would like to choose among the 500 sparse decisions indexed by λ_t. This may be done by first computing the corresponding satisfaction probabilities π_{λ_t} for each of the 500 decisions under consideration and then choosing one that satisfies and a pre-specified threshold κ. Recall that π_{λ_t} is defined in (2.4) as the probability that the regret (versus the dense target) is less than zero. The utility, as specified in Equation (2.10), is the next period log cumulative wealth.

Figure (2.3) displays the cross-sectional regret distributions (left vertical axis) and satisfaction probabilities π_{λ_t} (right vertical axis) for 300 of the sparse decisions in March 2002. As prescribed by the regret-based procedure, the investor uses this information to select a portfolio. The satisfaction probabilities span 0.4 to 0.5 indicating that the decisions in this investing period are all quite similar. Guided by this figure, we choose a $\kappa = 42.5\%$ threshold to construct an example sequence of selected portfolios.

Once the static threshold is selected, we can iterate the selection procedure through time. At each time t, the investor is confronted with ex ante

regret information provided by a cross-sectional plot like Figure (2.3) and selects a portfolio that satisfies the threshold κ. Once the sequence of decisions is constructed, we can look at how the regret distribution varies *over time*.

Figure (2.4) shows precisely how the regret of the selected decisions evolve over time. This example demonstrates how both the mean (black line) and variance (surrounding shaded black regions) of regret can vary substantially. Notice that the regret is close to zero with small variance for most periods of time. However, surrounding the financial crisis in 2009, the mean increases and then drops below zero and the variance increases. When regret is negative, the utility of sparse portfolio decision exceeds that of the dense portfolio. During crisis periods shortly into 2009, sparse portfolio decisions appear to be preferred (as measured by ex ante investor utility) to the dense portfolio. Nonetheless, this drop in mean is accompanied by increased variance which informs the investor to be wary of the precision of her regret assessment.

2.3.4 Case study: Selection among a large set of sparse decisions

The purpose of the following case study is to demonstrate regret-based portfolio selection for a layman investor. We assume our investor would like to hold a broad market fund (SPY) and a couple more diversifying positions in other funds. Additionally, we consider a scenario where the investor cannot hold negative positions in funds; i.e., short selling is prohibited. Therefore, we consider decisions of only positive weights: $w_t \geq 0 \ \forall t$. We construct a set of portfolio decisions for a layman investor using the following rules for a sparse

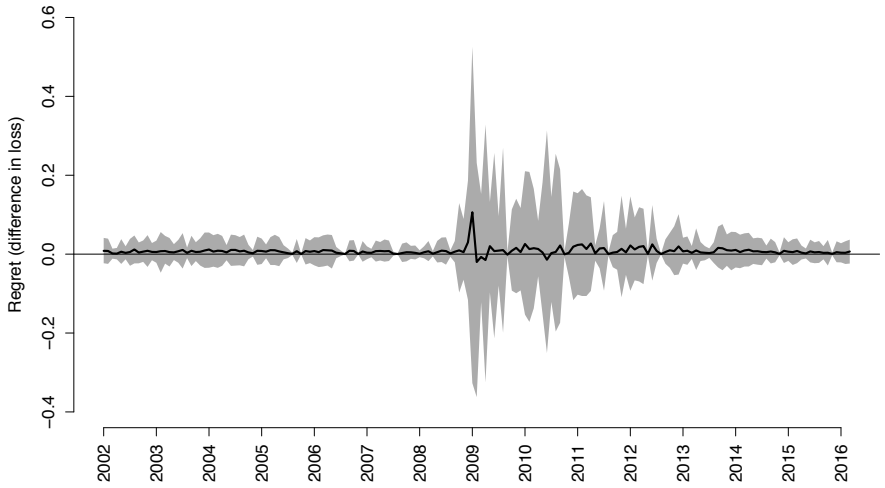


Figure 2.4: The evolution of the ex ante regret distributions for the sparse long/short portfolio decision given by a $\kappa = 42.5\%$ threshold and versus the unpenalized Kelly optimal target. The mean regret is shown by the lines and the surrounding areas represent the evolving centered 60% posterior credible intervals.

portfolio with $q < N$ funds:

1. $\geq 25\%$ of the portfolio is invested in SPY, a broad market fund tracking an index comprised of the 500 largest US companies by market capitalization.

2. $\geq 25\%$ of the portfolio is diversified across the $q - 1$ non-market funds in the following way: The $q - 1$ non-market funds each have weights $\geq \frac{25}{q-1}\%$.

We consider portfolios of two, three, four, and five funds, all of which include SPY. Each of these sparse portfolios are optimized using the unpenalized *Kelly optimal* loss as the objective (Loss (2.13) without the complexity function) and constraints defined as above. Since our data has 24 funds excluding SPY, enumeration of decisions in this way results in $\Sigma_{i=1}^{4}\binom{24}{i} = 12,950$ sparse portfolios to select among. Enumeration of sparse decisions implies a complexity function that measures the *cardinality* or number of funds included in a portfolio. Since the complexity function is now implicit in the sparse enumeration, λ_t may be thought of as a convenient indexing of each possible portfolio decision.

As presented in the initial example, we must specify a target decision which then defines the regret random variable defined in Equation (2.3). We consider two targets at opposite ends of the sparsity spectrum for the empirical analysis.

1. **Dense target**: The unpenalized *Kelly optimal* decision; a portfolio optimized over all available assets. Define the Kelly optimal decision as $w_t^* = \arg\min_{w_t \geq 0} \mathring{\mathcal{L}}(w_t)$ where $\mathring{\mathcal{L}}(w_t) = \mathbb{E}[\mathring{\mathcal{L}}(w_t, \tilde{R}_t)] = \frac{1}{2} w_t^T \overline{\Sigma}_t^{\text{nc}} w_t - w_t^T \overline{\mu}_t$; the optimal decision in absence of the penalty function. This is the same target used in the "ℓ_1 penalty" example presented above, now with a positivity constraint on the weights.

2. **Sparse target**: The *market*; a portfolio composed of one asset representing broad exposure to the financial market. We choose SPY as the market fund.

The choice of each target will give an investor vastly different perspectives on sparse portfolio selection. In the dense target case, the investor desires a sparse portfolio decision that is close (in terms of regret) to a potentially unattainable decision involving all possible funds. The sparse target turns this approach on its head. In this case, the sparse target approach will inform the investor of the added benefit (if any) in diversifying away from a broad market fund.

Each of the 12,950 sparse decisions has a probability of satisfaction versus a target (π_{λ_t}) which can be readily calculated via simulation at each point in time using Equations (2.3) and (2.4) and the distribution of future returns given by Models (2.5) and (2.6). In Figure (2.5), we show histograms of the satisfaction probabilities for March 2002 across all 12,950 sparse decisions. It is related to Figure (2.2) in that the satisfaction probabilities corresponding to the right vertical axis are shown in histogram form, now for various targets. The probabilities versus the dense (SPY) target are shown in black (gray). The dense target is the *dense* portfolio optimized over all 25 funds. Satisfaction versus this dense portfolio decision are gathered at smaller probabilities when compared with the SPY portfolio decision. Of course, the satisfaction rate versus a diversified dense portfolio will intuitively be lower than versus a sparse portfolio of a single fund.

Figure (2.5) aids in the proper choice of the regret threshold κ. When evaluated under the next period cumulative wealth utility, all long only portfolio decisions are somewhat similar. Thus, the regret (difference in loss) will

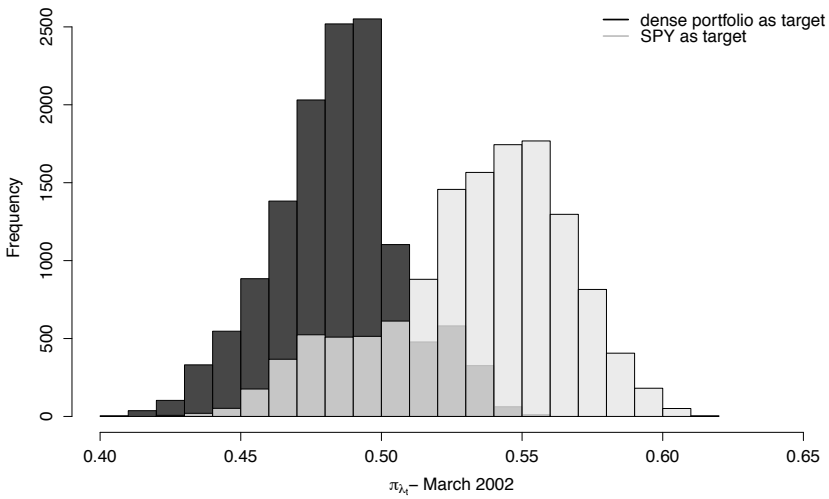


Figure 2.5: Histograms of the satisfaction probabilities (π_{λ_t}) for two target decisions: The dense portfolio (black) and SPY (i.e., sparse target, in gray). These are shown for the March 2002 investing period and are distributions across all 12,950 enumerated sparse decisions.

generally be gathered around values close to zero, and the satisfaction probabilities will be gathered around value close to 0.5. As a next step, we select a κ and present the resulting dynamic portfolio decision for the dense target and SPY target.

We show the selected sparse decisions with the $\kappa = 45\%$ threshold for the dense (portfolio optimized over all funds in white columns) and sparse (portfolio of only SPY in gray columns) target decisions in Table (2.2). Each of these decisions possesses the property that, at each point in time, the satisfaction probability of investing in this decision versus the respective target is at least 45%. The portfolios are updated on a monthly basis; the table

displays annual weights for brevity. There are many decisions that will satisfy this threshold for each of the target (see for example the probability mass above 0.45 in Figure (2.5)). In this case, we have added flexibility in which sparse decision to choose. We construct the sparse decisions so that at most one fund is selected or removed from month to month. For example, if the current portfolio at time t has SPY, OEF, and IVV, two admissible portfolios at $t+1$ could include the funds be SPY, OEF, IVV, and EWG *or* SPY and OEF assuming they both also satisfy the κ threshold.

Dates	SPY		EZU		EWU		EWY		EWG		EWJ		OEF		IVV		IVE		EFA		IWP	
2003	25	75	-	-	58	-	-	-	-	-	-	-	-	8.3	-	-	-	-	-	-	-	8.3
2004	25	75	-	-	43	-	-	-	-	-	20	-	-	-	6.2	-	-	-	-	-	-	-
2005	25	75	-	-	25	-	-	-	6.2	-	13	-	-	8.3	-	-	-	-	-	-	-	-
2006	62	75	-	-	-	-	-	-	6.2	-	19	-	-	12	-	-	-	-	-	-	-	-
2007	75	75	-	-	-	-	25	-	-	8.3	-	-	-	-	-	-	-	-	-	-	-	8.3
2008	44	75	-	-	-	-	-	12	8.3	13	21	-	-	-	-	-	-	-	26	-	-	-
2009	30	45	-	-	-	-	6.2	-	-	-	41	-	-	-	-	34	-	-	17	21	6.3	-
2010	75	55	-	-	-	-	8.3	-	-	-	-	-	-	-	-	26	-	-	-	11	8.3	-
2011	58	57	-	-	25	-	-	-	-	-	-	-	-	-	-	26	-	-	-	-	8.3	-
2012	29	25	8.3	-	-	-	-	-	-	-	54	-	-	56	-	6.2	-	-	-	-	-	-
2013	34	25	-	-	-	-	-	6.2	-	6.2	49	-	-	56	-	-	-	-	-	-	8.3	-
2014	25	75	-	-	-	-	-	-	-	-	37	-	26	-	-	-	-	-	6.2	-	-	-
2015	45	25	-	-	-	-	-	-	-	-	39	36	-	27	-	-	8.3	-	-	6.2	8.3	-
2016	35	75	-	-	-	-	-	-	-	-	40	-	-	-	17	-	-	8.3	-	8.3	8.3	8.3

Dates	IWR		IWF		IWN		IWM		IYW		IYR		RSP		EEM		IWO		IWV	
2003	-	-	-	-	8.3	-	-	-	-	8.3	-	-	8.3	-	-	-	-	-	-	-
2004	-	-	-	12	-	-	-	-	-	12	-	-	6.2	-	-	-	-	-	-	-
2005	-	-	-	8.3	-	-	-	-	-	8.3	30	-	-	-	-	-	-	-	-	-
2006	-	-	6.3	-	-	-	6.2	-	-	12	-	-	-	-	-	-	-	-	-	-
2007	-	-	-	-	-	-	-	-	-	8.3	-	-	-	-	-	-	-	-	-	-
2008	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2009	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2010	-	-	-	-	-	-	-	-	8.3	8.3	-	-	-	-	-	-	-	-	-	-
2011	-	-	-	-	-	-	-	-	8.3	8.3	-	-	-	-	-	-	-	-	-	8.3
2012	-	-	-	6.2	-	-	-	-	8.3	-	-	-	-	-	-	-	-	-	-	6.3
2013	-	-	-	-	-	-	-	-	8.3	-	-	-	-	-	-	6.3	-	-	-	-
2014	6.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	12	-	12	-	-
2015	-	-	-	-	-	-	-	-	-	6.2	-	-	-	-	-	-	-	-	-	-
2016	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 2.2: Sparse portfolio decisions (in percent) for DLMs (2.5) and (2.6) with $\delta_F = \delta_\epsilon = 0.97$ and $\delta_c = \delta_\beta = 0.9925$. Shown are the selected portfolio decisions for the two targets: dense portfolio (left column in white) and SPY (right column in gray). Note that annual weights are shown for brevity although portfolios are updated monthly. In this dynamic portfolio selection, the regret threshold is $\kappa = 45\%$ for both targets.

In Table (2.2), the sparse decision for the SPY target has larger allocations to SPY over the trading period compared with the sparse decision

for the dense portfolio target. Also, it possesses a consistent allocation to the US technology sector fund IYW. In contrast, the sparse decision for the dense target often possesses a significant allocation to the Japanese equity specific fund EWJ.

Figure (2.6) displays the evolution of the regret distributions for the sparse decisions shown in Table (2.2). The lines are the expected regret, and the surrounding areas correspond to the centered 60% posterior credible intervals. The expected regret for both decisions remains close to zero and for most investing periods is slightly above zero; this is by construction since we choose the sparse decision that satisfies the $\kappa = 45\%$ threshold at each point in time. Overall, these decisions do not result in much regret. Indeed, many of the enumerated long only decisions appear similar in terms of the next period log wealth utility.

The variance of the regret distributions in Figure (2.6) changes substantially over the investing period. The range of log cumulative wealth difference for the "dense portfolio as target" at the beginning is large ($\sim$ 0.98 to 1.02 on the cumulative wealth scale). The sparse decision for the dense target collapses in variance around 2005 exactly when the sparse decision is very close to the dense portfolio. Notice also that the variance of regret for the sparse decision with SPY as the target is, in general, smaller than the dense target decision. Since all of the enumerated decisions have at least 25% of the portfolio allocated to SPY (the target itself) with other diversifying positions, it is intuitive that the uncertainty in regret should be smaller.

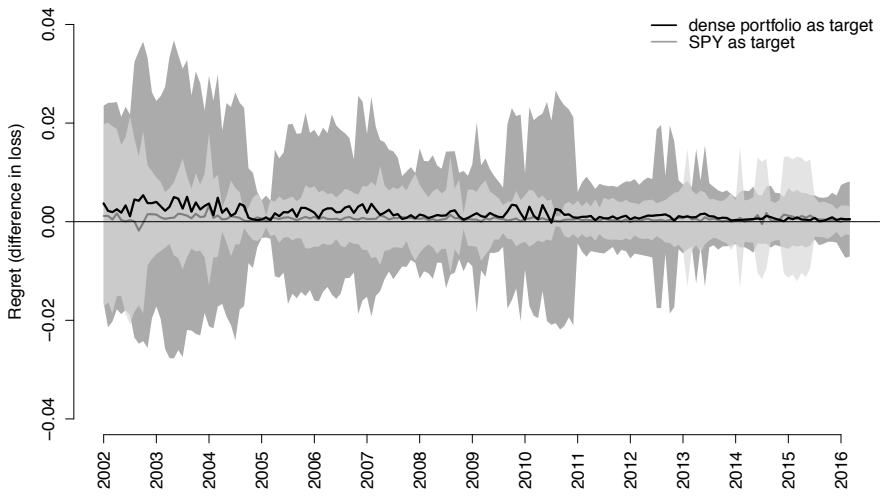


Figure 2.6: The evolution of the ex ante regret distributions for the sparse decisions in Table (2.2) versus the two targets: dense portfolio (black) and SPY (gray). The mean regret is shown by the lines and the surrounding areas represent the evolving centered 60% posterior credible intervals.

The evolution of regret for the sparse decision with SPY as the target sheds light on another question: Are there diversification benefits of allocating to other funds in consideration? Selecting among the 12,950 sparse decisions including up to four non-SPY funds, the expected regret appears to be essentially zero (see the gray line in Figure (2.6)). This analysis suggests that under the log cumulative wealth utility and considering the large set of enumerated decisions defined at the beginning of this section, the best sparse decision from an *ex ante* perspective may be SPY itself!

The ex ante evolution of other metrics, such as the portfolio Sharpe ratio, may be studied for the sparse decisions displayed in Table (2.2). The

Sharpe ratio is not a utility since it is not a function of future returns $\tilde{R}_t$. However, it is a function of our model parameters whose uncertainty is characterized by the posterior distribution. Specifically, define the predictive portfolio Sharpe ratio:

$$\begin{aligned} \mathcal{SR}(w_t, \Theta_t) &= w_t^T \mu_t / (w_t^T \Sigma_t w_t)^{1/2}, \\ \rho_{\mathcal{SR}}(w^*_{\lambda_t}, w^*_t, \Theta_t) &= \mathcal{SR}(w^*_t, \Theta_t) - \mathcal{SR}(w^*_{\lambda_t}, \Theta_t), \end{aligned} \tag{2.15}$$

where $\rho_{\mathcal{SR}}(\cdot)$ is predictive in the sense that future returns $\tilde{R}_t$ conditional on the model parameters are integrated out. This portfolio metric differs from the Kelly criterion loss in that it focuses on a ratio of the portfolio expected return and variance. It may be used as an exploratory tool to accompany selection from regret-based portfolio selection.

We utilize this "difference in Sharpe ratio" distribution in an exploratory fashion in Figure (2.7) shown on an annualized scale. The evolution of the difference in Sharpe ratio is similar to the regret in Figure (2.6). In this case, a larger positive difference in Sharpe ratio means the selected sparse decision possesses a smaller return-risk tradeoff compared to the target decision. The sparse decision for the dense target is larger variance and trends around larger positive values compared with the sparse decision for the SPY target. The rationale for these features is similar: The enumerated sparse decisions are constructed to contain SPY, so the Sharpe ratios (like the loss) of the sparse decision and the SPY target decision will often be close. Following the financial crisis around 2009, the difference in Sharpe ratio stabilizes at lower values for both sparse decisions.

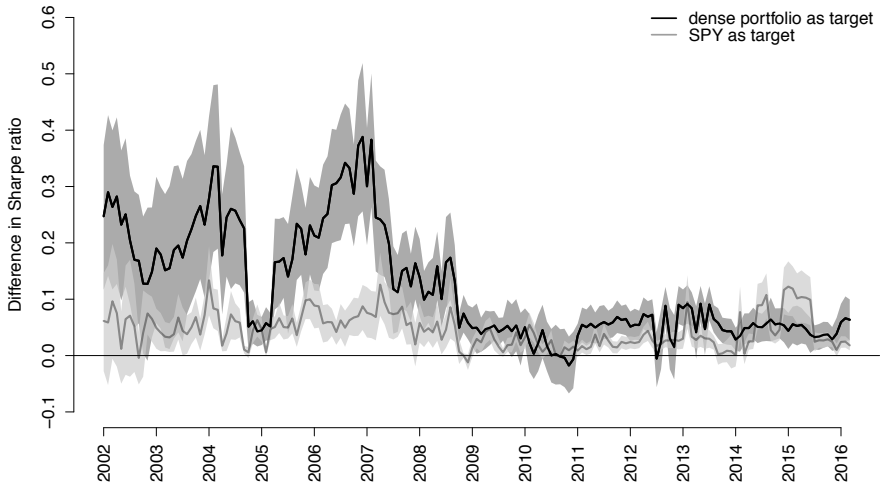


Figure 2.7: The evolution of the ex ante "Difference in annualized Sharpe ratio" distributions for the sparse decisions in Table (2.2) versus the two targets: dense portfolio (black) and SPY (gray). The mean regret is shown by the lines and the surrounding areas represent the evolving centered 60% posterior credible intervals.

2.3.4.1 What happens when κ is varied?

The selection of dynamic portfolio decisions will change based on the regret threshold κ. In Figure (2.8), we show how expected regret and difference in Sharpe ratio (on an annualized scale) change for selected sparse decisions using the SPY target. The evolution of these metrics is shown for sparse decisions constructed using three κ thresholds: $\kappa = 45\%$ (black), 50% (dark gray), and 55% (light gray). The black lines in both figures correspond to the "SPY as target" paths in Figures (2.6) and (2.7), now compared to other κ choices.

Since κ is a lower bound on the satisfaction probability, increasing this lower bound should lead to dynamic sparse decisions with generally smaller regret. In other words, if the investor would like to be satisfied with higher probability, a "lower regret"-sequence of sparse decisions should be selected. Figure (2.8) demonstrates this when SPY is the target. Larger κ generally lead to smaller expected regret and difference in Sharpe ratio paths. The $\kappa = 55\%$ sparse decision leads to expected regret and difference in Sharpe ratio that are mostly negative from 2002 through 2016, indicating that portfolios with SPY and diversifying funds may be preferred to just SPY alone at this high satisfaction threshold. However, these differences in expectation are still close to zero and small, especially for the evolution of expected regret.

2.3.4.2 Enumerated decisions without using the utility

The enumerated decisions considered up to this point are constructed by optimizing the integrated approximate loss. An investor might prefer to construct decisions without any consideration for utility and statistical model. An equal-weighted portfolio (where each of the N assets is given weight $1/N$) is one such example of a "utility agnostic" decision. Financial practitioners often advocate for this decision because of its out of sample performance and purity in not involving errors in the statistical model and optimization procedure [DeMiguel et al., 2007]. The regret-based procedure can readily accommodate a set of decisions with these characteristics as well.

In the following analysis we consider the set of sparse enumerated equal-

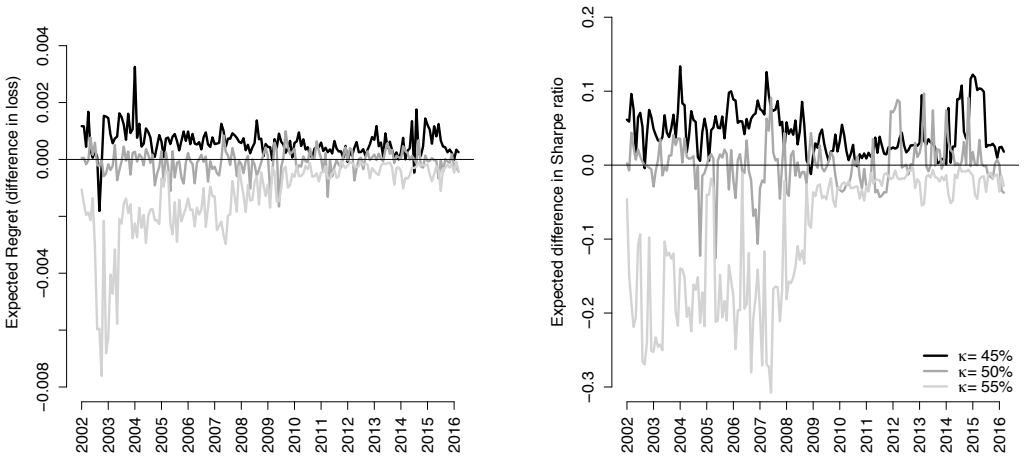


Figure 2.8: Expected regret (left) and expected difference in Sharpe ratio (right) for sparse decisions with SPY as the target. These metrics are shown for three regret thresholds (the lower bound on probability of satisfaction): $\kappa = 45\%$ (black), 50% (dark gray), and 55% (light gray). Note that as the lower bound on the probability of satisfaction increases, both the expected regret and difference in Sharpe ratio tend to decrease for the selected sparse decisions.

weighted portfolios with up to four funds. This amounts to $\Sigma_{i=1}^{4}\binom{25}{i} = 15,275$ decisions to choose among at each point in time. We choose the "dense $1/N$" portfolio as the target decision. This target has 4% invested in each of the 25 funds. To remain consistent with the previous analysis, we consider selection when $\kappa = 45\%$.

The weights for the selected portfolio decision are shown in Table (2.3). At the $\kappa = 45\%$ threshold, all portfolios have either three or four funds included, and the portfolio decisions have sustained exposures to EWJ (Japanese equity) and IYW (technology) throughout the investing period.

Dates	SPY	DIA	EZU	EWU	EWY	EWJ	OEF	EEM	IVE	EFA	IWP	IWR	IWF	IWO	IWM	IYW	XLI
2003	-	-	-	-	-	-	-	-	-	25	25	-	-	-	25	25	-
2004	-	-	-	-	33	33	-	-	-	-	-	-	-	-	-	33	-
2005	-	-	-	-	25	-	25	-	-	-	-	-	-	-	-	25	25
2006	-	-	-	33	-	-	-	-	-	-	-	33	-	-	-	33	-
2007	-	-	-	33	-	-	-	-	-	-	33	-	33	-	-	-	-
2008	-	-	25	-	-	-	25	-	-	-	25	-	25	-	-	-	-
2009	-	33	-	-	-	33	-	-	-	-	-	-	-	-	-	33	-
2010	-	-	-	-	-	33	-	33	-	-	-	-	-	-	-	33	-
2011	-	-	-	-	-	-	-	33	-	33	33	-	-	-	-	-	-
2012	-	-	-	33	33	-	33	-	-	-	-	-	-	-	-	-	-
2013	-	-	-	-	-	-	33	-	33	-	-	-	-	33	-	-	-
2014	33	-	-	-	-	-	-	-	-	33	-	-	-	33	-	-	-
2015	-	-	25	-	-	25	-	-	25	-	-	-	-	-	25	-	-
2016	-	-	-	33	-	33	-	-	33	-	-	-	-	-	-	-	-

Table 2.3: Sparse portfolio decision (in rounded percent) for DLMs (2.5) and (2.6) with $\delta_F = \delta_\epsilon = 0.97$ and $\delta_c = \delta_\beta = 0.9925$. Each point in time represents an equal-weighted portfolio and corresponding λ_t such that the decision satisfies the $\kappa = 45\%$ threshold. The target decision is the equal-weighted portfolio of all 25 funds – also known as the dense $1/N$ portfolio. Note that annual weights are shown for brevity although portfolios are updated monthly.

This approach to equal-weighted portfolio selection possesses innovative and important features that should be highlighted. While traditional "$1/N$" approaches avoid the investor's utility and model for future asset returns altogether, the regret-based procedure still accounts for both. The decisions themselves may be constructed without a utility or model in mind, but the characterization of regret must involve the utility and model. Regret is computed by the difference in *utility* between the target and sparse decisions, and it is a random variable that may be simulated using the *model* for future asset returns. Regardless of the set of decisions considered by the investor, the utility and model will always play a crucial role in a regret-based selection

procedure.

2.3.4.3 Ex post decision analysis

In this section, we consider the realized performance of the three sparse portfolio decisions presented in Tables (2.2) and (2.3) relative to their target decisions. We present out of sample statistics for the six decisions in Table (2.4). Shown is the annualized Sharpe ratio, standard deviation of return, and mean return.

The sparse enumerated decision for the dense target performs similarly to the dense target. This is comforting – this sparse decision is a dynamic portfolio that is allocated to the market (SPY) and at most four other diversifying positions, and its out of sample performance is comparable to the dense portfolio optimized over all 25 funds at each time.

	out of sample statistics		
	Sharpe ratio	s.d.	mean return
sparse enumerated - dense as target	0.40	14.98	6.02
dense	0.45	14.41	6.47
sparse enumerated - SPY as target	0.43	14.65	6.28
SPY	0.43	14.63	6.28
sparse EW enumerated - dense $1/N$ as target	0.49	16.71	8.15
dense $1/N$	0.44	16.47	7.32

Table 2.4: Comparison of out of sample statistics for the six portfolio strategies considered over the investing period February 2002 to May 2016. The three solid lines correspond to the sparse portfolio decisions presented in Tables (2.2) and (2.3). The three dotted lines correspond to the target decisions used for the regret-based selection procedure. All statistics are presented on an annualized scale. "EW" refers to the equal-weighted portfolio decision.

The sparse enumerated decision for the SPY target is equally interesting. Since the SPY target is a sparse decision itself, comparison of it with selected sparse decisions provides insight into whether or not one should diversify away from investing in just the *Market*. The out of sample performance shown in rows three and four of Table (2.4) display similar performance of the sparse enumerated decision and SPY. Even after considering 12,950 sparse decisions containing up to four funds other than SPY – the diversification benefits of exposure beyond SPY are negligible. The decisions that are *ex ante* better than SPY with 45% probability turn out to help out little *ex post*. Note that this conclusion is with respect to the next period cumulative wealth utility. Future work will involve consideration of other utilities and compare how their selection and ex ante/post analyses differ.

While the sparse optimal strategies both underperform their targets, the sparse equal-weighted strategy slightly outperforms its dense $1/N$ target. This is shown in rows five and six of Table (2.4). Interestingly, its out of sample performance even exceeds its sparse optimal counterparts shown in rows one and three in the Table.

2.4 Discussion

This chapter presents a new approach to portfolio selection based on an investor-specified regret tolerance. A loss function defined by the expected portfolio growth rate is used in tandem with a new procedure that separates statistical inference from selection of a sparse dynamic portfolio. We illustrate

the procedure using a set of exchange-traded funds. After analyzing two target decisions: (i) A dense portfolio of all available assets, and (ii) A portfolio comprised of a single market fund, we find that selected sparse decisions differ little from their targets in terms of utility; especially after taking into account uncertainty. This finding persists ex post, and a variety of sparse decisions perform similarly to their target decisions on a risk adjusted return (or Sharpe ratio) basis.

The procedure offers a fresh approach to portfolio selection. While traditional approaches typically focus on either the careful modeling of parameters or the optimization procedure used to calculate portfolio weights, regret-based selection combines both through analysis of the regret random variable. Portfolio decisions that are *parsimonious* in nature are then evaluated in a framework that incorporates *uncertainty* in the investor's utility.

Areas of future research include alternative utility specifications. Two relevant examples are: (i) incorporation of fees and (ii) minimization of transaction costs. In each case, a variant of Loss function (2.10) may be considered. Fees of the funds can be incorporated directly into the vector of future returns. For example, suppose a vector of expense ratios (percentage fee charged of total assets managed) of all funds were given by τ. The vector of future returns within the loss function may be adjusted by τ to reflect an investor's sensitivity to fees:

$$\mathcal{L}(w_t, \tilde{R}_t) = -\log\left(1 + \sum_{k=1}^{N} w_t^k(\tilde{R}_t^k - \tau_k)\right), \tag{2.16}$$

where $\tilde{R}_t^k - \tau_k$ is the net return on investment in for fund k.

Sensitivity to transaction costs can be similarly accounted for by modifying the complexity (penalty) function Φ. This can be accomplished by penalizing the difference in consecutive weight vectors through time, $w_t - w_{t-1}$. An example penalty function would look like:

$$\Phi(\lambda_t^1, \lambda_t^2, w_t) = \lambda_t^1 \left\| w_t \right\|_1 + \lambda_t^2 \left\| w_t - w_{t-1} \right\|_1 . \tag{2.17}$$

This penalty is designed to encourage sparsity as well as slow movement in portfolio positions over time so as to avoid frequent buying and selling of assets. It poses an interesting challenge since there are two penalty parameters (λ_t^1 and λ_t^2) that must be chosen. This is precisely where the regret-based framework has merit. Portfolio decisions indexed by these two penalties can be mapped to a digestible single probability of regret. Then, selection of an appropriate $\{\lambda_t^1, \lambda_t^2\}$ pair can be done in this intuitive "regret space".

The remainder of this section takes a step back and discusses the modularity and important features of regret-based portfolio selection. The methodology is intended to be general – the particular loss, model and dataset used for the empirical analysis are only chosen to demonstrate how the procedure works in practice. The primitive components are: (*i*) a utility function characterizing investor preferences, (*ii*) a complexity function measuring how "simple" a portfolio decision is, (*iii*) a statistical model, and (*iv*) the investor's regret tolerance; where regret is defined as a difference in utility. The regret tolerance stitches together the first two primitives by answering the question: How does

the investor view the *tradeoff* between her utility and portfolio complexity? Using the utility and posterior distribution defined by the statistical model (primitive three), one can construct a mapping between a set of penalty parameters $\{\lambda_t\}$ and probabilities of regret (as displayed by the right vertical axis in Figure (2.2)). However, this is not enough. The collection of portfolio decisions indexed by λ_t must be distilled down to one. The fourth primitive accomplishes this by placing an upper bound on the probability of regret; a portfolio that satisfies this upper bound is selected. By incorporating the four primitives, the main (and surprising) feature of this methodology is that a *static* regret threshold produces a sequence of *dynamic* portfolio decisions, one for each investing period.

Chapter 3

Regularization in Asset Return Models: Seemingly Unrelated Regressions and Monotonic Function Estimation

Analysis and text in this chapter closely follows Puelz et al. [2017]. The first part develop a variable selection approach from seemingly unrelated regression models and applies it to factor selection in asset pricing. The second part considers asset return prediction and monotonic function estimation.

3.1 Introduction and overview

This chapter develops a method for parsimoniously summarizing the shared dependence of many individual response variables upon a common set of predictor variables drawn at random. The focus is on multivariate Gaussian linear models where an analyst wants to find, among p available predictors X, a subset which work well for predicting $q > 1$ response variables Y. The multivariate normal linear model assumes that a set of responses $\{Y_j\}_{j=1}^q$ are linearly related to a shared set of covariates $\{X_i\}_{i=1}^p$ via

$$Y_j = \beta_{j1}X_1 + \cdots + \beta_{jp}X_p + \epsilon_j, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}(0, \Psi), \tag{3.1}$$

where Ψ is a non-diagonal covariance matrix.

Bayesian variable selection in (single-response) linear models is the subject of a vast literature, from prior specification on parameters [Bayarri et al., 2012a] and models [Scott and Berger, 2006] to efficient search strategies over the model space [George and McCulloch, 1993, Hans et al., 2007]. For a more complete set of references, we refer the reader to the reviews of Clyde and George [2004] and Hahn and Carvalho [2015]. By comparison, variable selection has not been widely studied in concurrent regression models, perhaps because it is natural simply to apply existing variable selection methods to each univariate regression individually. Indeed, such joint regression models go by the name "seemingly unrelated regressions" (SUR) in the Bayesian econometrics literature, reflecting the fact that the regression coefficients from each of the separate regressions can be obtained in isolation from one another (i.e., conducting estimation as if Ψ were diagonal). However, allowing non-diagonal Ψ can lead to more efficient estimation [Zellner, 1962] and can similarly impact variable selection [Brown et al., 1998, Wang, 2010].

This chapter differs from Brown et al. [1998] and Wang [2010] in that we focus on the case where all or some of the predictor variables (the regressors, or covariates) are treated as random as opposed to fixed. Our goal will be to summarize codependence among multiple responses in *subsequent* periods, making the uncertainty in future realizations highly central to our selection objective. This approach is natural in many contexts (e.g., macroeconomic models) where the purpose of selection is inherently forward-looking. Measurement errors in the predictors may also contribute to randomness, so this

approach is naturally applicable in an "errors-in-variables" context. To our knowledge, no existing variable selection methods are suitable in this context. The new approach is based on the sparse summary perspective outlined in Hahn and Carvalho [2015], which applies Bayesian decision theory to summarize complex posterior distributions. By using a utility function that explicitly rewards sparse summaries, a high dimensional posterior distribution is collapsed into a more interpretable sequence of sparse point summaries.

A related approach to variable selection in multivariate Gaussian models is the Gaussian graphical model framework [Jones et al., 2005b, Dobra et al., 2004, Wang and West, 2009]. In that approach, the full conditional distributions are represented in terms of a sparse $(p+q)$-by-$(p+q)$ precision matrix. By contrast, we partition the model into response and predictor variable blocks, leading to a distinct selection criterion that narrowly considers the p-by-q covariance between Y to X.

This chapter is structured as follows. Section 2 describes the methodology. A methods overview is presented followed by three subsections discussing the details of the approach. Section 3 presents a simulation study utilizing the methodology with comparisons to alternative approaches. Section 4 demonstrates how the methodology works in practice by considering an application in asset pricing. Section 5 concludes.

3.2 Posterior summary variable selection

3.2.1 Methods overview

Posterior summary variable selection consists of three phases: *model specification and fitting*, *utility specification*, and *graphical summary*. Each of these steps is outlined below. Additional details of the implementation are described in Appendices (1.1) and (1.2).

Step 1: Model specification and fitting

The statistical model may be described compositionally as $p(Y, X) = p(Y|X)p(X)$. For $(Y, X) \sim \mathrm{N}(\mu, \Sigma)$, the regression model (3.1) implies Σ has the following block structure:

$$\Sigma = \left[\begin{array}{c|c} \boldsymbol{\beta}^T\Sigma_x\boldsymbol{\beta} + \Psi & (\Sigma_x\boldsymbol{\beta})^T \\ \hline \Sigma_x\boldsymbol{\beta} & \Sigma_x \end{array}\right]. \tag{3.2}$$

We denote the unknown parameters for the full joint model as $\Theta = \{\mu_x, \mu_y, \Sigma_x, \boldsymbol{\beta}, \Psi\}$ where $\mu = (\mu_y^T, \mu_x^T)^T$ and $\Sigma_x = \mathrm{cov}(X)$.

For a given prior choice $p(\Theta)$, posterior samples of all model parameters are computed by routine Monte Carlo methods, primarily Gibbs sampling. Details of the specific modeling choices and associated posterior sampling strategies are described in Appendix (1.1).

A notable feature of our approach is that *steps 2* (and *3*) will be unaffected by modeling choices made in *step 1* except insofar as they lead to

different posterior distributions. In short, *step 1* is "obtain a posterior distribution"; posterior samples then become inputs to *step 2*.

Step 2: Utility specification

For our utility function we use the log-density of the regression $p(Y|X)$ above. It is convenient to work in terms of negative utility, or loss:

$$\mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}) = \frac{1}{2}(\tilde{Y} - \boldsymbol{\gamma}\tilde{X})^T \Omega (\tilde{Y} - \boldsymbol{\gamma}\tilde{X}),$$

where $\Omega = \Psi^{-1}$. Note that this log-density is being used in a descriptive capacity, not an inferential one; that is, all posterior inferences are based on the posterior distribution from *step 1*. The "action" $\boldsymbol{\gamma}$ is regarded as a point estimate of the regression parameters $\boldsymbol{\beta}$, which would be a good fit to *future* data $(\tilde{Y}, \tilde{X})$ drawn from the same model as the observed data given by $\mathbf{Y} \in \mathbb{R}^{N \times q}$ and $\mathbf{X} \in \mathbb{R}^{N \times p}$.

Taking expectations over the posterior distribution of all unknowns,

$$p(\tilde{Y}, \tilde{X}, \Theta | \mathbf{Y}, \mathbf{X}) = p(\tilde{Y} | \tilde{X}, \Theta) p(\tilde{X} | \Theta) p(\Theta | \mathbf{Y}, \mathbf{X}),$$

yields expected loss

$$\mathcal{L}(\boldsymbol{\gamma}) \equiv \mathbb{E}[\mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma})] = \mathrm{tr}[M \boldsymbol{\gamma} S \boldsymbol{\gamma}^T] - 2\mathrm{tr}[A \boldsymbol{\gamma}^T] + \text{constant},$$

where $A = \mathbb{E}[\Omega \tilde{Y} \tilde{X}^T]$, $S = \mathbb{E}[\tilde{X} \tilde{X}^T] = \overline{\Sigma_x}$, and $M = \overline{\Omega}$, the overlines denote posterior means, and the final term is a constant with respect to $\boldsymbol{\gamma}$.

Finally, we add an explicit penalty, reflecting our preference for sparse summaries:

$$\mathcal{L}_\lambda(\boldsymbol{\gamma}) \equiv \mathrm{tr}[M\boldsymbol{\gamma}S\boldsymbol{\gamma}^T] - 2\mathrm{tr}[A\boldsymbol{\gamma}^T] + \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1, \tag{3.3}$$

where $\left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1$ sums the absolute values of components in $\mathbf{vec}(\boldsymbol{\gamma})$. In practice, it is well known that the ℓ_1 penalty selects relevant components by shrinking irrelevant ones to zero.

Step 3: Graphical summary

Traditional applications of Bayesian decision theory derive *point-estimates* by minimizing expected loss for certain loss functions. The present goal is not an *estimator* per se, but a parsimonious summary of information contained in a complicated, high dimensional posterior distribution. This distinction is worth emphasizing because we have not one, but rather a continuum of loss functions, indexed by the penalty parameter λ. This class of loss functions can be used to summarize the posterior distribution as follows.

Using available convex optimization techniques, expression (3.3) can be optimized efficiently for a range of λ values simultaneously. Posterior graphical summaries consist of two components. First, graphs depicting which response variables have non-zero $\boldsymbol{\gamma}^*_\lambda$ coefficients on which predictor variables can be produced for any given λ. Second, posterior distributions of the quantity

$$\Delta_\lambda = \mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}^*_\lambda) - \mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}^*)$$

can be used to gauge the impact λ has on the descriptive capacity of γ^*_λ. Here, $\boldsymbol{\gamma}^* = \gamma^*_{\lambda=0}$ is the unpenalized optimal solution to the minimization of loss (3.3). Note that these graphs defined by γ^*_λ provide appropriate variable selection for SUR models – different sets of predictors are connected to each of the responses.

The statistical model is given in equations (3.1) and (4.6); prior specification and model fitting details can be found in Appendix (1.1). To briefly summarize, we use a multivariate version of the priors presented in George and McCulloch [1993] and similar to Brown et al. [1998] and Wang [2010] for the exercises in this chapter. We choose the g-prior parameter using an empirical Bayes procedure, and the marginal distribution of the predictors is modeled via a Gaussian linear latent factor model. In the following three subsections, we flesh out the details of *steps 2* and *3*, which represent the main contributions of this chapter.

3.2.2 Deriving the sparsifying expected utility function

Define the optimal posterior summary as the $\boldsymbol{\gamma}^*$ minimizing some expected loss $\mathcal{L}_\lambda(\boldsymbol{\gamma}) = \mathbb{E}[\mathcal{L}_\lambda(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma})]$. Here, the expectation is taken over the joint posterior predictive and posterior distribution: $p(\tilde{Y}, \tilde{X}, \Theta \mid \mathbf{Y}, \mathbf{X})$.

As described in the previous section, our loss takes the form of a penalized log conditional distribution:

$$\mathcal{L}_\lambda(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}) \equiv \frac{1}{2}(\tilde{Y} - \boldsymbol{\gamma}\tilde{X})^T \Omega (\tilde{Y} - \boldsymbol{\gamma}\tilde{X}) + \lambda \left\| \mathbf{vec}(\boldsymbol{\gamma}) \right\|_1, \qquad (3.4)$$

where $\Omega = \Psi^{-1}$, $\|\mathbf{vec}(\boldsymbol{\gamma})\|_1 = \sum_j |\mathbf{vec}(\boldsymbol{\gamma})_j|$, and $\mathbf{vec}(\boldsymbol{\gamma})$ is the vectorization of the action matrix $\boldsymbol{\gamma}$. The first term of this loss measures the distance (weighted by the precision Ω) between the linear predictor $\boldsymbol{\gamma}\tilde{X}$ and a future response $\tilde{Y}$. The second term promotes a sparse optimal summary, $\boldsymbol{\gamma}$. The penalty parameter λ determines the relative importance of these two components. Expanding the quadratic form gives:

$$\begin{aligned}\mathcal{L}_\lambda(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}) &= \frac{1}{2}\left(\tilde{Y}^T\Omega\tilde{Y} - 2\tilde{X}^T\boldsymbol{\gamma}^T\Omega\tilde{Y} + \tilde{X}^T\boldsymbol{\gamma}^T\Omega\boldsymbol{\gamma}\tilde{X}\right) + \lambda\|\mathbf{vec}(\boldsymbol{\gamma})\|_1 \\ &= \left(\tilde{X}^T\boldsymbol{\gamma}^T\Omega\boldsymbol{\gamma}\tilde{X} - 2\tilde{X}^T\boldsymbol{\gamma}^T\Omega\tilde{Y}\right) + \lambda\|\mathbf{vec}(\boldsymbol{\gamma})\|_1 + \text{constant}.\end{aligned}$$

Integrating over $(\tilde{Y}, \tilde{X}, \Theta \mid \mathbf{Y}, \mathbf{X})$ (and dropping the constant) gives:

$$\begin{aligned}\mathcal{L}_\lambda(\boldsymbol{\gamma}) &= \mathbb{E}[\mathcal{L}_\lambda(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma})] \\ &= \mathbb{E}\left[\mathrm{tr}[\boldsymbol{\gamma}^T\Omega\boldsymbol{\gamma}\tilde{X}\tilde{X}^T]\right] - 2\mathbb{E}\left[\mathrm{tr}[\boldsymbol{\gamma}^T\Omega\tilde{Y}\tilde{X}^T]\right] + \lambda\|\mathbf{vec}(\boldsymbol{\gamma})\|_1, \\ &= \mathbb{E}\left[\mathrm{tr}[\boldsymbol{\gamma}^T\Omega\boldsymbol{\gamma}S]\right] - 2\mathrm{tr}[A\boldsymbol{\gamma}^T] + \lambda\|\mathbf{vec}(\boldsymbol{\gamma})\|_1, \\ &= \mathrm{tr}[M\boldsymbol{\gamma}S\boldsymbol{\gamma}^T] - 2\mathrm{tr}[A\boldsymbol{\gamma}^T] + \lambda\|\mathbf{vec}(\boldsymbol{\gamma})\|_1,\end{aligned} \tag{3.5}$$

where:

$$\begin{aligned}A &\equiv \mathbb{E}[\Omega\tilde{Y}\tilde{X}^T], \\ S &\equiv \mathbb{E}[\tilde{X}\tilde{X}^T] = \overline{\Sigma}_x, \\ M &\equiv \overline{\Omega},\end{aligned} \tag{3.6}$$

and the overlines denote posterior means. Define the Cholesky decompositions $M = LL^T$ and $S = QQ^T$. Expression (3.5) can be formulated in the form of a standard penalized regression problem:

$$\mathcal{L}_\lambda(\boldsymbol{\gamma}) = \left\|\left[Q^T \otimes L^T\right]\mathbf{vec}(\boldsymbol{\gamma}) - \mathbf{vec}(L^{-1}AQ^{-T})\right\|_2^2 + \lambda\|\mathbf{vec}(\boldsymbol{\gamma})\|_1, \tag{3.7}$$

with "covariates" $Q^T \otimes L^T$, "data" $L^{-1}AQ^{-T}$, and regression coefficients $\boldsymbol{\gamma}$ (see Appendix 1.2 for details). Accordingly, (3.7) can be optimized using existing software such as the `lars` R package of Efron et al. [2004a] and still yield sparse solutions.

The `lars` formulation of the utility function provides fast computation as well as flexibility. For example, suppose we wish to always include certain (potentially different) predictors in each SUR equation. This can be easily achieved by removing the ℓ_1 penalty on the relevant components of $\mathbf{vec}(\boldsymbol{\gamma})$ (since $\boldsymbol{\gamma}$ represents the dependence structure between the responses and predictors) by zeroing the appropriate penalty parameters λ.

3.2.2.1 What if only a subset of predictors are random?

Building on the previous derivation, we consider a scenario where some predictors are known, or fixed, and the remainder are random. This may occur when one would like to condition on a particular value of a predictor at some fixed future value. In this case, an expected utility function can be derived in a similar manner to the random predictors case.

Let the covariates X be divided into two pieces, those that are considered random: $X_r \in \mathbb{R}^{p_r}$, and those that are considered fixed: $X_f \in \mathbb{R}^{p_f}$, so that the column vector $X = [X_r^T \; X_f^T]^T \in \mathbb{R}^p$ and $p = p_r + p_f$. So, future values of the covariates are given by $\tilde{X} = [\tilde{X}_r^T \; X_f^T]^T$.

Conditioning on the fixed covariates, the distribution of unknowns is: $p(\tilde{Y}, \tilde{X}_r, \Theta | X_f)$ where Θ is a vector of parameters from a specified model. If

we assume conditional independence, then we can write:

$$p(\tilde{Y}, \tilde{X}_r, \Theta|X_f) = p(\tilde{Y}|\tilde{X}_r, X_f, \Theta)p(\tilde{X}_r|X_f, \Theta)p(\Theta|X_f).$$

where, as before, $p(\Theta|X_f)$ is the posterior distribution of model parameters conditional on the fixed covariates. Following *step 1* of the methodology, any models may be chosen for the conditional $Y|X_r, X_f$ and the marginal $X_r|X_f$. For example, in the case of X following a multivariate normal distribution implied by a latent factor regression model, we automatically know the conditionals including $X_r|X_f$.

Define the following block structure for the action, $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_r & \boldsymbol{\gamma}_f \end{bmatrix},$$

so that $\boldsymbol{\gamma}_r \in \mathbb{R}^{q \times p_r}$ and $\boldsymbol{\gamma}_f \in \mathbb{R}^{q \times p_f}$. We expand out (3.4) and drop terms that don't involve the action $\boldsymbol{\gamma}$:

$$\begin{aligned}\mathcal{L}_\lambda(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}) =& \frac{1}{2}\left(\tilde{X}_r^T\boldsymbol{\gamma}_r^T\Omega\boldsymbol{\gamma}_r\tilde{X}_r + X_f^T\boldsymbol{\gamma}_f^T\Omega\boldsymbol{\gamma}_f X_f - 2\tilde{X}_r^T\boldsymbol{\gamma}_r^T\Omega\tilde{Y} - 2X_f^T\boldsymbol{\gamma}_f^T\Omega\tilde{Y}\right) \\ &+ \lambda\left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1 + \text{constant}.\end{aligned}$$

Taking expectations over $p(\tilde{Y}, \tilde{X}_r, \Theta|X_f)$ and dropping the one-half and constant, we obtain the integrated loss function:

$$\begin{aligned}\mathcal{L}_\lambda(\boldsymbol{\gamma}) =& \mathbb{E}\left[\text{tr}[\boldsymbol{\gamma}_r^T\Omega\boldsymbol{\gamma}_r\tilde{X}_r\tilde{X}_r^T]\right] - 2\mathbb{E}\left[\text{tr}[\boldsymbol{\gamma}_r^T\Omega\tilde{Y}\tilde{X}_r^T]\right] + \mathbb{E}\left[\text{tr}[\boldsymbol{\gamma}_f^T\Omega\boldsymbol{\gamma}_f X_f X_f^T]\right] \\ &- 2\mathbb{E}\left[\text{tr}[\boldsymbol{\gamma}_f^T\Omega\tilde{Y}X_f^T]\right] + \lambda\left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1.\end{aligned}$$

We simplify the expectations in a similar way to our derivation of the original loss function presented at the beginning of Section (3.2.2).

$$\begin{aligned}\mathcal{L}_\lambda(\boldsymbol{\gamma}) =& \text{tr}[M\boldsymbol{\gamma}_r S_r \boldsymbol{\gamma}_r^T] - 2\text{tr}[A_r\boldsymbol{\gamma}_r^T] + \text{tr}[M\boldsymbol{\gamma}_f S_f \boldsymbol{\gamma}_f^T] - 2\text{tr}[A_f\boldsymbol{\gamma}_f^T] \\ &+ \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1,\end{aligned}$$

where:

$$\begin{aligned}A_r &\equiv \mathbb{E}[\Omega\tilde{Y}\tilde{X}_r^T], \quad A_f \equiv \mathbb{E}[\Omega\tilde{Y}\tilde{X}_f^T] \\ S_r &\equiv \mathbb{E}[\tilde{X}_r\tilde{X}_r^T], \quad S_f = X_f X_f^T \\ M &\equiv \overline{\Omega}\end{aligned}$$

Combining the matrix traces, we simplify the loss function as follows:

$$\mathcal{L}_\lambda(\boldsymbol{\gamma}) = \text{tr}[M\boldsymbol{\gamma}\dot{S}\boldsymbol{\gamma}^T] - 2\text{tr}[\dot{A}\boldsymbol{\gamma}^T] + \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1,$$

where:

$$\dot{S} \equiv \begin{bmatrix} S_r & 0 \\ 0 & S_f \end{bmatrix}, \quad \dot{A} \equiv \begin{bmatrix} A_r \\ A_f \end{bmatrix}.$$

This form is analogous to the loss function derived when all predictors are assumed random, now for a case when fixed *and* random predictors are present. This can similarly be formulated into a penalized regression problem. The full derivation of the lasso form of this problem is presented in Appendix (1.2).

The effect on *step 3* of assuming only a subset of predictors are random is intuitive. Less statistical uncertainty will propagate into the Δ_λ metric, and the algorithm will favor denser graphs. This will be shown in the results section where we study two extremes: all random predictors and all fixed predictors.

3.2.3 Sparsity-utility trade-off plots

Rather than attempting to determine an "optimal" value of λ, we advocate displaying plots that reflect the utility attenuation due to λ-induced sparsification. We define the "loss gap" between a λ-sparse solution, $\mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}^*_\lambda)$, and the optimal unpenalized (non-sparse, $\lambda = 0$) summary, $\mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}^*)$ as

$$\Delta_\lambda = \mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}^*_\lambda) - \mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}^*).$$

As a function of $(\tilde{Y}, \tilde{X}, \Theta)$, Δ_λ is itself a random variable which we can sample by obtaining posterior draws from $p(\tilde{Y}, \tilde{X}, \Theta \mid \mathbf{Y}, \mathbf{X})$. The posterior distribution(s) of Δ_λ (for various λ) therefore reflects the deterioration in utility attributable to "sparsification". Plotting these distributions as a function of λ allows one to visualize this trade-off. Specifically, $\pi_\lambda \equiv \Pr(\Delta_\lambda < 0 \mid \mathbf{Y}, \mathbf{X})$ is the (posterior) probability that the λ-sparse summary is no worse than the non-sparse summary. Using this framework, a useful heuristic for obtaining a single sparse summary is to report the sparsest model (associated with the highest λ) such that π_λ is higher than some pre-determined threshold, κ; we adopt this approach in our application section.

We propose summarizing the posterior distribution of Δ_λ via two types of plots. First, one can examine posterior means and credible intervals of Δ_λ

for a sequence of models indexed by λ. Similarly, one can plot π_λ across the same sequence of models. Also, for a fixed value of λ, one can produce graphs where nodes represent predictor variables and response variables and an edge is drawn between nodes whenever the corresponding element of γ_λ^* is non-zero. All three types of plots are exhibited in Section (3.4).

3.2.4 Relation to previous methods

Loss function (3.7) is similar in form to the univariate *DSS* (decoupled shrinkage and selection) strategy developed by Hahn and Carvalho [2015]. Our approach generalizes Hahn and Carvalho [2015] by optimizing over the matrix $\boldsymbol{\gamma} \in \mathbb{R}^{qxp}$ rather than a single vector of regression coefficients, extending the sparse summary utility approach to seemingly unrelated regression models [Brown et al., 1998, Wang, 2010]. Additionally, the present method considers random predictors, $\tilde{X}$, whereas Hahn and Carvalho [2015] considered only a matrix of fixed design points. The impact of accounting for random predictors on the posterior summary variable selection procedure is examined in more detail in the application section.

To the best of our knowledge, the most comparable method for analyzing sparse linear covariance structures are the SUR models described in Wang [2010], which utilize independent point-mass priors for each element of $\boldsymbol{\beta}$. Our method differs from this approach for the following reasons. A sparse SUR model provides posterior draws of the coefficient matrix $\boldsymbol{\beta}$, but as in the simpler linear regression case described in Hahn and Carvalho [2015], obtaining

a sparse summary is non-trivial. Two common approaches for extracting a summary from posterior draws of a sparse SUR model are either to report a maximum a posteriori estimate (MAP) or to hard-threshold posterior inclusion probabilities of matrix components of $\boldsymbol{\beta}$ describing the model sparsity pattern. Neither approach is fully satisfactory; the MAP estimate is not well-motivated if the goal is future prediction and approaches based on thresholding the edge inclusion probabilities fail to take into account co-dependence between individual edges coming in and out of the model together. By contrast, our method begins with a principled loss function; by focusing on the expected log-density of future predictions, our approach synthesizes information from all model parameters simultaneously in gauging how important they are for prediction. A comprehensive simulation comparing inclusion probability thresholding and our approach is presented in Section (3.3).

3.3 Simulation study

In this section, we present a simulation study to compare our posterior model selection summary to that of the median (posterior) probability model (using the model of Wang [2010]) for a fixed data generating process (DGP). This comparison is not meant to demonstrate the superiority of one method over the other, but rather to highlight that the two methods can give quite distinct summaries. More specifically, we observe that the median probability model (MPM) can differ substantially from our penalized utility summary when the predictor variables are highly correlated.

The data generating process is the following two equation seemingly unrelated regression model:

$$\begin{aligned} X &\sim N(0, \Sigma_x^{\text{sim}}), \\ Y|X &\sim N(\boldsymbol{\beta}_{\text{sim}}^T X, \Psi^{\text{sim}}), \end{aligned} \tag{3.8}$$

where:

$$\Sigma_x^{\text{sim}} = \begin{bmatrix} 1 & 0 & 0.9 & 0 & 0 \\ 0 & 1 & 0 & 0.8 & -0.3 \\ 0.9 & 0 & 1 & 0 & 0 \\ 0 & 0.8 & 0 & 1 & 0 \\ 0 & -0.3 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta}_{\text{sim}} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \Psi^{\text{sim}} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Figure (3.1) graphically depicts the association structure encoded by $\boldsymbol{\beta}^{\text{sim}}$. In words, Y_1 is related to $\{X_1, X_2\}$ and Y_2 is related to $\{X_1, X_3, X_4\}$. Estimation of $\boldsymbol{\beta}^{\text{sim}}$ is complicated by the large positive correlation between $\{X_1, X_3\}$ and $\{X_2, X_4\}$ and negative correlation between $\{X_2, X_5\}$, as well as the non-diagonality of the residual variance Ψ^{sim}. We simulate 500 data sets, each with 50 samples of (Y, X).

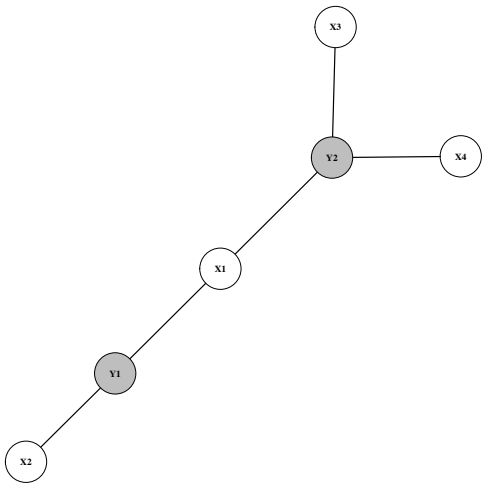
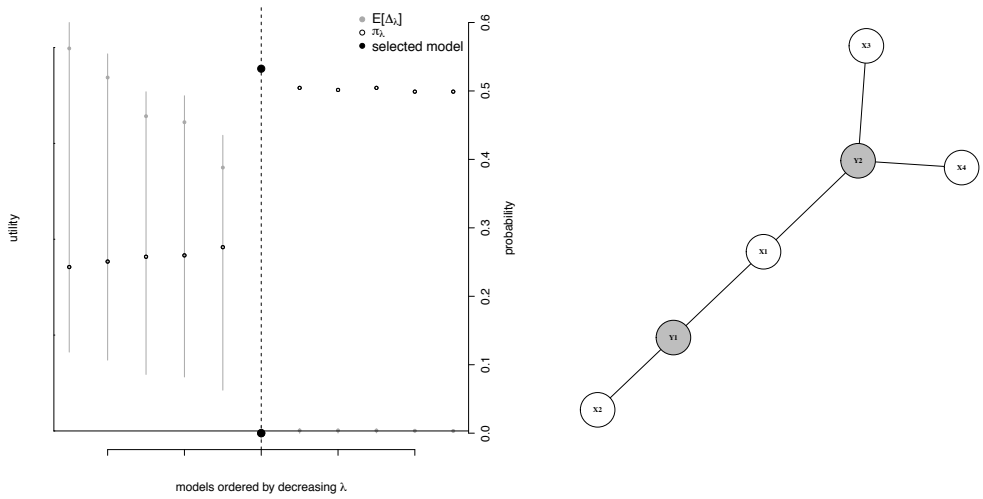


Figure 3.1: Graphical representation of the true graph defined by the coefficient matrix $\boldsymbol{\beta}^{\text{sim}}$.

As outlined in Section (3.2), the first step of our analysis consists of fitting a Bayesian model. We fit model (3.1) using a multivariate version of the priors presented in George and McCulloch [1993] and similar to Brown et al. [1998] and Wang [2010]. Specifically, we use conjugate normal g-priors on the regression coefficients and choose the g parameter by an empirical Bayes procedure. The marginal distribution of the predictors is modeled via a Gaussian latent factor model. Details for all models and fitting algorithms are given in Appendix 1.1. Note that this is one of many such priors a researcher may choose.

Figures (3.2) displays an example solution path for one of the data sets

that selected the true model. The left axis is on the utility scale and shows the Δ_λ metric for decreasing penalty parameters λ. The right axis show the probability that the sparsified model is no worse than the saturated model π_λ. Posterior summary variable selection correctly identified the true model in 258 out of 500 simulated data sets at a 20% posterior uncertainty interval. Notice that the true model has a considerable jump in utility and π_λ on the left plot in Figures (3.2). In addition, the true model is contained in 400 out of the 500 posterior summary variable solution paths. This implies that our utility chooses the true model as the *best* 5-edge graph out of all possible graphs of equal size in 90% of the simulated data sets.



Figures 3.2: **(left)** Example of evaluation of Δ_λ and π_λ along the solution path for one of the simulated data sets where the true graph was correctly selected. Uncertainty bands are 20% posterior intervals on the Δ_λ metric. The large black dot and associated dashed line represent the graph selected and shown on the right. **(right)** The most selected graph for simulated data. This is the true graph given by $\boldsymbol{\beta}^{\text{sim}}$ and was selected for 258 out of the 500 simulated data sets and is present in 400 out of 500 posterior summary solution paths. The responses and predictors are colored in gray and white, respectively. Edges represent nonzero components of the optimal action, $\boldsymbol{\gamma}$.

Next, we compare these results to a related method, based on the model of Wang [2010], which extends the stochastic search variable selection of George and McCulloch [1993] to seemingly unrelated regression models. Their model allows for sparsity in the $\boldsymbol{\beta}$ matrix and provides an inclusion probability for each entry in the matrix representing an edge in the graph. Details of the specific priors used may be found in Wang [2010]. Although this model uses point-mass priors and produces posterior samples across various sparse

regressions, the most common posterior summary, the (arithmetical) mean, generally produces *non-sparse* summaries. By contrast, among widely used summaries, the median probability model does provide a sparse point summary. In Barbieri and Berger [2004], the median probability model is shown to be optimal under squared error prediction loss, as long as the predictor variables are mutually orthogonal. In practice, the median probability model is defined as the model containing all and only those variables with marginal inclusion probabilities greater than 1/2, whether or not the orthogonality condition is satisfied.

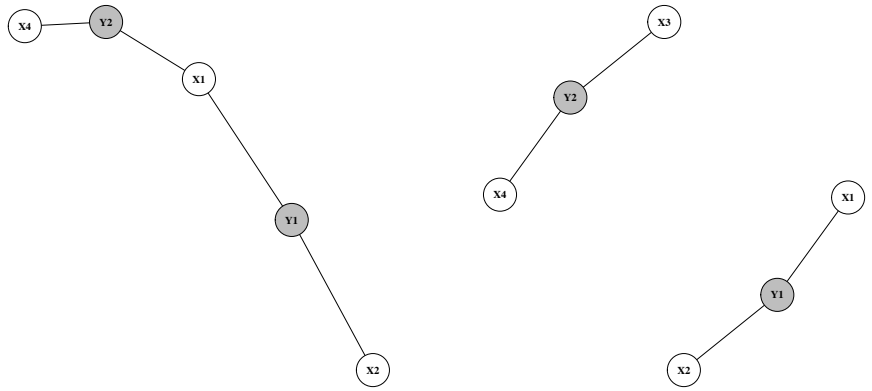
	covariate				
response	X_1	X_2	X_3	X_4	X_5
Y_1	0.8540	0.9588	0.1381	0.0335	0.0039
Y_2	0.5511	0.0467	0.5779	0.9492	0.0097

Table 3.1: Average edge inclusion probabilities for the $\boldsymbol{\beta}$ matrix across the 500 simulated data sets.

Table (3.1) shows the average edge inclusion probabilities for the $\boldsymbol{\beta}$ matrix across the 500 simulated data sets. Edge $\{Y_1, X_3\}$ was slightly sampled less than others, but the correlation in the predictors prohibits the inference from accurately ruling this out completely. The strong dependence between $\{X_1, X_3\}$ effects the inconclusive sampling of edges $\{Y_2, X_1\}$ and $\{Y_2, X_3\}$ as well – often when one edge is sampled, the other is excluded. Overall, the inclusion probabilities vary widely depending on the simulated data set.

Figures (3.3) depicts the most common MPMs across the 500 data sets.

The left graph was selected in 181 out of the 500 simulated data sets, and the right graph was selected in 149 out of the 500 data sets. The true graph shown in Figure (3.1) was only selected in 55 out of the 500 data sets.



Figures 3.3: The two most frequently appearing median probability models from the sparse SUR inference on each of the 500 simulated data sets. The left graph was selected in 181 out of the 500 simulated data sets, and the right graph was selected in 149 out of the 500 data sets.

To obtain a sense of how dissimilar the MPM summary can be compared to our approach, we tally how often the MPM appears in the posterior summary solution path displayed in, for example, the left side of Figures (3.2). A selected graph depends on the posterior uncertainty interval of Δ_λ; where a larger interval leads to sparser graphs. Therefore, if the MPM is contained in the solution path, it is a desirable model under our utility function modulo a degree of statistical uncertainty. We find that the MPM is contained in only

241 out of the 500 posterior summary solution paths. In other words, in 259 out of 500 solution paths, our utility function prefers a *different* model over the MPM of equal size (where size is measured by number of of edges). Of the 241 occasions that they coincided, 55 of those recovered the true structure. We speculate that the difference between the two approach is largely due to strong correlation between predictors X_1 and X_3; our utility function explicitly considers this structure whereas the MPM formulation does not. In a similar simulation study with orthogonal predictors, the MPM recovers the true sparse structure in 488 out of 500 simulated data sets.

3.4 Applications

In this section, the sparse posterior summary method is applied to a data set from the finance (asset pricing) literature. A key component of our analysis will be a comparison between the posterior summaries obtained when the predictors are drawn at random versus when they are assumed fixed.

The response variables are returns on 25 tradable portfolios and our predictor variables are returns on 10 other portfolios thought to be of theoretical importance. In the asset pricing literature [Ross, 1976], the response portfolios represent assets to be priced (so-called *test assets*) and the predictor portfolios represent distinct sources of variation (so-called *risk factors*). More specifically, the test assets Y represent trading strategies based on company size (total value of stock shares) and book-to-market (the ratio of the company's accounting valuation to its size); see Fama and French [1992] and

Fama and French [2015] for details. Roughly, these assets serve as a lower-dimensional proxy for the stock market. The risk factors are also portfolios, but ones which are thought to represent *distinct* sources of risk. What constitutes a distinct source of risk is widely debated, and many such factors have been proposed in the literature [Cochrane, 2011]. Moreover, finding a small subset of factors (even from these 10) is useful for a finance researcher by providing *ease of interpretation*. If 3 factors are good enough predictively *and* easier for the finance researcher to grasp mentally, then this dimension reduction is useful; even in this moderately sized problem. We use monthly data from July 1963 through February 2015, obtained from Ken French's website:

`http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/`.

Our analysis investigates which subset of risk factors are most relevant (as defined by our utility function). As our initial candidates, we consider factors known in previous literature as: market, size, value, direct profitability, investment, momentum, short term reversal, long term reversal, betting against beta, and quality minus junk. Each factor is constructed by cross-sectionally sorting stocks by various characteristics of a company and forming linear combinations based on these sorts. For example, the value factor is constructed using the book-to-market ratio of a company. A high ratio indicates the company's stock is a "value stock" while a low ratio leads to a "growth stock" assessment. Essentially, the value factor is a portfolio built by going long stocks with high book-to-market ratio and shorting stocks with low book-to-market ratio. For detailed definitions of the first five factors, see Fama and

French [2015]. In the figures to follow, each is labeled as, for example, "Size2 BM3," to denote the portfolio buying stocks in the second quintile of size and the third quintile of book-to-market ratio.

Recent related work includes Ericsson and Karlsson [2004] and Harvey and Liu [2015]. Ericsson and Karlsson [2004] follow a Bayesian model selection approach based off of inclusion probabilities, representing the preliminary inference step of our methodology. Harvey and Liu [2015] take a different approach that utilizes multiple hypothesis testing and bootstrapping.

3.4.1 Results

As outlined in Section (3.2.1), the first step of our analysis consists of fitting a Bayesian model. We fit model (3.1) using a multivariate version of the priors presented in George and McCulloch [1993] and similar to Brown et al. [1998] and Wang [2010]. Specifically, we use conjugate normal g-priors on the regression coefficients and choose the g parameter by an empirical Bayes procedure. The marginal distribution of the predictors are modeled via a Gaussian latent factor model. Note that this is one of many such priors a researcher may choose. The advantage of posterior summary variable selection is that any reasonable statistical model for the joint (Y, X) may be chosen.

Recalling the block structure for the covariance of the full joint distribution of (Y, X) from expression (4.6) we obtain posterior samples of Σ by sampling the conditional model parameters using a matrix-variate stochastic search algorithm (described below) and sampling the covariance of X from a

latent factor model where it is marginally normally distributed. To reiterate our procedure is

- Σ_x is sampled from independent latent factor model,
- $\boldsymbol{\beta}$ is sampled from matrix-variate MCMC,
- Ψ is sampled from matrix-variate MCMC.

The conditional model for $Y|X$ also includes the sampling of an indicator variable α that records if a given variable is non-zero (included in the model). In our simulation and application results, we fix α to the identity vector. This is done to emphasize that even when dense models are sampled in the inference step, our procedure has the ability to select a sparse set of predictors. Details of the model fitting algorithm may be found in Appendix (1.1).

In the subsections to follow, we will show a panel consisting of two figures. First, we plot the expectation of Δ_λ (and associated posterior credible interval) across a range of λ penalties. Recall, Δ_λ is the "loss gap" between a sparse summary and the best non-sparse (saturated) summary, meaning that smaller values are "better". Additionally, we plot the probability that a given model is no worse than the saturated model π_λ on this same figure, where "no worse" means $\Delta_\lambda < 0$. Note that even for very weak penalties (small λ), the distribution of Δ_λ will have non-zero variance and therefore even if it is

centered about zero, some mass can be expected to fall above zero; practically, this means that $\pi_\lambda > 0.5$ is a very high score.

Second, we display a summary graph of the selected variables for the $\kappa = 12.5\%$ threshold. Recall that this is the highest penalty (sparsest graph) that is no worse than the saturated model with 12.5% posterior probability. For these graphs, the response and predictor variables are colored gray and white, respectively. A test asset label of, for example, "Size2 BM3," denotes the portfolio that buys stocks in the second quintile of size and the third quintile of book-to-market ratio. The predictors without connections to the responses under the optimal graph are not displayed.

These panels of two figures are shown in two scenarios:

1. Random predictors.

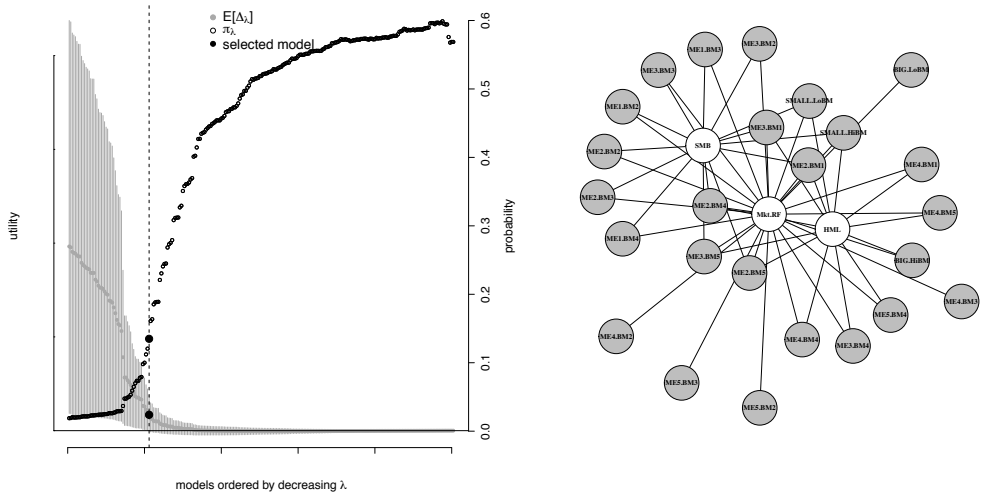
2. Fixed predictors.

3.4.1.1 Random predictors

This section introduces our baseline example where the risk factors (predictors) are random. We evaluate the set of potential models by analyzing plots such as the left plot in Figures (3.4). This shows Δ_λ and π_λ evaluated across a range of λ values. Additionally, we display the posterior uncertainty in the Δ_λ metric with gray vertical uncertainty bands; these are the centered $P\%$ posterior credible intervals where $\kappa = (1 - P)/2$. As the accuracy of the sparsified solution increases, the posterior of Δ_λ concentrates around zero

by construction, and the probability of the model being no worse than the saturated model, π_λ, increases. We choose the sparsest model such that its corresponding $\pi_\lambda > \kappa = 12.5\%$. This model is displayed on the right in Figures (3.4) – also referred to as the "graphical summary".

The selected set of factors are the market (Mkt.RF), value (HML), and size (SMB). This three factor model is no worse than the saturated model with 12.5% posterior probability where all test assets are connected to all risk factors. Note also that in our selected model almost every test asset is distinctly tied to one of either value or size and the market factor. These are the three factors of Ken French and Eugene Fama's pricing model developed in Fama and French [1992]. They are known throughout the finance community as being "fundamental dimensions" of the financial market, and our procedure is consistent with this widely held belief at a small κ level.



Figures 3.4: **(left)** Evaluation of Δ_λ and π_λ along the solution path for the 25 size/value portfolios modeled by the 10 factors. An analyst may use this plot to select a particular model. Uncertainty bands are 75% posterior intervals on the Δ_λ metric. The large black dot and associated dashed line represents the model selected and shown on the right. **(right)** The selected model for 25 size/value portfolios modeled by the 10 factors. The responses and predictors are colored in gray and white, respectively. Edges represent nonzero components of the optimal action, $\boldsymbol{\gamma}$.

The characteristics of the test assets in the selected graph from Figure (3.4) are also important to highlight. The test portfolios that invest in small companies ("Size1" and "Size2") are primarily connected to the SMB factor which is designed as a proxy for the risk of small companies. Similarly, the test portfolios that invest in high book-to-market companies ("BM4" and "BM5") have connections to the HML factor which is built on the idea that companies whose book value exceeds the market's perceived value should generate a distinct source of risk. As previously noted, all of the test portfolios are

connected to the market factor suggesting that it is a relevant predictor even for the sparse $\kappa = 12.5\%$ selection criterion.

In Figure (3.5), we examine how different choices of the κ threshold change the selected set of risk factors. In this analysis, there is a trade-off between the posterior probability of being "close" to the saturated model and the utility's preference for sparsity. When the threshold is low ($\kappa = 2$ and 12.5%) the summarization procedure selects relatively sparse graphs with up to three factors (Mkt.RF, HML, and SMB). The market (Mkt.RF) and size (SMB) factors appear first, connected to a small number of the test assets ($\kappa = 2\%$). As the threshold is increased, the point summary becomes denser and correspondingly more predictively accurate (as measured by the utility function). The value factor (HML) enters at $\kappa = 12.5\%$ and quality minus junk (QMJ), investment (CMA), and profitability (RMW) factors enter at $\kappa = 32.5\%$. The graph for $\kappa = 32.5\%$ excluding QMJ is essentially the new five factor model proposed by Fama and French [2015]. The five Fama-French factors (plus OMJ, BAB and LTR with a small number of connections) persist up to the $\kappa = 47.5\%$ threshold. This indicates that, up to a high posterior probability, the five factor model of Fama and French [2015] does no worse than an asset pricing model with all ten factors connected to all test assets.

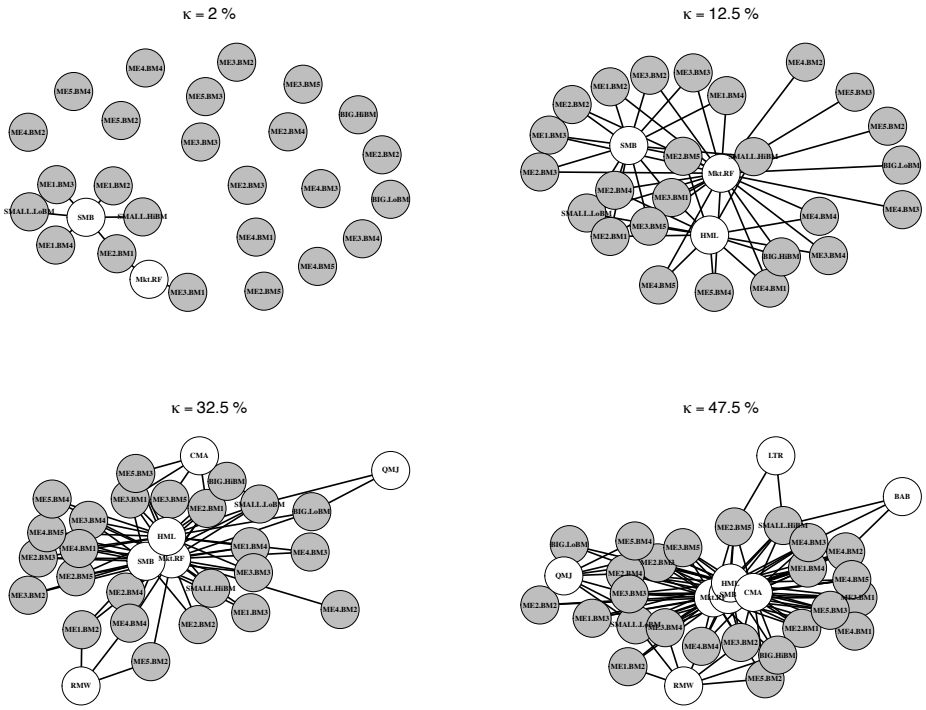


Figure 3.5: Sequence of selected models for varying threshold level κ under the assumption of **random predictors**.

Notice also that our summarization procedure displays the specific relationship between the factors and test assets through the connections. Using this approach, the analyst is able to identify which predictors drive variation in which responses and at what thresholds they may be relevant. This feature is significant for summarization problems where individual characteristics of the test portfolios and their joint dependence on the risk factors may be a priori unclear.

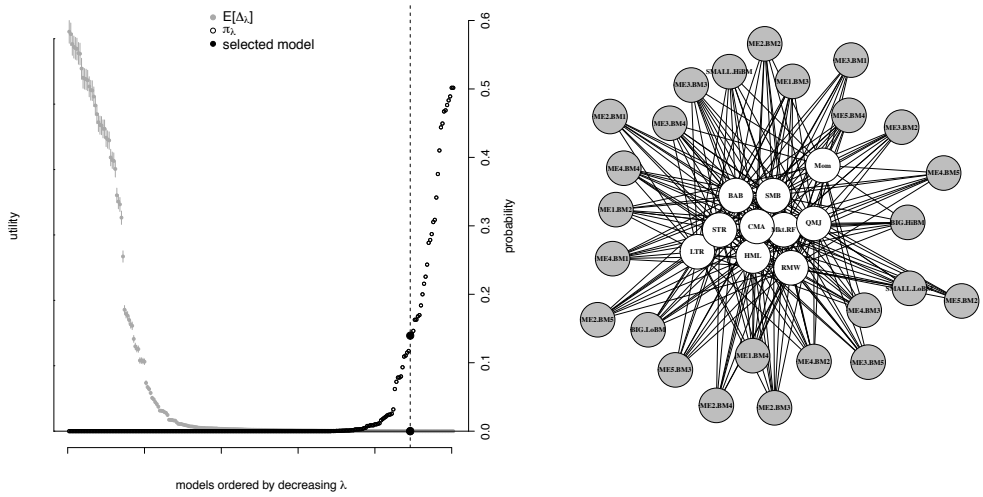
As κ approaches the 50% threshold ($\kappa = 47.5\%$ in Figure (3.5), the model summary includes eight of ten factors. Requesting a summary with this level of certainty results in little sparsification. However, an additional contribution of a factor results in minor increases in out utility. Sparse posterior summarization applied in this context allows an analyst to study the impact of risk factors on pricing while taking uncertainty into account. Coming to a similar conclusion via common alternative techniques (e.g., component-wise ordinary least squares combined with thresholding by t-statistics) is comparatively ad hoc; our method is simply a perspicuous summary of a posterior distribution. Likewise, applying sparse regression techniques based on ℓ_1 penalized likelihood methods would not take into account the residual correlation Ψ, nor would that approach naturally accommodate random predictors.

3.4.1.2 Fixed predictors

In this section, we consider posterior summarization with the loss function derived under the assumption of *fixed predictors*. The analogous loss function when the predictor matrix is fixed at pre specified points $\mathbf{X}$ is:

$$\mathcal{L}_\lambda(\boldsymbol{\gamma}) = \left\| \left[Q_f^T \otimes L^T \right] \mathbf{vec}(\boldsymbol{\gamma}) - \mathbf{vec}(L^{-1} A_f Q_f^{-T}) \right\|_2^2 + \lambda \left\| \mathbf{vec}(\boldsymbol{\gamma}) \right\|_1, \quad (3.9)$$

with $Q_f Q_f^T = \mathbf{X}^T\mathbf{X}$, $A_f = \mathbb{E}[\Omega \tilde{\mathbf{Y}}^T \mathbf{X}]$, and $M = \overline{\Omega} = LL^T$; compare to (3.6) and (3.7). The derivation of (3.9) is similar to the presentation in Section (3.2) and may be found in Appendix (1.3).



Figures 3.6: **(left)** Evaluation of Δ_λ and π_λ along the solution path for the 25 size/value portfolios modeled by the 10 factors under the assumption of **fixed predictors**. An analyst may use this plot to select a particular model. Uncertainty bands are 75% posterior intervals on the Δ_λ metric. The large black dot and associated dashed line represents the model selected and shown on the right. **(right)** The selected model for 25 size/value portfolios modeled by the 10 factors. The responses and predictors are colored in gray and white, respectively. Edges represent nonzero components of the optimal action, $\boldsymbol{\gamma}$.

The corresponding version of the loss gap is

$$\Delta_\lambda = \mathcal{L}(\tilde{\mathbf{Y}}, \mathbf{X}, \Theta, \boldsymbol{\gamma}_\lambda^*) - \mathcal{L}(\tilde{\mathbf{Y}}, \mathbf{X}, \Theta, \boldsymbol{\gamma}^*).$$

which has distribution induced by the posterior over $(\tilde{\mathbf{Y}}, \Theta)$ rather than $(\tilde{Y}, \tilde{X}, \Theta)$ as before. By fixing $\mathbf{X}$, the posterior of Δ_λ has smaller dispersion which results in denser summaries for the same level of κ. For example, compare how dense the graph in Figures (3.4) is relative to the graph in Figures (3.6). The denser graph in Figures (3.6) contains all ten potential risk factors compared

to just three in Figures (3.4), which correspond to the Fama-French factors described in Fama and French [1992]. Recall that both graphs represent the sparsest model such that the probability of being no worse than the saturated model is greater than $\kappa = 12.5\%$ — the difference is that one of the graphs defines "worse-than" in terms of a fixed set of risk factor returns while the other acknowledges that those returns are themselves uncertain in future periods.

Figure (3.7) demonstrates this problem for several choices of the uncertainty level. Regardless of the uncertainty level chosen, the selected models contain most (if not all) of the ten factors and many edges. In fact, it is difficult to distinguish even the $\kappa = 2\%$ and $\kappa = 47.5\%$ models.

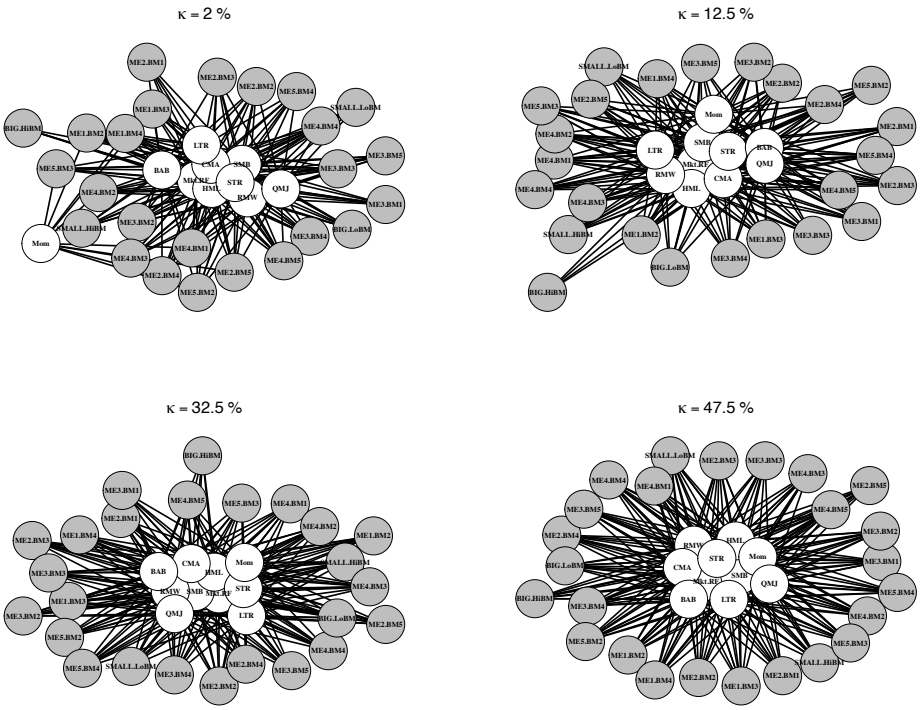


Figure 3.7: Sequence of selected models for varying threshold level κ under the assumption of **fixed predictors**.

3.5 Discussion

In this chapter, we propose a model selection summary for multivariate linear models when future realizations of the predictors are unknown. Such models are widely used in many areas of science and economics, including genetics and asset pricing. Our utility-based sparse posterior summary procedure is a multivariate extension of the "decoupling shrinkage and selection" methodology of Hahn and Carvalho [2015]. The approach we develop has three

steps: (*i*) fit a Bayesian model, (*ii*) specify a utility function with a sparsity-inducing penalty term and optimize its expectation, and (*iii*) graphically summarize the posterior impact (in terms of utility) of the sparsity penalty. Our utility function is based on the kernel of the conditional distribution responses given the predictors and can be formulated as a tractable convex program. We demonstrate how our procedure may be used in asset pricing under a variety of modeling choices.

The remainder of this discussion takes a step back from the specifics of the seemingly unrelated regressions model and considers a broader role for utility-based posterior summaries.

A paradox of applied Bayesian analysis is that posterior distributions based on relatively intuitive models like the SUR model are often just as complicated as the data itself. For Bayesian analysis to become a routine tool for practical inquiry, methods for summarizing posterior distributions must be developed apace with the models themselves. A natural starting point for developing such methods is decision theory, which suggests developing loss functions specifically geared towards practical posterior summary. As a matter of practical data analysis, articulating an apt loss function has been sorely neglected relative to the effort typically lavished on the model specification stage, specifically prior specification. Ironically (but not surprisingly) our application demonstrates that one's utility function has a dominant effect on the posterior summaries obtained relative to which prior distribution is used.

This chapter makes two contributions to this area of "utility design".

First, we propose that the likelihood function has a role to play in posterior summary apart from its role in inference. That is, one of the great practical virtues of likelihood-based statistics is that the likelihood serves to summarize the data by way of the corresponding point estimates. By using the log-density as our utility function applied to *future* data, we revive the fundamental summarizing role of the likelihood. Additionally, note that this approach allows three distinct roles for parameters. First, all parameters of the model appear in defining the posterior predictive distribution. Second, some parameters appear in *defining* the loss function; Ψ plays this role in our analysis. Third, some parameters define the action space. In this framework there are no "nuisance" parameters that vanish from the estimator as soon as a marginal posterior is obtained. Once the likelihood-based utility is specified, it is a natural next step to consider augmenting the utility to enforce particular features of the desired point summary. For example, our analysis above was based on a utility that explicitly rewards sparsity of the resulting summary. A traditional instance of this idea is the definition of high posterior density regions, which are defined as the *shortest, contiguous* interval that contains a prescribed fraction of the posterior mass.

Our second contribution is to consider not just one, but a range, of utility functions and to examine the posterior distributions of the corresponding posterior loss. Specifically, we compare the utility of a sparsified summary to the utility of the optimal non-sparse summary. Interestingly, these utilities are random variables themselves (defined by the posterior distribution) and

examining their distributions provides a fundamentally Bayesian way to measure the extent to which the sparsity preference is driving one's conclusions. The idea of comparing a hypothetical continuum of decision-makers based on the posterior distribution of their respective utilities represents a principled Bayesian approach to exploratory data analysis. This is an area of ongoing research.

3.6 Ongoing work in asset return modeling

Another exciting area of work in asset return modeling is related to the unpublished manuscript Carvalho et al. [2018]. In this paper, the goal is to model asset returns at time t as a function of firm level characteristics at time $t-1$ (such as the book-to-market or market capitalization of a firm). Thus, the framework here requires a *predictive* regression. Letting R_{it} be the excess return of firm i at time t and X_{it-1} a vector of characteristics at time $t-1$, our goal is to describe the conditional expectation function (CEF):

$$\mathbb{E}[R_{it} \mid \mathrm{X}_{it-1}]. \tag{3.10}$$

It has been common practice in the finance literature to study predictors of returns in one of two ways: (i) A predictive linear regression for modeling the cross section of observed returns and firms characteristics or (ii) Portfolio sorts on firm characteristics. The former implies that the relationship between firm characteristics and returns are linear and stationary. The latter is analogous to "building factors" on characteristics (and estimating step func-

tions for the CEF) like those explored in the seemingly unrelated regression work above. Both approaches have shortcomings.

In the regression approach, recent research suggests that relationships between returns and firm characteristics should be nonlinear and time-varying. See Freyberger et al. [2017] for a recent finance paper exploring these ideas. Equally complicating is the existence of many observed characteristics. Which characteristics matter for the purposes of explaining returns, and during what time periods are they relevant?

In the portfolio sorting approach, the large dimensionality of the "characteristic space" poses a challenge. To understand this, we briefly describe portfolio sorts. A portfolio sort involves organizing firms by characteristics and examining the returns of firms in a given quantile. It is often used to build explanatory factors of returns based on a given firm characteristic. Prototypical examples of factors are the *size* and *value* factors unveiled in Fama and French [1992]. Both of these factors are portfolios formed from companies after sorting them along the dimension of interest; for size sorting is done on market capitalization and for value sorting is done on book-to-market. The sorting is accomplished with characteristic information up to time $t-1$, and then the firm returns in, for example, each quintile or decile may be examined. The factors themselves are constructed by forming long-short portfolios of firms in the lowest and highest quantiles (i.e. the size factor buys low market cap firms and sells high market cap firms). As discussed in Freyberger et al. [2017], these sorts correspond to estimating the conditional mean function of

returns as a step function. In a loose sense, sorting is a special case of nonparametric regression. Additionally, we often see a monotonic relationship between characteristics and returns when a sort is done. See Cattaneo et al. [2018] for a nice connection between portfolio sorting and nonparametric estimation.

Suppose we have ten characteristics ($\mathrm{X}_{it-1} \in \mathbb{R}^{10}$) and would like to sort stocks jointly into quintile portfolios across these characteristics to understand returns at time t. This would involve constructing $5^{10} = 9,765,625$ distinct portfolios which is larger than the observed number of stocks at any time in the US stock market. More problematic, sorting offers little guidance as to which characteristics are truly relevant to returns. See Fama and French [2015] for a discussion of the drawbacks of portfolio sorting in the context of building their five factor model.

Our goal is to build a regression model to describe $\mathbb{E}[R_{it} \mid \mathrm{X}_{it-1}]$ that:

1. Models nonlinear relationships between characteristics and returns.
2. Incorporates shrinkage to bias away from irrelevant characteristics.
3. Imposes model structure through monotonicity across characteristic dimensions.

With these features in mind, our model will combine the attractive nonparametric elements of portfolio sorting and structure imposed by traditional regression techniques.

3.6.1 Previous literature

This research is motivated by the work of Freyberger et al. [2017]. They develop an additive quadratic spline model on 36 characteristics, and they eliminate weaker characteristics using techniques similar to the LASSO regression of Tibshirani [1996]. Their developments are important because splines allow them to infer a nonlinear return surface, and penalized regression manages complexity as more characteristics are added.

We agree with the sentiment of Freyberger et al. [2017], but the flexibility of their modeling approach violates a priori beliefs gathered from asset pricing theory that most characteristics have a *monotonic* relationship with returns. Our approach will incorporate monotonic structure at the characteristic level if it is supported by existing finance theory or empirical evidence. Additionally, we allow for time variation in our model while Freyberger et al. [2017] can only incorporate dynamics through rolling window estimation.

Shively et al. [2009] develop of model for smooth, monotonic, quadratic splines. Using carefully designed priors, their model can remove spline knots if the data suggests they are irrelevant. In other words, if the true relationship between the response and a covariate is linear, the coefficients associated with all quadratic knots should be shrunk toward zero in this model. Therefore, we can specify a more-than-sufficient number of knots and the model will correct our choice. We model time variation by merging Shively's monotonic spline model with McCarthy et al. [2016] who provide a general method for discounting by power weighting the likelihood density. Other papers related

to dynamic functional estimation include Kowal et al. [2017].

We are especially interested in knowing which firm characteristics *matter* for returns at each point in time. In order to do this, we adapt the decoupling shrinkage and selection methodology set out in Hahn and Carvalho [2015] and developed for applications in Puelz et al. [2016, 2017, 2018] to our monotonic, quadratic spline model. The upcoming section discusses the model and some preliminary results.

3.6.2 A nonlinear, monotonic conditional expectation function

We assume expected returns are equal to the sum of quadratic splines of firm characteristics. Therefore, our additive model is:

$$\mathbb{E}[R_{it} \mid \mathrm{X}_{it-1}] = \alpha_t + \sum_{k=1}^{K} f_{kt}(x_{ki,t-1}) \tag{3.11}$$

where R_{it} is the time t return for firm i, and α_t is the intercept term for time t. $\mathrm{X}_{it-1} = (x_{1i,t-1}, \ldots, x_{Ki,t-1})$ is a K length vector of firm i's characteristics at time $t-1$, where each characteristic is individually ranked across firms, resulting in $x_{ki,t-1} \in [0,1]$. Defining n_t as the number of firm observations at time t, we can write the formula for the ranked characteristics as $x_{ki,t-1} = \frac{\mathrm{rank}_{k,t-1}(\mathrm{characteristic}_{ki,t-1})}{n_t+1}$. The resulting characteristic is the empirical quantile of firm i's k^{th} characteristic at the beginning of time t. This is the approach of Freyberger et al. [2017] and closely follows portfolio sorting methods. However, inputting the characteristic quantiles leads to interpretation of the intercept in Model (3.11) as the expected return for the "perfectly minimum" firm along all

quantile dimensions. This interpretation is unmotivated, and the parameter may be difficult to learn since there is little or no data at the extremes of characteristics in the cross section. Therefore, we make a novel adjustment by instead defining:

$$x_{ki,t-1} = \frac{\text{rank}_{k,t-1}(\text{characteristic}_{ki,t-1})}{n_t + 1} - 0.5. \tag{3.12}$$

This shift by 0.5 results in $x_{ki,t-1} \in [-0.5, 0.5]$, and the intercept is interpreted as the expected return for the "perfectly median" firm, that is, a firm that has the median value across all characteristics.

We now describe the structure of the spline model. Letting f_{kt} be the quadratic spline for characteristic k at time t, we drop the kt subscripts for clarity. For a given series of $\grave{m}$ nonpositive knots ($\grave{x}_{\grave{m}} < ... < \grave{x}_1 < 0$) and $\acute{m}$ nonnegative knots ($0 < \acute{x}_1 < ... < \acute{x}_{\acute{m}}$), we set

$$\begin{aligned} f(x) =& \beta_1 x \\ &+ \beta_2 (x)_-^2 + \beta_3 (x - \grave{x}_1)_-^2 + ... + \beta_{\grave{m}+2} (x - \grave{x}_{\grave{m}})_-^2 \\ &+ \beta_{\grave{m}+3} (x)_+^2 + \beta_{\grave{m}+4} (x - \acute{x}_1)_+^2 + ... + \beta_{\grave{m}+\acute{m}+3} (x - \acute{x}_{\acute{m}})_+^2 \end{aligned}$$

where the $(y)_+ = \max(0, y)$ and $(y)_- = min(0, y)$. This can be abbreviated as

$$f_{kt}(x) = \mathrm{X}_k^T \beta_{kt} \tag{3.13}$$

where X_k is the carefully constructed quadratic spline basis.

We create these splines to be nondecreasing (without loss of generality) using the ideas of Shively et al. [2009], Section 3, though we have both

positive and negative knots (they limit $x \in [0,1]$). By definition, the spline is monotonic nondecreasing if the first derivative is nonnegative for all x:

$$f'(x) \geq 0. \tag{3.14}$$

This yields $\grave{m} + \acute{m} + 3$ conditions to satisfy:

$$0 \leq f'(-0.5) = \beta_1 + 2\beta_2(-0.5) + 2\beta_3(-0.5 - \grave{x}_1) + ... + 2\beta_{\grave{m}+2}(-0.5 - \grave{x}_{\grave{m}})$$

$$0 \leq f'(\grave{x}_{\grave{m}}) = \beta_1 + 2\beta_2(\grave{x}_{\grave{m}}) + 2\beta_3(\grave{x}_{\grave{m}} - \grave{x}_1) + ... + 2\beta_{\grave{m}+1}(\grave{x}_{\grave{m}} - \grave{x}_{\grave{m}-1})$$

$$\vdots$$

$$0 \leq f'(\grave{x}_2) = \beta_1 + 2\beta_2(\grave{x}_2) + 2\beta_3(\grave{x}_2 - \grave{x}_1)$$

$$0 \leq f'(\grave{x}_1) = \beta_1 + 2\beta_2(\grave{x}_1)$$

$$0 \leq f'(0) = \beta_1$$

$$0 \leq f'(\acute{x}_1) = \beta_1 + 2\beta_{\grave{m}+3}(\acute{x}_1)$$

$$0 \leq f'(\acute{x}_2) = \beta_1 + 2\beta_{\grave{m}+3}(\acute{x}_2) + 2\beta_{\grave{m}+4}(\acute{x}_2 - \acute{x}_1)$$

$$\vdots$$

$$0 \leq f'(\acute{x}_{\acute{m}}) = \beta_1 + 2\beta_{\grave{m}+3}(\acute{x}_{\acute{m}}) + 2\beta_{\grave{m}+4}(\acute{x}_{\acute{m}} - \acute{x}_1) + ... + 2\beta_{\grave{m}+\acute{m}+2}(\acute{x}_{\acute{m}} - \acute{x}_{\acute{m}-1})$$

$$0 \leq f'(0.5) = \beta_1 + 2\beta_{\grave{m}+3}(0.5) + 2\beta_{\grave{m}+4}(0.5 - \acute{x}_1) + ... + 2\beta_{\grave{m}+\acute{m}+3}(0.5 - \acute{x}_{\acute{m}})$$

which can be vectorized into a lower triangular matrix $\mathbf{L}$ multiplied by the coefficient vector β. Reintroducing the kt subscripts, these nonnegativity constraints for each characteristic k at time t may be written as:

$$\mathbf{0} \leq \mathbf{L}\beta_{kt} = \gamma_{kt} \tag{3.15}$$

where we don't include a subscript kt on $\mathbf{L}$ since each characteristic k is given the same knots spanning the unit interval in our method. Moreover, we see that $\mathbf{L}$ acts as a projection matrix, projecting our more complicated constraints on β_{kt} to the simple nonnegative constraints on γ_{kt}. Hence

$$\begin{aligned} f_{kt}(x) &= \mathrm{X}_k^T \beta_{kt} \\ &= \mathrm{X}_k^T \mathbf{L}^{-1} \mathbf{L} \beta_{kt} \\ &= \mathrm{X}_k^T \mathbf{L}^{-1} \gamma_{kt} \\ &= w_k^T \gamma_{kt} \end{aligned}$$

where $w_k^T = \mathrm{X}_k^T \mathbf{L}^{-1}$ is now our modified spline basis. Overloading time and firm subscripts and superscripts on w_k^T above, we write Equation (3.11) as:

$$\mathbb{E}[R_{it} \mid \mathrm{X}_{it-1}] = \alpha_t + \sum_{k=1}^{K} w_{kt-1}^{iT} \gamma_{kt}. \tag{3.16}$$

Here, information about the lagged characteristics X_{it-1} are incorporated into w_{kt-1}^i. Note that this modeling approach allows one to specify individually the monotonicity of each covariate (characteristic) through the design of $\mathbf{L}$. Additionally, an analyst may decide to not specify monotonicity constraints. In this case, a small modification is made to the sampler, described in subsequent sections.

3.6.3 The model

With Equation (3.16) introduced, the model may be written as:

$$R_{it} = \alpha_t + \sum_{k=1}^{K} w_{kt-1}^{iT} \gamma_{kt} + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma_t^2). \tag{3.17}$$

We introduce an indicator variable that determines whether or not the regression coefficients are nonzero. Since these coefficients correspond to knots in the spline basis, this structure allows the model to select the proper knots for the splines. Formally, let j denote the vector index for the j^{th} knot. Let $I_{jkt} = 1$ indicate that $\gamma_{jkt} > 0$ and $I_{jkt} = 0$ indicate that $\gamma_{jkt} = 0$. Thus, I_{jkt} is a Bernoulli random variable with prior probability $P(I_{jkt} = 1) = p_{jkt}$. This leads us to the conditional prior on γ_{jkt}:

$$(\gamma_{jkt}|I_{jkt} = 1, \cdot) \sim N_+(0, c_k\sigma_t^2). \tag{3.18}$$

This setup permits the data to select the knots for the splines. By over-specifying the number of potential knots, the data will inform the model as to which should be included ($I_{jkt} = 1$) or not ($I_{jkt} = 0$).

Following Shively et al. [2009], we place uninformative priors on $\alpha \sim N(0, 10^{10})$ and $\sigma^2 \sim U(0, 10^3)$ as well as set $p_{jkt} = 0.2$, $\forall j, k, t$. In summary, the fully specified model for the vector of n_t firm returns R_t is:

$$\begin{aligned}
\mathrm{R}_t &\sim \mathrm{N}\left(\alpha_t \mathbf{1}_{n_t} + \mathbf{X}_{t-1}\boldsymbol{\beta_t},\ \sigma_t^2 \mathbb{I}_{n_t}\right), \\
&\quad \text{with } \mathbf{X}_{t-1}\boldsymbol{\beta}_t = \mathbf{W}_{t-1}\boldsymbol{\gamma}_t \\
\alpha_t &\sim N(0, 10^{10}) \\
\sigma_t^2 &\sim U(0, 10^3) \\
(\gamma_{jkt}|I_{jkt} = 1) &\sim N_+(0, c_k\sigma_t^2) \\
(\gamma_{jkt}|I_{jkt} = 0) &= 0 \\
I_{jkt} &\sim Bn(p_{jkt} = 0.2)
\end{aligned}$$

where $\mathbf{X}_{t-1}\boldsymbol{\beta}_t = \mathbf{X}_{t-1}\text{diag}_K(\mathbf{L})^{-1}\text{diag}_K(\mathbf{L})\boldsymbol{\beta}_t = \mathbf{W}_{t-1}\boldsymbol{\gamma}_t$. Note that $\text{diag}_K(\mathbf{L})$ is a block diagonal matrix of size $K(\grave{m}+\acute{m}+3)\times K(\grave{m}+\acute{m}+3)$ where each lower triangular block is the projection matrix $\mathbf{L}$. Also, $\mathbf{X}_{t-1}$ is matrix of size $n_t \times K(\grave{m}+\acute{m}+3)$ and $\boldsymbol{\beta}_t$ is vector of size $K(\grave{m}+\acute{m}+3)$. Therefore, each firm is given a row in $\mathbf{X}_{t-1}$, and each $\grave{m}+\acute{m}+3$ block of $\boldsymbol{\beta}_t$ corresponds to the coefficients on the spline basis for a particular characteristic, k. Incorporating the intercept directly into the characteristic matrix, we can write the generating model compactly as:

$$\mathrm{R}_t \sim \mathrm{N}\left(\mathbb{X}_{t-1}\mathbf{B}_t,\ \sigma_t^2\mathbb{I}_{n_t}\right) \tag{3.19}$$

where

$$\begin{aligned} \mathbb{X}_{t-1} &= \begin{bmatrix}\mathbf{1}_{n_t} & \mathbf{X}_{t-1}\end{bmatrix} \\ \mathbf{B}_t &= \begin{bmatrix}\alpha_t & \boldsymbol{\beta}_t\end{bmatrix} \end{aligned} \tag{3.20}$$

3.6.4 Time dynamics

Let Θ_t be our model parameters at time t. To allow the model's parameters over time, we use the power-weighted likelihood approach of McCarthy et al. [2016]. Their approach provides the ability to evolve parameters over time without specifying an explicit evolution equation. For $\delta_t \in [0,1]$, such that $\delta_1 \leq \delta_2 \leq ... \leq \delta_\tau$, the likelihood at time $\tau \in \{1,...,T\}$ discounts the impact of past data. Therefore the power-weighted density posterior distribution of the parameter vector at time τ is written as follows:

$$p(\Theta_\tau \mid R_{1:\tau}) \propto p(\Theta_\tau)\prod_{t=1}^{\tau} p(\boldsymbol{r}_t|\Theta_\tau)^{\delta_t}. \tag{3.21}$$

We choose values of δ_t that are exponentially decreasing in time $\delta_t = \delta^t$ for some fixed value $\delta \in [0, 1]$. McCarthy et al. [2016] discuss how this fixed δ may be chosen to maximize the one-step ahead predictive likelihood of observed data. As a first step, we consider a range of δ's. In the following subsection, we derive a Gibbs sampler used to explore the joint posterior.

3.6.5 Parameter sampling

To sample all parameters at time $\tau \in \{1, ..., T\}$, iterate through the following, conditional upon the most recent draws of other parameters:

$$\begin{aligned}
\alpha_\tau \mid \cdot &\sim N\left(\frac{V}{\sigma^2}\sum_{t=1}^{\tau}\delta_t \mathbf{1}_{n_t}^T\left[\mathrm{R}_t - \mathbf{W}_{t-1}\boldsymbol{\gamma}_t\right],\quad V = \left[\frac{1}{\sigma^2}\sum_{t=1}^{\tau}\delta_t n_t + \frac{1}{10^{10}}\right]^{-1}\right)\\
\sigma_\tau^2 \mid \cdot &\sim IG(a, b)\\
I_{jk\tau} \mid \cdot &\sim Bn(p^*_{jk\tau})\\
\gamma_{jk\tau} \mid \cdot &\sim \begin{cases} 0 & \text{if } I_{jkt} = 0\\ N_+\left(V_{jk\tau}\sum_{t=1}^{\tau}\delta_t \mathrm{E}_{jkt}^T w_{jkt},\quad \sigma^2 V_{jk\tau}\right) & \text{if } I_{jkt} = 1\end{cases}
\end{aligned}$$

where:

$$\begin{aligned}
a &= \left[\frac{1}{2}\left(\sum_{t=1}^{\tau} n_t\delta_t + \sum_{j=1}^{m+2}\sum_{k=1}^{K} I_{jk}\right) - 1\right],\\
b &= \frac{1}{2}\left[\sum_{t=1}^{\tau}\delta_t\left(\mathrm{R}_t - \alpha_\tau \mathbf{1}_{n_t} - \mathbf{W}_{t-1}\boldsymbol{\gamma}_t\right)^T(\prime\prime) + {c_k}^{-1}\boldsymbol{\gamma}_t^T\boldsymbol{\gamma}_t\right].
\end{aligned}$$

for $j = 1, ..., m+2$ and $k = 1, ..., K$:

$$\mathrm{E}_{jkt} = \mathrm{R}_t - \alpha \mathbf{1}_{n_t} - [\mathbf{W}_{t-1}\boldsymbol{\gamma}_t]_{-jk} \quad \text{and} \quad V_{jk\tau} = \left(\sum_{t=1}^{\tau} \delta_t w_{jkt}^T w_{jkt} + c_k^{-1} \right)^{-1}$$

$$p^*_{jk\tau} = \frac{\rho_1 \rho_2}{\rho_1 \rho_2 + (1 - p_{jkt})}$$

$$\rho_1 = 2 p_{jkt} c_k^{-\frac{1}{2}} V_{jk\tau}^{\frac{1}{2}} exp \left\{ -\frac{1}{2\sigma^2} \left[-V_{jk\tau} \left(\sum_{t=1}^{\tau} \delta_t \mathrm{E}_{jkt}^T w_{jkt} \right)^2 \right] \right\},$$

$$\rho_2 = 1 - \Phi \left(0 \left| V_{jk\tau} \sum_{t=1}^{\tau} \delta_t \mathrm{E}_{jkt}^T w_{jkt},\ \sigma^2 V_{jk\tau} \right. \right).$$

The current draw for $\gamma_{jk\tau}$ assumes that the conditional expectation function is monotone *increasing* with respect to characteristic k. Monotone decreasing can be enforced by drawing from the negative half of the truncated normal, and constraints are removed by drawing from the "full" version of the normal distribution specified in the $\gamma_{jk\tau}$ draw.

3.6.6 Motivating examples

In this section, we present simulated examples to demonstrate the methodology. We will focus on the benefits of certain features of the model, including monotonic constraints and the ability to model dynamics using a power-weighted density approach.

3.6.6.1 Why monotonicity?

The answer to this question harkens back to the the first chapter of this thesis which discussed the bias-variance tradeoff. A function with more

flexibility may be able to weave through the observed data more effectively, but this increased freedom comes at a price, and the price is increased variance.

Figure (3.8) displays this empirically. Here, we use our methodology on a single set of data. The true function is given by the dotted line. It is a parabola for a portion of the x-space, and the remainder of the function is a cubic. Gaussian noise is added to this function and 200 data points are generated and given by the gray dots.

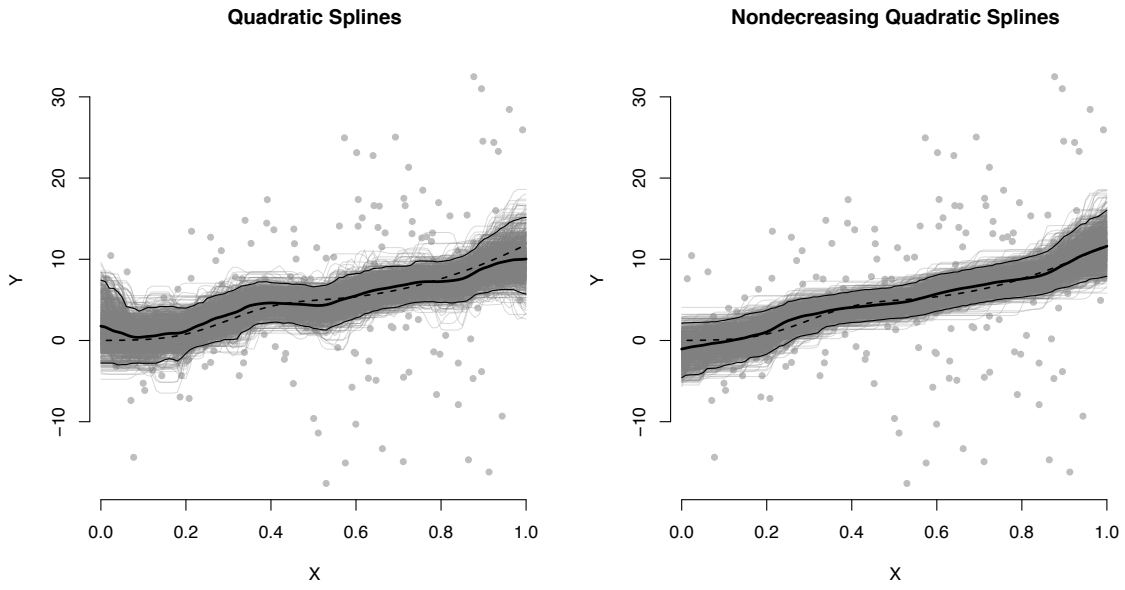


Figure 3.8: Comparison of quadratic spline fit (left) and monotonic quadratic spline fit (right) along with posterior means (black line) and individual MCMC draws (gray lines). The true function is given by the dotted black line.

The left graph displays the estimated function in black as well as draws from the sampler in the gray lines and 95% posterior credible interval around

the estimated function (which is the posterior mean). It is a quadratic spline fit when *no monotonicity* is enforced. The right graph displays the estimated function with monotonicity enforced.

Notice how the credible intervals are wider when monotonicity is not enforced (left) compared with a monotonicity constraint (right). This is especially pronounced in high noise, low signal estimation tasks like those encountered in finance. The quadratic spline fit is influenced by noise in the data. The benefits of introducing bias via the monotonicity constraint are clear – increased model structure allows the sampled and estimated functions to be "better behaved", and the result is inference with lower variance estimates. Since finance theory often predicts the direction of monotonicity for a firm characteristic on expected return, it is beneficial to utilize this information in our methodology.

How does the estimation methodology work on actual finance data? Figure (3.9) shows a function fit to "momentum" and excess return monthly data. Momentum is calculated as the cumulative return of a firm from the past 12 months through 2 months prior to month t. Jegadeesh and Titman [1993] famously documented that "past winners" (firms that do well in the past) tend to outperform in the future and "past losers" (firms that historically have low returns) tend to underperform in the future. Thus, we enforce increasing monotonicity in the momentum characteristic. The power-weighted density approach is used, and the data begins in January 1965. The function estimate has been smoothed over 157 months of data, but most weight is given to data in

the estimation month: January 1978. Notice the drastic difference in credible intervals between the quadratic spline and monotonic quadratic spline fits. Since previous empirical work tells us that past winners have higher future returns than past losers, applying this constraint during inference results in a much lower variance estimate and identification of the return-characteristic relationship.

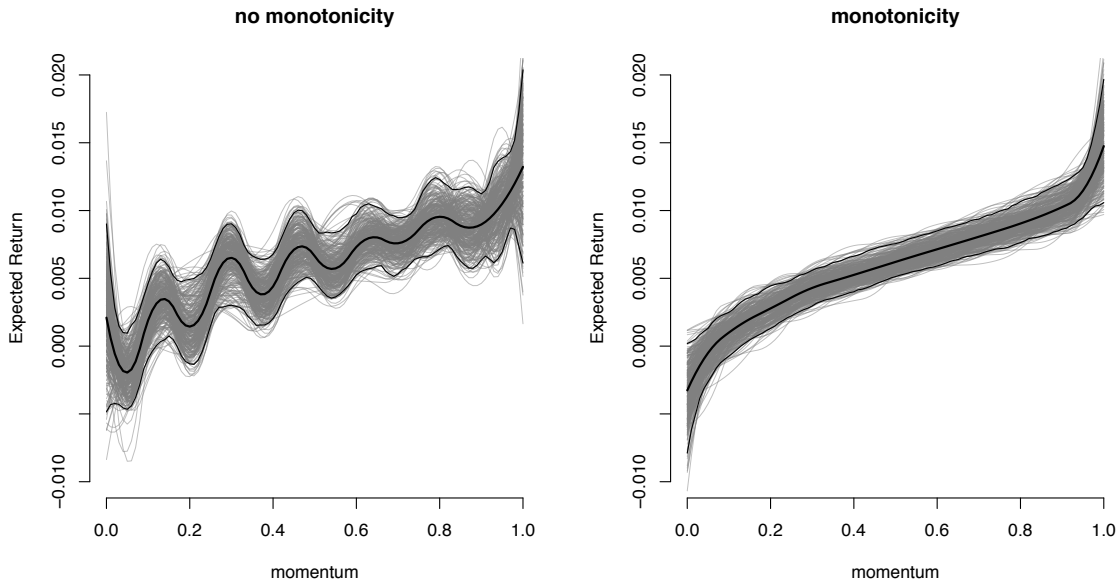


Figure 3.9: Comparison of quadratic spline fit (left) and monotonic quadratic spline fit (right) along with posterior means (black line) and individual MCMC draws (gray lines). This is based on monthly momentum and excess return data. Shown is the function fit for January 1978.

3.6.6.2 Why time dynamics?

A second feature of the methodology is the ability to estimate nonlinear, monotonic functions that are dynamic. This is important in a finance context because we would like to determine which characteristics are relevant to returns at different points in time. Our goal is to incorporate sensible time dynamics while maintaining interpretability and model structure.

Figure (3.10) demonstrates our methodology utilitizing power-weighted density for dynamics [McCarthy et al., 2016] versus a commonly used alternative – updating the historical average. The figure considers a true function that is a parabola flattening over 11 time periods to a line $f(x) = 0$. At each time period, 100 data points are generated from each of these functions with the addition of Gaussian noise.

The more faded data and lines correspond to functions and samples (respectively) that are further back in time. We show the function estimates and draws for a historical average estimate (dotted black line) and a power-weighted density estimate (solid black line) at time 11. The historical average estimate treats each new cross sectional data sample equivalent in terms of weighting, and the estimated function is markedly bias towards earlier data generated from the higher curvature parabolas. The historical average estimate is achieved by setting the density discount parameter $\delta = 1$. Another issue with this approach is that future observed data has an increasingly diminished effect on the function estimate.

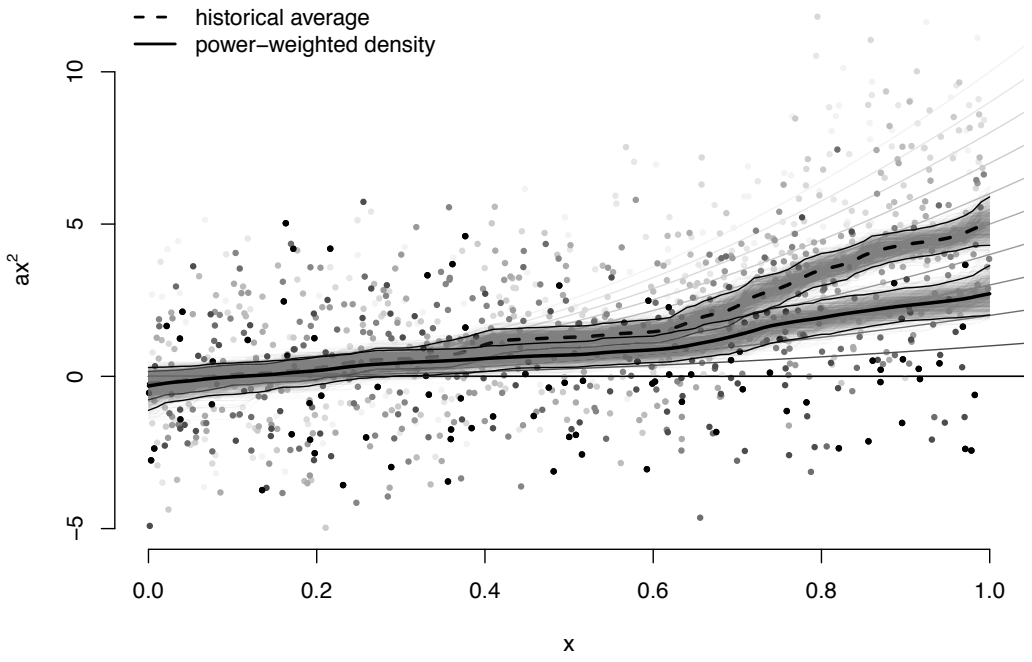


Figure 3.10: This figure demonstrates the dynamic estimation approach of the model. The true generating function is a parabola ax^2 that flattens over 11 time points (a starts at 10 and increments to 0). All 11 true parabolas are shown by the gray lines that fade to white for functions further back in time. Gaussian noise is added to the true functions, and 100 points are generated for each of the 11 time periods – each cross sectional sample is also shown by the gray dots that fade to white for data further back in time. Displayed are two monotonic function estimations at time point 11: (i) Historical average given by the dotted line, and (ii) Power-weighted density estimation given by the solid line.

In contrast, the power-weighted density estimate differentially weights the likelihood based on when the data is observed. Past data receives smaller

weight than more recent data, and the effect is a flatter function estimate at time 11. Additionally, time dynamics of future function estimates are ensured since the strictly largest weight will always be given to the most recently observed data.

In Figure (3.11), we use monthly financial data to show the dynamics of the CEF when modeled by the momentum characteristic. The left figure displays the function estimate for January 1978 (also shown in Figure (3.9)), and the right figure displays the estimate at January 2014. These dates were chosen before and after the concept of momentum appeared in the finance literature around 1980. Comparing these estimates provides evidence that the "momentum effect" – where past winners continue to win and past losers continue to lose – may be diminishing over time!

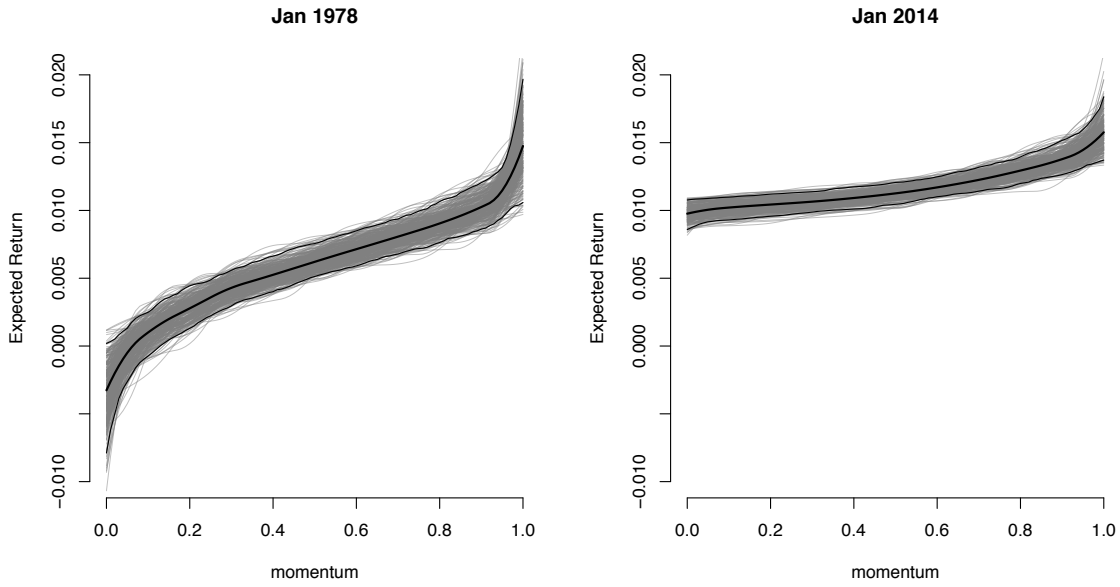


Figure 3.11: Comparison of January 1978 fit (left) and January 2014 fit (right) along with posterior means (black line) and individual MCMC draws (gray lines). This is based on monthly momentum and excess return data.

3.6.7 Utility-based selection of characteristics

We conclude this section by discussing future work that will incorporate utility-based selection into the proposed methodology. A key practical question in finance is which firm characteristics are necessary for return prediction. In our application, we will consider a recent version of the data used in Freyberger et al. [2017] which includes 36 firm characteristics in the cross section of returns. By fitting our monotonic quadratic spline model on all 36 characteristics at every point in time, we obtain posterior draws of model parameters from a relatively complex posterior distribution. This posterior can

be parsimoniously summarized using the procedure described below.

Recall the compactly written version of our model:

$$\mathrm{R}_t \sim \mathrm{N}\left(\mathbb{X}_{t-1}\mathbf{B}_t,\ \sigma_t^2\mathbb{I}_{n_t}\right) \tag{3.22}$$

where

$$\begin{aligned}\mathbb{X}_{t-1} &= \begin{bmatrix}\mathbf{1}_{n_t} & \mathbf{X}_{t-1}\end{bmatrix} \\ \mathbf{B}_t &= \begin{bmatrix}\alpha_t & \boldsymbol{\beta}_t.\end{bmatrix}\end{aligned} \tag{3.23}$$

$\mathbf{X}_{t-1}$ is an $n_t \times K(\grave{m}+\acute{m}+3)$ matrix representing the spline basis with K characteristics and $\grave{m}+\acute{m}+3$ basis vectors for each characteristic. It is structured so that the first $\grave{m}+\acute{m}+3$ columns correspond to the basis for characteristic 1, the next $\grave{m}+\acute{m}+3$ for characteristic 2, and so forth. Thus, $\boldsymbol{\beta}_t$ is a stacked $K(\grave{m}+\acute{m}+3)$ length vector whose coefficients are organized in the same way: The k^{th} $\grave{m}+\acute{m}+3$ sub-vector corresponds to coefficients on characteristic k's spline basis. With the model specified, we walk through the primitives and procedure of utility-based selection. The methodology here is conditional on the observed characteristic values $\mathbb{X}_{t-1}$, so we will see that the approach is very similar to Hahn and Carvalho [2015], now applied to monotonic quadratic splines.

First, we specify the log density of Regression (3.22) as our utility function.

$$\mathcal{L}_{\lambda_t}(\tilde{\mathrm{R}}_t, \mathbf{A}_t, \Theta_t) = \frac{1}{2}(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t)^T(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t) \tag{3.24}$$

where $\tilde{\mathrm{R}}_t$ is future return data at time t, Θ_t is a vector of model parameters, and $\mathbf{A}_t$ is the "action" to be taken by the data analyst. This action is intended to

represent a sparse summary of the regression vector $\mathbf{B}_t$. In order to encourage sparsity in $\mathbf{A}_t$, we include an additional penalty function Φ with parameter λ_t:

$$\mathcal{L}_{\lambda_t}(\tilde{\mathrm{R}}_t, \mathbf{A}_t, \Theta_t) = \frac{1}{2}(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t)^T(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t) + \Phi(\lambda_t, \mathbf{A}_t). \tag{3.25}$$

The second step of the selection methodology is to integrate the loss function over all uncertainty given by the joint distribution of asset returns and model parameters, conditioned on observed data: $p(\tilde{\mathrm{R}}_t, \Theta_t \mid \mathbf{R}_t) = p(\tilde{\mathrm{R}}_t \mid \Theta_t, \mathbf{R}_t)p(\Theta_t \mid \mathbf{R}_t)$. We do this integration in two steps, first over $\tilde{\mathrm{R}}_t \mid \Theta_t$ and second over Θ_t:

$$\begin{aligned}
\mathcal{L}_{\lambda_t}(\mathbf{A}_t) &= \mathbb{E}_{\Theta_t}\mathbb{E}_{\tilde{\mathrm{R}}_t|\Theta_t}\left[\frac{1}{2}(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t)^T(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t) + \Phi(\lambda_t, \mathbf{A}_t)\right] \\
&= \mathbb{E}_{\Theta_t}\mathbb{E}_{\tilde{\mathrm{R}}_t|\Theta_t}\left[\frac{1}{2}(\tilde{\mathrm{R}}_t^T\tilde{\mathrm{R}}_t - 2\tilde{\mathrm{R}}_t^T\mathbb{X}_{t-1}\mathbf{A}_t + \mathbf{A}_t^T\mathbb{X}_{t-1}^T\mathbb{X}_{t-1}\mathbf{A}_t)\right] + \Phi(\lambda_t, \mathbf{A}_t) \\
&\propto \mathbb{E}_{\Theta_t}\left[2\mathbb{E}_{\tilde{\mathrm{R}}_t|\Theta_t}[\tilde{\mathrm{R}}_t^T]^T\mathbb{X}_{t-1}\mathbf{A}_t + \mathbf{A}_t^T\mathbb{X}_{t-1}^T\mathbb{X}_{t-1}\mathbf{A}_t\right] + \Phi(\lambda_t, \mathbf{A}_t) + \text{constants} \\
&= \mathbb{E}_{\Theta_t}\left[2\mathbf{B}_t^T\mathbb{X}_{t-1}^T\mathbb{X}_{t-1}\mathbf{A}_t\right] + \mathbf{A}_t^T\mathbb{X}_{t-1}^T\mathbb{X}_{t-1}\mathbf{A}_t + \Phi(\lambda_t, \mathbf{A}_t) + \text{constants} \\
&= 2\overline{\mathbf{B}}_t^T\mathbb{X}_{t-1}^T\mathbb{X}_{t-1}\mathbf{A}_t + \mathbf{A}_t^T\mathbb{X}_{t-1}^T\mathbb{X}_{t-1}\mathbf{A}_t + \Phi(\lambda_t, \mathbf{A}_t) + \text{constants}.
\end{aligned} \tag{3.26}$$

In the third line, we drop the one-half and collect all terms not involving the action $\mathbf{A}_t$ into "constants." After integrating over the joint distribution of returns and parameters, we notice that the posterior mean of the coefficients appears in the first term, while the expectations pass over the second and third terms.

We complete the square and drop constants to obtain the final form of

the integrated loss function:

$$\mathcal{L}_{\lambda_t}(\mathbf{A}_t) = \left\|\mathbb{X}_{t-1}\mathbf{A}_t - \mathbb{X}_{t-1}\overline{\mathbf{B}}_t\right\|_2^2 + \Phi(\lambda_t, \mathbf{A}_t) \tag{3.27}$$

For a fixed time t, Loss (3.27) is exactly the same as the one derived for linear regression models in Hahn and Carvalho [2015]. The third and final step is to choose a penalty function Φ and optimize the loss function for a range of λ_t for each time t. This is an area of ongoing development.

One interesting choice is $\Phi(\lambda_t, \mathbf{A}_t) = \lambda_t \sum_{k=1}^{K} \left\|\mathbf{A}_t^k\right\|_2^2$ where $\mathbf{A}_t^k$ is the k^{th} $\grave{m} + \acute{m} + 3$ block of the vector $\mathbf{A}_t$ after neglecting the intercept. The group lasso algorithm of Yuan and Lin [2006] can then be used to minimize the integrated loss. A similar alternative is the sparse group lasso discussed in Simon et al. [2013] where an ℓ_1 norm is added: $\Phi(\lambda_{1t}, \lambda_{2t}, \mathbf{A}_t) = \lambda_{1t} \sum_{k=1}^{K} \left\|\mathbf{A}_t^k\right\|_2^2 + \lambda_{2t} \left\|\mathbf{A}_t\right\|_1$. Both optimization approaches are useful for our methodology. They provide a clear way to undertake variable selection in spline models while taking uncertainty in future data and model parameters into account.

In order to see this, recall the structure of the sparse action $\mathbf{A}_t$. It is a $K(\grave{m} + \acute{m} + 3) + 1$ length vector where the k^{th} $\grave{m} + \acute{m} + 3$ block (excluding the intercept) corresponds to the spline basis for firm characterstic k. By using the approach outlined in Yuan and Lin [2006] or Simon et al. [2013], we can group together the spline bases for each characteristic. Then, Loss (3.27) is minimized for varying penalty parameter choices. We are then able to look at a range of monotonic quadratic spline models built from one characteristic

up to the 36 characteristics available. Analogous to the SUR model selection and portfolio selection work discussed in previous chapters, these models are optimal under our choice of utility and fixed level of regularization given by the penalty parameter, and we can compare them in light of predictive uncertainty.

An important feature of this approach is the ability to identify important return predictors and how this set may vary over time. The time variation and connection across time periods is driven by the power-weighted density approach and embedded in the posterior. Therefore, although the minimization of the integrated loss is performed myopically at each point in time, the variation of optimal sparse models across time may be studied.

As a final example, we revisit our monthly finance data and model the CEF with 36 firm characteristics. In Figure (3.12), we show the partial function fits for each of the characteristics including the intercept. We enforce monotonicity on each individual characteristic if there is finance theory or empirics that suggest a monotonic relationship with expected return.

These partial relationships are fascinating to study from a finance perspective since they provide information about the relationship between each characteristic and expected return *conditional* on all other characteristics. Notice that some characteristics are flat and thus have a negligible relationship with return – **c**, **rna**, **roa**, **roe**, and **free cf** – while others have substantial relationships, including **beme** (book-to-market) and **lme** (market capitalization). We will use utility-based selection at this stage to select which characteristics have the most influence on the CEF while taking predictive uncertainty into

account.

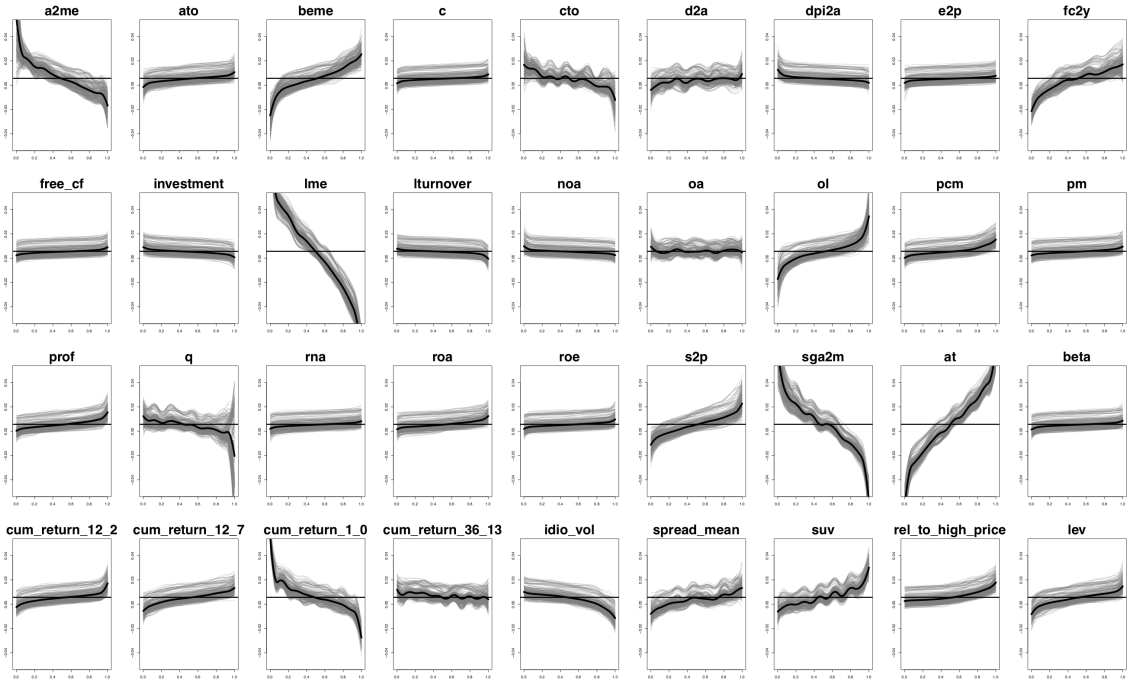


Figure 3.12: Partial function estimates for the CEF modeled with 36 firm characteristics. The intercept is plotted as a black horizontal line in each subgraph. Posterior means are shown in black and individual MCMC draws are shown in gray. This is based on monthly data from January 1965 to January 1978.

Chapter 4

Regularization and Confounding in Linear Regression for Treatment Effect Estimation

Some analysis and text in this chapter follows Hahn et al. [2018a]. We will reintroduce the main ideas described in that paper and discuss ongoing research.

4.1 Introduction

While the previous chapters focused on regularization and its purpose within a formal model selection setting, this chapter considers its use as solely a shrinkage/bias inducing technology. Specifically, we are interested in regularization's role in treatment effect estimation with observational (and potentially clustered) data.

A treatment effect – the amount a response variable would change if a treatment variable were changed by one unit – is appropriately estimated only when all other *confounding* variables are taken into account. Confounding variables are given such a name because they explain a portion of the correlation between the treatment and response variables; effectively masking (confounding) the true relationship the data analyst wishes to measure. The

models considered in this chapter specify a linear relationship between the response Y_i, and the treatment and covariates Z_i and X_i respectively:

$$Y_i = \alpha Z_i + \mathrm{X}_i^T \beta + \nu_i. \tag{4.1}$$

For notation, let letters denote vectors, boldfaced letters denote matrices, and italicized letters denote scalars. Let β be a p-length vector of coefficient parameters, and α be the scalar treatment effect parameter. The errors ν are normally distributed with zero mean and unknown variance. In observational studies, there may be many covariates, i.e.: p may be large. For example, corporate finance studies often involve observations that are firms and covariates that are firm characteristics taken from financial statements. The relationship of interest may be the effect of a firm's cash flow (Z_i) on its debt-to-equity ratio (Y_i), and there are a plethora of firm characteristics one can include as part of X_i. Although including all covariates may mitigate bias in the estimate for α, this is at the expense of increased variance of the estimator for α. Practically, interval lengths of the treatment effect for this naive approach will be large, and discovering statistical significance will be difficult.

One solution to the "many covariate" problem is to hand-select a subset of variables from X_i to control for in Model (4.1) and toss out the remaining covariates. Leamer [1983] describes how this procedure is an unsatisfyingly ad-hoc reaction to a practical data analysis issue. After hand-selecting covariates, how does the analyst truly know if all information from X_i is taken into account? This chapter provides a solution to this problem using statistical

regularization. Specifically, we propose using information from marginal likelihoods to narrow down our potentially large list of covariates. The aim is to replace an ad-hoc selection approach with one that is informed by the data, and our desire is to appeal to a broad base of researchers estimating linear treatment effects from observational data.

Exploring the use of regularization in treatment effect estimation and providing a procedure was the main contribution of Hahn et al. [2018a]. We extend these ideas to an empirical-Bayes setting and where model errors may be dependent across clusters of data. The forthcoming sections are speculative, but compare this new approach to Hahn et al. [2018a] and discuss areas of future work.

4.1.1 Previous literature

Treatment effect estimation is an important topic with a deep literature base. This work focuses on one slice: The use of Bayesian regularized regression models for effect estimation. Li and Tobias [2014] and Heckman et al. [2014] provide review articles of Bayesian approaches to this problem. Careful attention will be given to the impact of regularization on the estimation of treatment effects and new ways for characterizing the estimates' standard errors.

Hahn et al. [2018a] contributed to the small but growing literature on Bayesian approaches to treatment effect estimation via linear regression with many potential controls. They proposed a conceptual and computational

refinement of ideas first explored in Wang et al. [2012], where Bayesian adjustment for confounding is addressed via hierarchical priors. Their method can be seen as an alternative to Wang et al. [2012], with certain conceptual and computational advantages, namely ease of prior specification and posterior sampling. Other related papers include Wang et al. [2015], Lefebvre et al. [2014] and Talbot et al. [2015]; see also Jacobi et al. [2016]. Zigler and Dominici [2014] and An [2010] focus on Bayesian propensity score models (for use with binary treatment variables). Wilson and Reich [2014] takes a decision theoretic approach to variable selection of controls. Again, each of these previous approaches cast the problem as one of selecting appropriate controls; posterior treatment effect estimates are obtained via model averaging. Here, we argue that if the goal is estimation of a certain regression parameter (corresponding to the treatment effect, provided the model is correctly specified), then questions about which specific variables are necessary controls is a means to an end rather than an end in itself. Other recent papers looking at regularized regression for treatment effect estimation include Ertefaie et al. [2015] and Ghosh et al. [2015], but even here the focus is on variable selection via the use of 1-norm penalties on the regression coefficients.

There are several books dealing with the broader topic of causal inference, including Imbens and Rubin [2015], Morgan and Winship [2014], and Angrist and Pischke [2008]. Similar to Wang et al. [2012] where there is a focus on the joint modeling of the treatment and response variables as a function of covariates, the following papers have approached the problem similarly:

Rosenbaum and Rubin [1983], Robins et al. [1992], and McCandless et al. [2009].

Equally important and vast is the literature dealing with clustered inference. We defer this literature review to a final section on the application of our approach to the clustered data setting.

4.2 Regularization-induced confounding

What happens when regularization is naively used in treatment effect estimation? In this section, we illustrate this important phenomena, referred to as "regularization-induced confounding" (RIC). Hahn et al. [2018a] and Hahn et al. [2017] provide intuition for this issue within Bayesian linear regression and heterogenous treatment effect estimation using random forests, respectively. We recapitulate their exposition here since RIC is a central issue of this chapter. It is expected that regularization will introduce bias in coefficient estimates from a regression. What is not obvious is that bias will still exist in an *unregularized* treatment effect estimate if the treatment and covariates are correlated. For illustration, suppose regularization is introduced via a ridge penalty over parameters. A similar theoretical demonstration of RIC is presented in Hahn et al. [2017].

Returning to Model (4.1):

$$Y_i = \alpha Z_i + \mathrm{X}_i^T \beta + \nu_i,$$

the overall goal is to properly estimate the treatment effect α. We assume

that the error term is mean zero Gaussian and a ridge estimator is placed on the regression coefficients. Define observed data as $\tilde{\mathbf{X}} = (\mathrm{Z} \ \mathbf{X})$ and consider a ridge matrix $\mathbf{M}$. Following the seminal work of Hoerl and Kennard [1970], the ridge estimator for coefficients $\theta = \begin{pmatrix} \alpha & \beta^T \end{pmatrix}^T$ is $\hat{\theta}_{\text{ridge}} = (\mathbf{M} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathrm{Y} = (\mathbb{I}_{p+1} - (\mathbf{M} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\mathbf{M})\hat{\theta}$ where $\hat{\theta} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathrm{Y}$ is the maximum likelihood estimator for θ. Taking expectation of the ridge estimator yields $\mathbb{E}[\hat{\theta}_{\text{ridge}}] = (\mathbb{I}_{p+1} - (\mathbf{M} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\mathbf{M})\theta$, so we have the bias as the second term within the parentheses:

$$\text{bias}(\hat{\theta}_{\text{ridge}}) = -(\mathbf{M} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\mathbf{M}\theta \tag{4.2}$$

Consider a diagonal ridge matrix that leaves the treatment effect unregularized $\mathbf{M} = \begin{bmatrix} 0 & 0 \\ 0 & \lambda\mathbb{I}_p \end{bmatrix}$. Using this ridge matrix and the block inversion formula for $(\mathbf{M} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}$, the bias for the treatment effect may be expressed as:

$$\text{bias}(\hat{\alpha}_{\text{ridge}}) = -(\mathrm{Z}^T\mathrm{Z})^{-1}\mathrm{Z}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}_p - \mathbf{X}^T\hat{\mathbf{X}}_{\mathrm{Z}}\right)^{-1}\lambda\beta \tag{4.3}$$

where $(\mathrm{Z}^T\mathrm{Z})^{-1}\mathrm{Z}^T\mathbf{X}$ is a p-length vector of coefficients from p univariate regressions of each X_j on Z and $\hat{\mathbf{X}}_{\mathrm{Z}} = \mathrm{Z}(\mathrm{Z}^T\mathrm{Z})^{-1}\mathrm{Z}^T\mathbf{X}$ are the predicted values from these regressions. For all $\lambda > 0$, we see that Equation (4.3) will be nonzero; especially in the case when the X_j's are correlated with the treatment Z (confounding exists). Also pointed out by Hahn et al. [2018a], the treatment effect bias is not a function of the true treatment α, but instead the unknown (and likely nonzero) coefficient vector β. Of course, the OLS estimate of α is obtained when $\lambda = 0$, in which case the estimator is unbiased. In sum, Equation

(4.3) analytically highlights the issue of bias in the treatment effect estimate should a practitioner choose to regularize a linear treatment effect model.

4.2.1 Mitigating regularization-induced confounding

How can we avoid RIC in treatment effect estimation from observational data? We will describe two approaches: (*i*) Controlling for the propensity of Z, and (*ii*) Replacing the treatment with a proxy for random treatment variation *excluding* X. Both of these approaches require the addition of an equation to Model (4.1) that accounts for the relationship between Z and X:

$$\begin{aligned} &\text{Selection equation:} \quad Z_i = \mathrm{X}_i^T\gamma + \epsilon_i, && \epsilon_i \sim N(0, \sigma_\epsilon^2) \\ &\text{Response equation:} \quad Y_i = \alpha Z_i + \mathrm{X}_i^T\beta + \nu_i, && \nu_i \sim N(0, \sigma_\nu^2), \end{aligned} \tag{4.4}$$

thereby learning about confounding through the parameter γ. The first is called the selection equation since it determines which X_i's should be "selected" for controls, and the second is the original response equation. First, we briefly show how including an estimate of the propensity function from the selection equation: $\hat{\mathrm{Z}} \approx \mathbf{X}\hat{\gamma}$ can mitigate bias from RIC. Suppose we augment our covariates with predicted values for the treatment $\hat{\mathbf{X}}_{\text{new}} = \left(\mathrm{Z}\ \hat{\mathrm{Z}}\ \mathbf{X}\right)$. Effectively, we are including information about the predictable variation in the treatment described by the original controls $\mathbf{X}$. Using the same calculations to arrive at Equation (4.3) now unpenalizing the coefficients associated with *both* Z and $\hat{\mathrm{Z}}$, the bias of the treatment effect can be written as:

$$\text{bias}(\hat{\alpha}_{\text{ridge}}) = -\{(\tilde{\mathrm{Z}}^T\tilde{\mathrm{Z}})^{-1}\tilde{\mathrm{Z}}^T\mathbf{X}\}_1 \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}_p - \mathbf{X}^T\hat{\mathbf{X}}_{\mathrm{Z}}\right)^{-1} \lambda\beta \tag{4.5}$$

where $\tilde{\mathrm{Z}} = \left(\mathrm{Z}\ \hat{\mathrm{Z}}\right)$ and $\{\cdot\}_1$ corresponds to the top row of the matrix $\{\cdot\}$. $\{(\tilde{\mathrm{Z}}^T\tilde{\mathrm{Z}})^{-1}\tilde{\mathrm{Z}}^T\mathbf{X}\}_1$ are the coefficients on Z in the two variable regressions of each X_i on $\left(\mathrm{Z}\ \hat{\mathrm{Z}}\right)$. Since the propensity estimate $\hat{\mathrm{Z}}$ accounts for variation in Z due to the controls, the coefficient on Z in these univariate regressions is approximately zero which renders the bias of the treatment effect close to zero. This feature will be illustrated in simulations to follow.

Hahn et al. [2018a] discuss a reparameterization of Model (4.6) that allows for regularization via Bayesian shrinkage priors in both equations while mitigating RIC – an alternative to controlling for the propensity of Z. The following parameter transformation

$$\begin{pmatrix} \alpha \\ \beta + \alpha\gamma \\ \gamma \end{pmatrix} \rightarrow \begin{pmatrix} \alpha \\ \beta_d \\ \beta_c \end{pmatrix}$$

results in a new formulation of Model (4.6):

$$\begin{aligned} &\text{Selection equation:} \quad Z_i = \mathrm{X}_i^T\beta_c + \epsilon_i, && \epsilon_i \sim N(0, \sigma_\epsilon^2) \\ &\text{Response equation:} \quad Y_i = \alpha(Z_i - \mathrm{X}_i^T\beta_c) + \mathrm{X}_i^T\beta_d + \nu_i, && \nu_i \sim N(0, \sigma_\nu^2). \end{aligned} \tag{4.6}$$

Conveniently, β_c and β_d nicely separate the roles covariates play in treatment effect estimation. A covariate X_{ij} that is distinctly predictive of the response will have $\beta_{dj} \neq 0$ and $\beta_{cj} = 0$. As common in medicine, this covariate may also be called prognostic. Alternatively, the covariate may be a confounder, in which case $\beta_{cj} \neq 0$, $\beta_{dj} \neq 0$. This formulation provides an intuitive interpretation of the treatment effect. The selection equation provides the variation of the treatment excluding X ($\epsilon_i = Z_i - \mathrm{X}_i^T\beta_c$) that is then used to infer α. In

other words, the residual ϵ_i may be thought of as a "randomized experiment" that we then use to infer the treatment effect.

Examining the bias of the treatment effect under a ridge matrix that leaves the treatment effect unregularized ($\mathbf{M} = \begin{bmatrix} 0 & 0 \\ 0 & \lambda \mathbb{I}_p \end{bmatrix}$) yields:

$$\text{bias}(\hat{\alpha}_{\text{ridge}}) = -(\mathrm{R}^T\mathrm{R})^{-1}\mathrm{R}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbb{I}_p - \mathbf{X}^T\hat{\mathbf{X}}_{\mathrm{Z}}\right)^{-1}\lambda\beta_d \tag{4.7}$$

where $\mathrm{R} = \mathrm{Z} - \mathbf{X}\beta_c$. Further, $(\mathrm{R}^T\mathrm{R})^{-1}\mathrm{R}^T\mathbf{X}$ will be close to the zero vector since $R_i = Z_i - \mathrm{X}_i^T\beta_c$ is independent of X_i. In this case, the treatment likelihood given by the selection equation is crucial in providing information on β_c and thus R_i.

4.3 Regularization using the marginal likelihood

What remains to be discussed is how an analyst should choose the level of regularization. In the ridge regression case, this would amount to choosing two λ's for ridge priors on the coefficients in the treatment and response models shown in Model (4.6). Hahn et al. [2018a] approach this in a Bayesian regression context by regularizing using a variant of the horseshoe prior on the regression coefficients from Carvalho et al. [2010b]:

$$\begin{aligned} \pi(\beta_j) &\propto \frac{1}{v}\log\left(1 + \frac{4}{(\beta_j/v)^2}\right), \\ \pi(v) &\sim \mathrm{C}^+(0,1), \end{aligned} \tag{4.8}$$

where v is a global scale parameter common across all elements $j = 1, \ldots p$, and $\mathrm{C}^+(0,1)$ denotes a folded standard Cauchy distribution. Such priors have

proven empirically to be a fine default choice for regression coefficients: they lack hyperparameters, forcefully separate strong from weak predictors, and exhibit robust predictive performance.

In contrast to Hahn et al. [2018a] choice of priors, we consider using ridge regression to regularize and computationally effective ways to choose the ridge parameter (amount of regularization). With the treatment effect estimate under a ridge prior available in closed form, we will choose the level regularization (in both the treatment and response models) by maximizing marginal likelihoods. To characterize uncertainty, we will use bootstrapping. As a result, our method allows for nonparametric calculation of standard errors for treatment effects when the data is assumed clustered.

4.3.1 Marginal likelihood

In this section, we show how to compute the marginal likelihood of data given a ridge prior. Suppose we have the regression model:

$$\begin{aligned} \mathrm{Y} &= \mathbf{X}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 \mathbb{I}_n), \\ \beta, \sigma^2 &\sim \mathrm{NIG}(0, \mathbf{V}, a, b), \end{aligned} \tag{4.9}$$

where we place a conjugate normal-inverse-gamma prior on the parameters. Let the posterior distribution of parameters be defined as $\beta, \sigma^2 \mid \mathrm{Y} \sim \mathrm{NIG}(\overline{\beta}, \overline{\mathbf{V}}, \overline{a}, \overline{b})$. The marginal likelihood may be expressed as a ratio of the posterior and prior normalizing constants multiplied by non-kernel likelihood constants:

$$p(\mathrm{Y}) = \frac{|\overline{\mathbf{V}}|^{1/2} b^a \Gamma(\overline{a})}{|\mathbf{V}|^{1/2} \overline{b}^{\overline{a}} \Gamma(a)} \frac{1}{\pi^{n/2}}. \tag{4.10}$$

The posterior mean of the coefficients is $\overline{\beta} = \overline{\mathbf{V}}\mathbf{X}^T\mathrm{Y}$ where $\overline{\mathbf{V}} = (\mathbf{V}^{-1} + \mathbf{X}^T\mathbf{X})^{-1}$. We can see how the marginal likelihood may be written in the form of (4.10) by considering the expression for $p(\mathrm{Y})$ directly:

$$p(\mathrm{Y}) = \int p(\mathrm{Y} \mid \beta, \sigma^2)p(\beta \mid \sigma^2)p(\sigma^2)d\beta d\sigma^2. \tag{4.11}$$

Let k_* and C_* denote the kernel and non-kernel constants for distribution $*$, respectively. Also, let Z_* denote the normalizing constant for distribution $*$. Factoring the pdfs in (4.11) into the kernel and non-kernel constants, we obtain:

$$p(\mathrm{Y}) = \int C_{\mathrm{Y}|\beta,\sigma^2}k_{\mathrm{Y}|\beta,\sigma^2}C_{\beta|\sigma^2}k_{\beta|\sigma^2}C_{\sigma^2}k_{\sigma^2}d\beta d\sigma^2. \tag{4.12}$$

Since $p(\beta \mid \sigma^2)$ and $p(\sigma^2)$ are distributions over parameters, the non-kernel constants correspond to the reciprocal of their respective normalizing constants. Also, $\int k_{\mathrm{Y}|\beta,\sigma^2}k_{\beta|\sigma^2}k_{\sigma^2}d\beta d\sigma^2$. is the normalizing constant for the posterior density $Z_{\beta,\sigma^2|\mathrm{Y}} = Z_{\beta|\sigma^2,\mathrm{Y}}Z_{\sigma^2|\mathrm{Y}}$, so we have:

$$\begin{aligned} p(\mathrm{Y}) &= C_{\mathrm{Y}|\beta,\sigma^2}C_{\beta|\sigma^2}C_{\sigma^2}\int k_{\mathrm{Y}|\beta,\sigma^2}k_{\beta|\sigma^2}k_{\sigma^2}d\beta d\sigma^2. \\ &= C_{\mathrm{Y}|\beta,\sigma^2}C_{\beta|\sigma^2}C_{\sigma^2}Z_{\beta|\sigma^2,\mathrm{Y}}Z_{\sigma^2|\mathrm{Y}} \\ &= C_{\mathrm{Y}|\beta,\sigma^2}\frac{Z_{\beta|\sigma^2,\mathrm{Y}}Z_{\sigma^2|\mathrm{Y}}}{Z_{\beta|\sigma^2}Z_{\sigma^2}}. \end{aligned} \tag{4.13}$$

Therefore, the marginal likelihood of the data may be written in terms of the non-kernel constants of the likelihood multiplied by the ratio of the posterior and prior normalizing constants.

4.3.2 Expressing the marginal likelihood using the SVD

Is there a way to make expression (4.10) easier to compute? We consider a singular value decomposition of $\mathbf{X}$ and its effect on Model (4.9) and Marginal likelihood (4.10). Decompose $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{W}^T$ such that $\mathbf{U} \in \mathbb{R}^{n\times n}$ and $\mathbf{W} \in \mathbb{R}^{p\times p}$ are orthonormal and whose columns are the eigenvectors of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$, respectively. $\mathbf{D} \in \mathbb{R}^{n\times p}$ is a diagonal matrix whose diagonal elements $(d_1, \ldots, d_q)$, $q = \min(n, p)$, are the square-rooted *nonzero* eigenvalues of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$. Let $\mathbf{Z} = \mathbf{U}\mathbf{D} = \mathbf{X}\mathbf{W}$. Defining the first q columns of $\mathbf{W}$ as $\mathbf{W}_{(q)} = \mathbf{W}^{(1:q)}$, with $\phi \in \mathbb{R}^{q\times 1}$, we define $\mathbf{W}_{(q)}\phi = \beta$. We also keep the q columns of $\mathbf{Z}$ that are associated with the q nonzero eigenvalues by defining $\mathbf{Z}_{(q)} = \mathbf{Z}^{(1:q)}$. The likelihood may be written as $\mathrm{Y} \sim N(\mathbf{Z}_{(q)}\phi = \mathbf{X}\mathbf{W}_{(q)}\phi, \sigma^2\mathbb{I}_n)$. The OLS estimated regression coefficients on the rotated data space are then given by $\hat{\phi} = (\mathbf{Z}_{(q)}^T\mathbf{Z}_{(q)})^{-1}\mathbf{Z}_{(q)}^T\mathrm{Y} = \mathrm{diag}(d_1^2, \ldots, d_q^2)^{-1}\mathbf{Z}_{(q)}^T\mathrm{Y}$. We put a Gaussian (ridge) prior on parameters of the rotated data space:

$$\phi, \sigma^2 \sim \mathrm{NIG}(0, \lambda^{-1}\mathbb{I}_q, a, b) \tag{4.14}$$

which implies a prior for $\beta, \sigma^2 \sim \mathrm{NIG}(0, \mathbf{W}_{(q)}\lambda^{-1}\mathbb{I}_q\mathbf{W}_{(q)}^T, a, b) = \mathrm{NIG}(0, \lambda^{-1}\mathbb{I}_p, a, b)$. Consideration of the prior on the rotated data model allows for easier computation of the marginal likelihood of the response data Y. Effectively, both priors on ϕ and β are diagonal ridge priors with a regularization parameter given by λ. Our goal in the following section will be to present a fast way to calibrate λ.

4.3.3 Empirical Bayes calibration of the ridge prior

The Marginal likelihood (4.10) may be written on the rotated data space.

$$p(\mathrm{Y}) = \frac{|\overline{\mathbf{V}}^{\phi}|^{1/2} b^{a} \Gamma(\overline{a})}{|\mathbf{V}^{\phi}|^{1/2} \overline{b}^{\overline{a}} \Gamma(a)} \frac{1}{\pi^{n/2}} \tag{4.15}$$

where $\overline{\mathbf{V}}^{\phi}$ and $\mathbf{V}^{\phi}$ are the posterior and prior variance parameters. Expression (4.15) follows the general form of the marginal likelihood (4.13) expressed above. The prior variance is defined in (4.14) as $\mathbf{V}^{\phi} = \lambda^{-1}\mathbb{I}_q$. Under the rotated data likelihood $\mathrm{Y} \sim N(\mathbf{Z}_{(q)}\phi = \mathbf{X}\mathbf{W}_{(q)}\phi, \sigma^2\mathbb{I}_n)$ and joint Prior (4.14), the joint posterior distribution may be written as:

$$\phi, \sigma^2 \sim \mathrm{NIG}(\overline{\phi}_{\lambda}, \overline{\mathbf{V}}^{\phi}, \overline{a}, \overline{b}_{\lambda}) \tag{4.16}$$

with marginal likelihood relevant parameters:

$$\begin{aligned} \overline{\mathbf{V}}^{\phi} &= \mathrm{diag}(\lambda + d_1^2, \ldots, \lambda + d_q^2)^{-1} \\ \overline{b}_{\lambda} &= b + \frac{1}{2}\left(\mathrm{Y}^T\mathrm{Y} - \Sigma_{k=1}^{q} \frac{\hat{\phi}_k^2 d_k^4}{\lambda + d_k^2}\right) \\ \overline{a} &= a + n/2. \end{aligned} \tag{4.17}$$

We are now able to write out an explicit form for the log marginal likelihood expressed in (4.15):

$$\log p(\mathrm{Y}) = \log|\overline{\mathbf{V}}^{\phi}|^{1/2} - \log|\mathbf{V}^{\phi}|^{1/2} + a\log b - \overline{a}\log\overline{b}_{\lambda} + \log\Gamma(\overline{a}) - \log\Gamma(a) - \frac{n}{2}\log\pi. \tag{4.18}$$

Plugging in the relevant quantities from (4.17), we obtain:

$$\log p(\mathrm{Y}) = \frac{1}{2}\Sigma_{k=1}^{q}\left[\log(\lambda) - \log(\lambda + d_k^2)\right] + a\log b - \overline{a}\log\left[b + \frac{1}{2}\left(\mathrm{Y}^T\mathrm{Y} - \Sigma_{k=1}^{q}\frac{\hat{\phi}_k^2 d_k^4}{\lambda + d_k^2}\right)\right] + \log\Gamma(\overline{a}) - \log\Gamma(a) - \frac{n}{2}\log\pi. \tag{4.19}$$

Dropping terms that don't involve λ, we obtain a function $f(\lambda) \propto \log p(\mathrm{Y})$:

$$f(\lambda) = \frac{1}{2}\Sigma_{k=1}^{q}\left[\log(\lambda) - \log(\lambda + d_k^2)\right] - \overline{a}\log\left[b + \frac{1}{2}\left(\mathrm{Y}^T\mathrm{Y} - \Sigma_{k=1}^{q}\frac{\hat{\phi}_k^2 d_k^4}{\lambda + d_k^2}\right)\right] \tag{4.20}$$

Finally, we assume a noninformative prior on σ^2: $a = b = 0_+$ where $0_+ = \lim_{t\downarrow 0} t$. This will simplify the function f to:

$$\begin{aligned} f(\lambda) &= q\log(\lambda) - \Sigma_{k=1}^{q}\log(\lambda + d_k^2) - n\log\left[\frac{1}{2}\left(\mathrm{Y}^T\mathrm{Y} - \Sigma_{k=1}^{q}\frac{\hat{\phi}_k^2 d_k^4}{\lambda + d_k^2}\right)\right] \\ &\propto q\log(\lambda) - \Sigma_{k=1}^{q}\log(\lambda + d_k^2) - n\log\left(\mathrm{Y}^T\mathrm{Y} - \Sigma_{k=1}^{q}\frac{\hat{\phi}_k^2 d_k^4}{\lambda + d_k^2}\right) \end{aligned} \tag{4.21}$$

Our goal is to choose λ so that (4.21) is maximized. In practice, we find that using existing R optimization functions such as `optim()` work well for optimizing (4.21). It is particularly fast since a non-negativity bound ($\lambda > 0$) may be placed on the search space.

4.3.4 Regularization for the treatment effect model

In this section, we link the regularization method presented in Section (4.3) with the two equation treatment effect model in (4.1). We describe the general procedure and present a simulation study.

With marginal likelihood (4.21) in hand, we are able to choose an appropriate λ for regularizing a linear regression. Our goal is to apply this procedure to each of the two regression models for treatment effect estimation restated below:

$$\begin{aligned} &\text{Selection equation:} \quad Z_i = \mathrm{X}_i^T\gamma + \epsilon_i, && \epsilon_i \sim N(0, \sigma_\epsilon^2) \\ &\text{Response equation:} \quad Y_i = \alpha Z_i + \mathrm{X}_i^T\beta + \nu_i, && \nu_i \sim N(0, \sigma_\nu^2). \end{aligned} \tag{4.22}$$

Recalling the original setup, X_i is a p-length vector where p is large, i.e., we have many covariates to control for in each regression. A ridge prior will be placed on the coefficients γ and β in each regression to regularize these large vectors. Also, we must be wary of regularization-induced confounding and mitigate it by including a propensity estimate, so the response equation is augmented to include this estimate from the first stage selection equation. In each regularization step, the λ parameters are chosen by the empirical Bayes approach described above. Specifically, the steps we follow are given below (all symbols correspond to observed data vectors of the covariates, treatment, and response):

1. Infer model $\mathrm{Z} \mid \mathbf{X}$ with a ridge prior on γ.
2. Extract predicted values $\hat{\mathrm{Z}}$ from step 1.
3. Define $\tilde{\mathbf{X}} = \left[\mathrm{Z}\ \hat{\mathrm{Z}}\ \mathbf{X}\right]$.
4. Infer model $\mathrm{Y} \mid \tilde{\mathbf{X}}$ with a flat prior on coefficients for Z and $\hat{\mathrm{Z}}$ and a ridge prior on coefficients for $\mathbf{X}$. This amounts to augmenting the response Model (4.22) with the additional "covariate" $\hat{\mathrm{Z}}$.
5. Extract treatment effect α as coefficient for Z.

4.3.5 Uncertainty characterization

Point estimates for the treatment effect in a ridge regression model are given in closed once the regularization parameter is known. For the regression of Y on Z and X, the ridge estimator for the coefficient vector is similar to OLS, with the addition of a "ridge matrix" given by $\mathbf{M}(\lambda)$:

$$\hat{\theta}(\lambda) = (\mathbf{X}^T\mathbf{X} + \mathbf{M}(\lambda))^{-1}\mathbf{X}^T\mathrm{Y}. \tag{4.23}$$

where $\theta = [\alpha \ \beta]$ is the overall coefficient vector. The ridge matrix is often taken to be diagonal with the first entry set to zero to leave the treatment effect estimate (coefficient on Z) unpenalized. The remaining diagonal entries are set equal to λ. Conditional on a calculated value of λ from on the marginal likelihood maximization, we can compute a point estimate of the treatment effect by extracting the first component of β from vector (4.23). In other words, $\hat{\alpha} = [\hat{\theta}(\lambda)]_1$ where $[\cdot]_1$ corresponds to the first component of the inputted vector.

In order to characterize uncertainty and compute a standard error for our treatment effect estimate, we encapsulate this procedure within a bootstrap. Bootstrapping is a simple technique for uncertainty characterization in which observed data is resampled with replacement. For each resample, the statistic of interest is computed and recorded. The process of resampling and recording many times builds a sampling distribution from which a standard error may be computed.

There are outstanding issues with this current approach. First, the calculation of λ for the full sample versus individual bootstrap samples will be materially different. For example, the λ for the full sample may be smaller then individual bootstrap samples since bootstrap samples have ties and omissions of observations. Therefore, the bootstrapped treatment effect estimates with "larger λ's" are often overshrunk, and the resulting sampling distribution is not centered around the full sample treatment effect estimate. To combat this, there are bias-correcting approaches that one can employ, such as those proposed in Efron [1987]. However, the formulas proposed here rely on the full sample statistic to be within the sampling domain of the bootstrapped estimates, which is not always the case for our procedure. In order to compute properly centered sampling distributions, we compute λ once using the full sample of data and use this same level of regularization for the bootstrapped samples as well. Although imprecise since uncertainty in the choice of λ itself is not taken into account, we will see in the simulated examples that we obtain close to expected coverage rates with this approach. In the analysis to follow, we will demonstrate the procedure on simulated data and compare to Hahn et al. [2018a].

4.4 Hahn et al. [2018a] simulation study

In this section, we demonstrate our procedure on simulated data using the data generating process from Hahn et al. [2018a].

We recapitulate the design of the simulation here. The simulation de-

signed to capture a variety of scenarios a data analyst may face. We consider the relative strengths of the confounding and direct effects as well as the number of such variables. Specifically, we use the two equation model that is reparameterized to generate our data. Recall that this reparameterization allows to explicitly control the direct and confounding effects through β_d and β_c respectively; making it ideal for generating data to test and understand our estimation methodology.

$$\begin{aligned} &\text{Selection equation:} \quad Z_i = \mathrm{X}_i^T \beta_c + \epsilon_i, && \epsilon_i \sim N(0, \sigma_\epsilon^2) \\ &\text{Response equation:} \quad Y_i = \alpha(Z_i - \mathrm{X}_i^T \beta_c) + \mathrm{X}_i^T \beta_d + \nu_i, && \nu_i \sim N(0, \sigma_\nu^2). \end{aligned} \tag{4.24}$$

We set the marginal variance of the treatment and response variables to one, $\mathrm{var}(Z) = \mathrm{var}(Y) = 1$, and we center and scale the control variables X to have mean zero and unit variance.

To ensure we consider a range of data compositions, we parametrize our simulations using an ANOVA style decomposition. Defining the ℓ-2 norms (squared Euclidean distance) of the confounding and direct effects as $\rho^2 = \|\beta_c\|_2^2$ and $\phi^2 = \|\beta_d\|_2^2$, we may decompose the marginal variances as

$$\begin{aligned} \mathrm{var}(Z) &= \rho^2 + \sigma_\epsilon^2 \\ \mathrm{var}(Y) &= \alpha^2(1 - \rho^2) + \phi^2 + \sigma_\nu^2, \\ &= \kappa^2 + \phi^2 + \sigma_\nu^2, \end{aligned} \tag{4.25}$$

because the control variables are standardized. Fixing the marginal variances to one implies $\sigma_\epsilon^2 = 1 - \rho^2$ and $\sigma_\nu^2 = 1 - \alpha^2(1 - \rho^2) - \phi^2$. This decomposition admits the following interpretation: ρ^2 is the percentage of the treatment's

variance due to confounding (strength of the confounding effect), ϕ^2 is the percentage of the response variance due to the direct impact of the control variables on the response (strength of the direct effect), and $\kappa^2 := \alpha^2(1-\rho^2)$ is the percentage of the response variance due to quasi-experimental variation of the treatment variable.

Next, observe that as the confounding becomes stronger (ρ^2 getting larger), the independent variation from which we infer the treatment effect ($Z - \mathrm{X}\beta_c$) becomes smaller ($1-\rho^2$). This means that for a fixed level of treatment effect, α, and a fixed marginal variance, stronger confounding makes treatment effect inference harder in that the residual variance becomes correspondingly larger: $1 - \alpha^2(1-\rho^2) - \phi^2$. This makes it more difficult to get a clear picture of whether or not the confounding *per se* is making the problem difficult, or if problems with strong confounding just happen to be more difficult in this artificial way. To avoid this problem, we fix $\kappa^2 := \alpha^2(1-\rho^2)$ to a constant, and allow α to vary as ρ^2 is varied. In this way we can examine the impact of confounding for a fixed difficulty of inference (as measured by the residual variance, which is held fixed at $1 - \kappa^2 - \phi^2$).

In Hahn et al. [2018a] simulations, they fix a decomposition of the response variance given in (4.25) and vary the strength of the confounding effect, ρ^2. In the exercise below, we fix ρ^2 and consider several other types of data generating processes. This amounts to specifying values for κ^2, ϕ^2, and σ^2_ν that sum to one and considering a value of ρ^2 between 0 and 1. Again, because $\kappa^2 = \alpha^2(1-\rho^2)$ is fixed, as ρ^2 varies, α will vary as well.

Next, the components of β_c and β_d must be specified. The nonzero entries of each identify which X_i's are confounders, direct effects, and both, as previously defined. We define the first k elements of X to be confounders, the next k to be both confounders *and* direct effects, and the final k elements to be direct effects. We achieve this in our simulation by setting $\beta_c^{1:2k}$ to ones and $\beta_d^{(k+1):3k} \sim \mathrm{N}(0,1)$. These vectors are then rescaled to have magnitudes ρ^2 and ϕ^2, respectively. This sets the overall β vector ($\beta = \beta_d - \alpha\beta_c$) to have $3k$ nonzero entries. (Note that under continuous priors for β_c and β_d, every variable is a confounder and no variables are strictly prognostic.)

4.4.1 DGP specifications and simulation results

Let n be the number of observations and p be the number of columns of X. In our simulations, we consider the following $\{n, p\}$ pairs: $\{n = 50, p = 30\}$, $\{n = 100, p = 30\}$, $\{n = 100, p = 60\}$, $\{n = 100, p = 95\}$, $\{n = 200, p = 175\}$, and $\{n = 300, p = 200\}$. Additionally, we focus on "strong confounding" situations, so we set $\rho^2 = 0.9$. Our results for weaker levels of confounding are similar to Hahn et al. [2018a], so we omit them for brevity. Finally, we consider the following response variance decompositions: $\{\kappa^2 = 0.05, \phi^2 = 0.7, \sigma_\nu^2 = 0.25\}$ and $\{\kappa^2 = 0.05, \phi^2 = 0.05, \sigma_\nu^2 = 0.9\}$ to mimic low and high residual noise environments, respectively. In the first scenario, the direct effect drives 70% of variance in the response while the treatment effect drives 5%. In the second scenario, the treatment and direct effects are weak while the noise is strong. In all simulated data sets, k as specified above is set to 3. These

two response variance decompositions are shown for the lower dimensional data sets ($n = 50, 100$ and $p = 30, 60$). For the higher dimensional data sets ($n = 100, p = 95$, $n = 200, p = 175$ and $n = 300, p = 200$), we only show results for the low noise decomposition ($\sigma_\nu^2 = 0.25$). In all DGP scenarios, we simulate 2000 data sets and display the average of the following metrics for the treatment effect estimate: Bias, coverage, interval length (I.L.), and mean squared error (MSE).

In each table, we show five separate estimation methods.

1. **New** – corresponds to the new methodology detailed in this thesis. Regularization through ridge regression and penalty parameter selection through marginal likelihood maximization. Standard errors are computed via bootstrapping.

2. **Shrinkage Bayes** – corresponds to the Bayesian approach described in Hahn et al. [2018a] that mitigates RIC using the two equation model (4.24).

3. **Naive Shrinkage Bayes** – corresponds to the naive regularization approach using Bayesian shrinkage priors; also presented in Hahn et al. [2018a]. In this case, shrinkage is only applied to regression coefficients in the response model.

4. **OLS** – ordinary least squares estimation controlling for all covariates.

5. **Oracle OLS** – ordinary least squares estimation controlling only for the $3k$ covariates that are confounders.

Posterior inference for the **Shrinkage Bayes** and **Naive Shrinkage Bayes** follows the methodology of Hahn et al. [2018a], including use of an elliptical slice sampler for posterior exploration of the regression coefficients. Hahn et al. [2018b] provides details on the algorithm, and it is easily implemented with the `bayeslm` package in R.

Tables (4.1) and (4.2) show results for the variance decompositions $\{\kappa^2 = 0.05, \phi^2 = 0.7, \sigma_\nu^2 = 0.25\}$ and $\{\kappa^2 = 0.05, \phi^2 = 0.05, \sigma_\nu^2 = 0.9\}$ respectively with $\{n = 50, p = 30\}$. This is the smallest data set considered. The four metrics we evaluate are bias, coverage, interval length (I.L.), and mean squared error (MSE). First, note the poor performance of the naive shrinkage Bayes approach. This method is severely biased due to regularization-induced confounding (RIC) and is the key finding in Hahn et al. [2018a]. Importantly, the interval for the naive approach is small too, indicating that this approach is confident about the wrong answer! This small interval length is a result of the shrinkage prior.

Note the differences and similarities between the new and shrinkage Bayes approach. These are two methods that mitigate RIC by considering *both* the treatment and response likelihoods conditioned on covariates; the former using ridge priors with the empirical Bayes choice for the regularization parameter, and the latter using Bayesian shrinkage priors. The interval length

for the new method is the largest, and is even slightly larger than OLS. The MSE is lowest for the shrinkage Bayes approach, and the new method does marginally better than OLS.

Method	Bias	Coverage	I.L.	MSE
New	0.003	0.972	1.7249	0.1319
Shrinkage Bayes	-0.0667	0.96	1.1595	0.0747
Naive Shrinkage Bayes	-0.5338	0.1165	0.4938	0.3216
OLS	0.0046	0.9305	1.4159	0.1394
Oracle OLS	0.0029	0.946	0.9751	0.0614

Table 4.1: $\mathbf{n = 50, p = 30, k = 3}$. $\kappa^2 = 0.05$. $\phi^2 = 0.7$. $\sigma_\nu^2 = 0.25$.

Method	Bias	Coverage	I.L.	MSE
New	0.0049	0.9305	2.5603	0.4621
Shrinkage Bayes	-0.1287	0.9365	1.8524	0.2357
Naive Shrinkage Bayes	-0.6218	0.0185	0.2739	0.4323
OLS	0.0022	0.939	2.7212	0.5056
Oracle OLS	0.0062	0.9425	1.8668	0.234

Table 4.2: $\mathbf{n = 50, p = 30, k = 3}$. $\kappa^2 = 0.05$. $\phi^2 = 0.05$. $\sigma_\nu^2 = 0.9$.

Tables (4.3), (4.4), (4.5), and (4.6) show results for the four combinations of the variance decompositions $\{\kappa^2 = 0.05, \phi^2 = 0.7, \sigma_\nu^2 = 0.25\}$ and $\{\kappa^2 = 0.05, \phi^2 = 0.05, \sigma_\nu^2 = 0.9\}$ and data dimensions $\{n = 100, p = 30\}$ and $\{n = 100, p = 60\}$. The rational for looking at these examples is to determine the effect of increasing the number of covariates relative to the number of observations. For the $p = 30$ scenarios shown in Tables (4.3) and (4.4), the new approach is essentially the same as OLS and performs ever so slightly worse than the shrinkage Bayes approach.

Similar conclusions may be drawn when looking at the $p = 60$ scenarios displayed in Tables (4.5) and (4.6). The new approach has gains in interval length and MSE relative to OLS, but is still inferior to the strong regularization imposed in the shrinkage Bayes approach.

Method	Bias	Coverage	I.L.	MSE
New	-9e-04	0.957	0.8082	0.0368
Shrinkage Bayes	-0.0091	0.9575	0.7381	0.031
Naive Shrinkage Bayes	-0.4648	0.213	0.4779	0.2506
OLS	-8e-04	0.942	0.7437	0.0369
Oracle OLS	-0.001	0.9455	0.6513	0.028

Table 4.3: $\mathbf{n = 100, p = 30, k = 3}$. $\kappa^2 = 0.05$. $\phi^2 = 0.7$. $\sigma_\nu^2 = 0.25$.

Method	Bias	Coverage	I.L.	MSE
New	0.0035	0.9495	1.3949	0.1248
Shrinkage Bayes	0.0019	0.948	1.2844	0.1083
Naive Shrinkage Bayes	-0.6164	0.001	0.1981	0.4017
OLS	0.0039	0.949	1.4135	0.1253
Oracle OLS	0.0045	0.948	1.2376	0.097

Table 4.4: $\mathbf{n = 100, p = 30, k = 3}$. $\kappa^2 = 0.05$. $\phi^2 = 0.05$. $\sigma_\nu^2 = 0.9$.

Method	Bias	Coverage	I.L.	MSE
New	-3e-04	0.9735	1.1666	0.064
Shrinkage Bayes	-4e-04	0.9475	0.7453	0.0362
Naive Shrinkage Bayes	-0.4833	0.1275	0.3401	0.2735
OLS	0.0029	0.9395	0.9911	0.0661
Oracle OLS	5e-04	0.9425	0.6527	0.0279

Table 4.5: $\mathbf{n = 100, p = 60, k = 3}$. $\kappa^2 = 0.05$. $\phi^2 = 0.7$. $\sigma_\nu^2 = 0.25$.

Method	Bias	Coverage	I.L.	MSE
New	0.0096	0.935	1.7725	0.2204
Shrinkage Bayes	-0.036	0.9395	1.2953	0.1179
Naive Shrinkage Bayes	-0.6106	0.005	0.2412	0.3978
OLS	0.0136	0.9435	1.8797	0.2375
Oracle OLS	8e-04	0.9565	1.2354	0.0986

Table 4.6: $\mathbf{n = 100, p = 60, k = 3}$. $\kappa^2 = 0.05$. $\phi^2 = 0.05$. $\sigma_\nu^2 = 0.9$.

Gains in the new method over shrinkage Bayes are noticeable in scenarios where the number of covariates approaches the number of observations. Tables (4.7) and (4.8) show results for data dimensions $\{n = 100, p = 95\}$ and $\{n = 200, p = 175\}$, respectively. In both scenarios, the response variance decomposition is $\{\kappa^2 = 0.05, \phi^2 = 0.7, \sigma_\nu^2 = 0.25\}$. Notice that the interval length in the shrinkage Bayes approach becomes exceedingly small, and the coverage moves lower than the expected 95%. In these many covariate scenarios, the benefits of "betting on sparsity" through regularization are traded off with a potential for too much bias.

Notice also that the MSE for the new method is decreased relative to OLS, and is closer to the shrinkage Bayes approach, especially for the scenario with 200 observations and 175 covariates. The new approach maintains minimal bias, improved MSE, and proper coverage relative to OLS and shrinkage Bayes.

Method	Bias	Coverage	I.L.	MSE
New	-0.003	0.9735	1.8165	0.1582
Shrinkage Bayes	-0.0782	0.7695	0.7083	0.0903
Naive Shrinkage Bayes	-0.4003	0.155	0.2867	0.217
OLS	-0.0109	0.87	3.1099	0.9439
Oracle OLS	-0.0013	0.9485	0.6538	0.0281

Table 4.7: $\mathbf{n = 100, p = 95, k = 3}$. $\kappa^2 = 0.05$. $\phi^2 = 0.7$. $\sigma_\nu^2 = 0.25$.

Method	Bias	Coverage	I.L.	MSE
New	-0.0033	0.97	1.1255	0.0612
Shrinkage Bayes	-0.0084	0.8385	0.5136	0.0342
Naive Shrinkage Bayes	-0.2494	0.258	0.2188	0.1217
OLS	-0.0099	0.938	1.2714	0.1119
Oracle OLS	-0.0024	0.9475	0.4505	0.0133

Table 4.8: $\mathbf{n = 200, p = 175, k = 3}$. $\kappa^2 = 0.05$. $\phi^2 = 0.7$. $\sigma_\nu^2 = 0.25$.

4.5 Clustered data

A common issue when calculating the standard error of statistical estimates is that data may be clustered. For example, we might wish to model city level crime rate data by several economic variables, but the errors in this model may be similar for all observations in a given state. In a regression setting where ordinary least squares is used and clustering is ignored, the resulting default standard errors will underestimate the true OLS standard errors. See, for example, Moulton [1986] and Moulton [1990].

One of the most common corrections is to compute cluster-robust standard errors. These provide a more flexible alternative to restrictive random-

effects models, and the adjustment was developed in White [2014] and Arellano [1987]. However, the limitation of cluster-robust standard errors is that they are only asymptotically justified. In other words, the correctness of the procedure assumes that the number of clusters goes to infinity. This may be problematic for policy related studies where there may be only a few clusters like states, counties, or regions.

When there are a small number of clusters, estimates of the standard errors are usually downward biased, see for example Kezdi [2003]. A natural next step would be to correct this bias, and several approaches in statistics attempt to do this, including Kauermann and Carroll [2001], Angrist and Lavy [2009], and Bell and McCaffrey [2002]. This will often make a difference, but hypothesis tests for significance based on a Wald statistic (where $\hat{\beta}$ is the estimate, β is the proposed value, and s.e.$(\hat{\beta})$ is the standard error of the estimate)

$$(\hat{\beta} - \beta)/\text{s.e.}(\hat{\beta}) \tag{4.26}$$

and standard normal critical values will still over-reject (see, for example, Bertrand et al. [2004]).

The goal of this work will be to propose an alternative to typical asymptotic correction methods for standard error calculation, and we attempt to do this using a bootstrap. In the context of clustered data, our bootstrap will involve resampling data in clusters, often called *block bootstrapping*. Cameron et al. [2008] provide a great review of bootstrap methods for computing stan-

dard errors, especially in the presence of clustered data. Efron [1987] also provides a bias correction procedure for bootstrap-based confidence intervals which we incorporate into our approach. This work will combine the new regularized estimation approach presented above with block bootstrapping for standard error calculation. Since the new approach is computationally efficient, our goal is to apply it to large data sets in social science and corporate finance that likely have clustered observations and a large number of covariates.

Chapter 5

Conclusion

This thesis introduced a new approach to model selection called utility-based selection (UBS) and applied it to common problems in econometrics and finance. The first chapter considered a venerable problem in finance of portfolio selection. We developed a methodology for dynamic portfolio construction that tied together a dynamic model for asset returns and the mean-variance portfolio criterion. The methodology emphasized an important distinction between statistical modeling of the optimization inputs and the optimization (utility specification) procedure. While most portfolio selection methods focus on one of these two "sub-problems," our method weaved both together in a principled way. There are several avenues for future work, including different utility and model specifications and designing the procedure to implementable for individual investors to compete with robo-investment advisors like Wealthfront and Betterment.

The second chapter addressed the modeling of asset returns in financial markets. The first section presented a seemingly unrelated regression model for describing the variation in asset returns with commonly used "factors" proposed in the finance literature. The selection procedure accounted for sta-

tistical uncertainty in responses *and* predictors. The second section described ongoing work in monotonic function estimation with an application to predictive regressions for describing returns by firm characteristics. Future work includes taking a deeper dive into the finance theory of asset pricing by considering, for example, the GRS test for pricing efficiency [Gibbons et al., 1989] as well as mutual fund benchmarking by calculating active alphas from passive benchmarks. There is exciting work ahead for the monotonic function estimation project.

The third chapter considered the use of regularization in treatment effect estimation. Regularization-induced confounding (RIC) – a pitfall of naively deploying regularization in treatment effect models and first introduced in Hahn et al. [2018a] – was reviewed. Two reparameterizations for mitigating RIC were discussed, and a new empirical-Bayes approach for treatment effect estimation was introduced. This new approach is encapsulated in a bootstrap for uncertainty characterization, and we presented several simulations to compare the new method to Hahn et al. [2018a] and other alternatives. The new approach performs best when the number of covariates approaches the number of observations. There is much promise for ongoing research. The next steps include uncertainty characterization when the data are clustered by using a block bootstrap.

The power of regularization to tame complex models and help with model interpretation is undeniable. This thesis developed ways in which regularization's effects can be studied in light of statistical uncertainty. Addition-

ally, we considered estimation tasks in econometrics where regularization must be carefully implemented. As more companies, policy makers, and managers use data to make decisions, we hope the developments in this thesis shed light on the tradeoffs between complexity, predictability, and model interpretability.

Appendix

Appendix 1

Regularization in SURs Appendix

1.1 Matrix-variate Stochastic Search

1.1.1 Modeling a full residual covariance matrix

In order to sample a full residual covariance matrix, we augment the predictor matrix with a latent factor f by substituting $\epsilon_j = b_j f + \tilde{\epsilon}_j$:

$$Y_j = \beta_{j1} X_1 + \cdots + \beta_{jp} X_p + b_j f + \tilde{\epsilon}_j, \qquad \tilde{\boldsymbol{\epsilon}} \sim \mathrm{N}(0, \tilde{\Psi}),$$

where $\tilde{\Psi}$ is now diagonal. Assuming that $f \sim N(0, 1)$ is shared among all response variables j and $\mathbf{b} \in \mathbb{R}^{qx1}$ is a vector of all coefficients b_j, the total residual variance may be expressed as:

$$\Psi = \mathbf{b}\mathbf{b}^T + \tilde{\Psi}.$$

We incorporate this latent factor model into the matrix-variate MCMC via a simple Gibbs step to draw posterior samples of f. This augmentation allows us to draw samples of Ψ that are not constrained to be diagonal.

1.1.2 Modeling the marginal distribution: A latent factor model

We model covariates via a latent factor model of the form:

$$\mathbf{X}_t = \mu_x + \mathbf{B}\mathbf{f}_t + \mathbf{v}_t$$

$$\mathbf{v}_t \sim \mathrm{N}(0, \boldsymbol{\Lambda}), \quad \mathbf{f}_t \sim \mathrm{N}(0, \mathbb{I}_k), \quad \mu_x \sim \mathrm{N}(0, \Phi)$$

where Λ is assumed diagonal and the set of k latent factors f_t are independent. The covariance of the covariates is constrained by the factor decomposition and takes the form:

$$\Sigma_x = \mathbf{B}\mathbf{B}^T + \Lambda.$$

Recall that this is only a potential choice for the $p(X)$ and it is chosen here primarily motivated by the applied context where financial assets tend to depend across each other through common factors. Our variable selection procedure would follow if any other choice was made at this point. To estimate this model, a convenient, efficient choice is the R package `bfa` [Murray, 2015]. The software allows us to sample the marginal covariance as well as the marginal mean via a simple Gibbs step assuming a normal prior on μ_x.

1.1.3 Modeling the conditional distribution: A matrix-variate stochastic search

We model the conditional distribution, $Y|X$, by developing a multivariate extension of stochastic search variable selection of George and McCulloch [1993]. Recall that the conditional model is: $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \sim \mathcal{N}(\mathbb{I}_{N\times N},\ \Psi_{q\times q})$. In order to sample different subsets of covariates (different models) during the posterior simulations, we introduce an additional parameter $\alpha \in \mathbb{R}^p$ that is a binary vector identifying a particular model. In other words, all entries i for which $\alpha_i = 1$ denote covariate i as included in model M_α. Specifically, we write the model identified by α as $M_\alpha : \mathbf{Y} - \mathbf{X}_\alpha\boldsymbol{\beta}_\alpha \sim \mathcal{N}(\mathbb{I}_{N\times N},\ \Psi_{q\times q})$. As in George and McCulloch [1993], we aim to explore the posterior on the model

space, $\mathbf{P}(M_\alpha \mid \mathbf{Y})$. Our algorithm explores this model space by calculating a Bayes factor for a particular model M_α. Given that the response $\mathbf{Y}$ is matrix instead of a vector, we derive the Bayes factor as a product of vector response Bayes factors. This is done by separating the marginal likelihood of the response matrix as a product of marginal likelihoods across the separate vector responses. This derivation requires our priors to be independent across the responses and is shown in the details to follow. It is important to note that we do not run a standard SSVS on each univariate response regression separately. Instead, we generalize George and McCulloch [1993] and require all covariates to be included or excluded from a model for each of the responses *simultaneously*.

The marginal likelihood requires priors for the parameters $\boldsymbol{\beta}$ and σ parameters in our model. We choose the standard g-prior for linear models because it permits an analytical solution for the marginal likelihood integral [Zellner, 1986a, Zellner and Siow, 1984, Liang et al., 2008a].

Our Gibbs sampling algorithm directly follows the stochastic search variable selection procedure described in George and McCulloch [1993] using these calculated Bayes factors, now adapted to a multivariate setting. The aim is to scan through all possible covariates and determine which ones to include in the model identified through the binary vector α. At each substep of the MCMC, we consider an individual covariate i within a specific model and compute its inclusion probability as a function of the model's prior probability

and the Bayes factors:

$$p_i = \frac{B_{a0}\mathbf{P}\left(M_{\alpha_a}\right)}{B_{a0}\mathbf{P}\left(M_{\alpha_a}\right) + B_{b0}\mathbf{P}\left(M_{\alpha_b}\right)}.$$

The Bayes factor B_{a0} is a ratio of marginal likelihoods for the model with covariate i included and the null model, and B_{b0} is the analogous Bayes factor for the model without covariate i. The prior on the model space, $\mathbf{P}\left(M_\gamma\right)$, can either be chosen to adjust for multiplicity or to be uniform - our results appear robust to both specifications. In this setting, adjusting for multiplicity amounts to putting equal prior mass on different sizes of models. In contrast, the uniform prior for models involving p covariates puts higher probability mass on larger models, reaching a maximum for models with $\binom{p}{2}$ covariates included. The details of the priors on the model space and parameters, including an empirical Bayes choice of the g-prior hyperparameter, are discussed below.

1.1.4 Details

Assume we have observed N realizations of data $(\mathbf{Y}, \mathbf{X})$. For model comparison, we calculate the Bayes factor with respect to the null model without any covariates. First, we calculate a marginal likelihood. This likelihood is obtained by integrating the full model over $\boldsymbol{\beta}_\alpha$ and σ multiplied by a prior, $\pi_\alpha\left(\boldsymbol{\beta}_\alpha, \sigma\right)$, for these parameters. A Bayes factor of a given model α versus the null model, $B_{\alpha 0} = \frac{m_\alpha(\mathbf{R})}{m_0(\mathbf{R})}$ with:

$$m_\alpha\left(\mathbf{Y}\right) = \int \text{Matrix Normal}_{N,q}\left(\mathbf{Y} \mid \mathbf{X}_\alpha\boldsymbol{\beta}_\alpha,\ \mathbb{I}_{NxN},\ \tilde{\Psi}_{q\times q}\right)\pi_\alpha\left(\boldsymbol{\beta}_\alpha, \sigma_i\right)d\boldsymbol{\beta}_\alpha d\sigma_i. \tag{1.1}$$

We assume independence of the priors across columns of $\mathbf{Y}$ so we can write the integrand in (1.1) as a product across each individual response vector:

$$\begin{aligned}
m_\alpha(\mathbf{Y}) &= \int \Pi_{i=1}^{q} N_N\left(Y^i \mid \mathbf{X}_\alpha \beta_\alpha^i,\ \sigma_i^2 \mathbb{I}_{NxN}\right) \pi_\alpha^i\left(\beta_\alpha^i, \sigma_i\right) d\beta_\alpha^i d\sigma_i \\
&\iff \\
m_\alpha(\mathbf{Y}) &= \int N_N\left(Y^1 \mid \mathbf{X}_\alpha \beta_\alpha^1,\ \sigma_1^2 \mathbb{I}_{NxN}\right) \pi_\alpha^1\left(\beta_\alpha^1, \sigma_1\right) d\beta_\alpha^1 d\sigma_1 \\
&\quad \times \cdots \times \int N_N\left(Y^q \mid \mathbf{X}_\alpha \beta_\alpha^q,\ \sigma_q^2 \mathbb{I}_{NxN}\right) \pi_\alpha^q\left(\beta_\alpha^q, \sigma_q\right) d\beta_\alpha^q d\sigma_q \\
&= m_\alpha\left(Y^1\right) \times \cdots \times m_\alpha\left(Y^q\right) \\
&= \Pi_{i=1}^{q} m_\alpha\left(Y^i\right),
\end{aligned}$$

with:

$$Y^i \sim N_N\left(\mathbf{X}_\alpha \beta_\alpha^i,\ \sigma_i^2 \mathbb{I}_{NxN}\right).$$

Therefore, the Bayes factor for this matrix-variate model is just a product of Bayes factors for the individual multivariate normal models.

$$B_{\alpha 0} = \widetilde{B}_{\alpha 0}^1 \times \cdots \times \widetilde{B}_{\alpha 0}^q$$

with:

$$\widetilde{B}_{\alpha 0}^i = \frac{m_\alpha\left(Y^i\right)}{m_0\left(Y^i\right)}.$$

The simplification of the marginal likelihood calculation is crucial for analytical simplicity and for the resulting SSVS algorithm to rely on techniques already developed for univariate response models. In order to calculate the integral for each Bayes factor, we need priors on the parameters $\boldsymbol{\beta}_\alpha$ and σ. Since the

priors are independent across the columns of $\mathbf{Y}$, we aim to define $\pi_\alpha^i(\beta_\alpha^i, \sigma_i)$ $\forall i \in \{1, ..., q\}$, which we express as the product: $\pi_\alpha^i(\sigma_i)\,\pi_\alpha^i(\beta_\alpha^i \mid \sigma_i)$. Motivated by the work on regression problems of Zellner, Jeffreys, and Siow, we choose a non-informative prior for σ_i and the popular g-prior for the conditional prior on β_α^i, [Zellner, 1986a], [Zellner and Siow, 1980], [Zellner and Siow, 1984], [Jeffreys, 1961]:

$$\pi_\alpha^i\left(\beta_\alpha^i, \sigma_i \mid g\right) = \sigma_i^{-1}\mathrm{N}_{k_\alpha}\left(\beta_\alpha^i \mid \mathbf{0}, g_\alpha^i\sigma_i^2(\mathbf{X}_\alpha^T(\mathbb{I} - N^{-1}\mathbf{1}\mathbf{1}^T)\mathbf{X}_\alpha)^{-1}\right). \tag{1.2}$$

Under this prior, we have an analytical form for the Bayes factor:

$$\begin{aligned} B_{\alpha 0} &= \widetilde{B}_{\alpha 0}^1 \times \cdots \times \widetilde{B}_{\alpha 0}^q \\ &= \Pi_{i=1}^q \frac{(1+g_\alpha^i)^{(N-k_\alpha-1)/2}}{\left(1+g_\alpha^i\frac{SSE_\alpha^i}{SSE_0^i}\right)^{(N+1)/2}}, \end{aligned} \tag{1.3}$$

where SSE_α^i and SSE_0^i are the sum of squared errors from the linear regression of column Y^i on covariates $\mathbf{X}_\alpha$ and k_α is the number of covariates in model M_α. We allow the hyper parameter g to vary across columns of $\mathbf{Y}$ and depend on the model, denoted by writing, g_α^i.

We aim to explore the posterior of the model space, given our data:

$$\mathbf{P}(M_\alpha \mid \mathbf{Y}) = \frac{B_{\alpha 0}\mathbf{P}(M_\alpha)}{\Sigma_\alpha B_{\alpha 0}\mathbf{P}(M_\alpha)},$$

where the denominator is a normalization factor. In the spirit of traditional stochastic search variable selection [Garcia-Donato and Martinez-Beneito, 2013], we propose the following Gibbs sampler to sample this posterior.

1.1.5 Gibbs Sampling Algorithm

Once the parameters $\boldsymbol{\beta}_\alpha$ and σ are integrated out, we know the form of the full conditional distributions for $\alpha_i \mid \alpha_1, \cdots, \alpha_{i-1}, \alpha_{i+1}, \cdots, \alpha_p$. We sample from these distributions as follows:

1. Choose column Y^i and consider two models α_a and α_b such that:

$$\alpha_a = (\alpha_1, \cdots, \alpha_{i-1}, 1, \alpha_{i+1}, \cdots, \alpha_p)$$
$$\alpha_b = (\alpha_1, \cdots, \alpha_{i-1}, 0, \alpha_{i+1}, \cdots, \alpha_p)$$

2. For each model, calculate B_{a0} and B_{b0} as defined by (1.3).

3. Sample

$$\alpha_i \mid \alpha_1, \cdots, \alpha_{i-1}, \alpha_{i+1}, \cdots, \alpha_p \sim Ber(p_i)$$

where

$$p_i = \frac{B_{a0}\mathbf{P}\left(M_{\alpha_a}\right)}{B_{a0}\mathbf{P}\left(M_{\alpha_a}\right) + B_{b0}\mathbf{P}\left(M_{\alpha_b}\right)},$$

Using this algorithm, we visit the most likely models given our set of responses. Under the model and prior specification, there are closed-form expressions for the posteriors of the model parameters β_α and σ.

1.1.6 Hyper Parameter for the g-prior

We use a local empirical Bayes to choose the hyper parameter for the g-prior in (1.2). Since we allow g to be a function of the columns of $\mathbf{Y}$ as

well as the model defined by α, we calculate a separate g for each univariate Bayes factor as in (1.2) above. An empirical Bayes estimate of g maximizes the marginal likelihood and is constrained to be non-negative. From Liang et al. [2008b], we have:

$$\hat{g}_\alpha^{EB(i)} = max\{F_\alpha^i - 1, 0\},$$
$$F_\alpha^i = \frac{R_\alpha^{2i}/k_\alpha}{(1 - R_\alpha^{2i})/(N - 1 - k_\alpha)}.$$

For univariate stochastic search, the literature recommends choosing a fixed g as the number of data points Garcia-Donato and Martinez-Beneito [2013]. However, the multivariate nature of our model induced by the vector-valued response makes this approach unreliable. Since each response has distinct statistical characteristics and correlations with the covariates, it is necessary to vary g among different sampled models and responses. We find that this approach provides sufficiently stable estimation of the inclusion probabilities for the covariates.

1.2 Derivation of lasso form

In this section of the Appendix, we derive the penalized objective (lasso) forms of the utility functions. After integration over $p(\tilde{Y}, \tilde{X}, \Theta|\mathbf{Y}, \mathbf{X})$, the utility takes the form (from equation (3.5)):

$$\mathcal{L}(\boldsymbol{\gamma}) = \text{tr}[M\boldsymbol{\gamma}S\boldsymbol{\gamma}^T] - 2\text{tr}[A\boldsymbol{\gamma}^T] + \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1,$$

where $A = \mathbb{E}[\Omega\tilde{Y}\tilde{X}^T]$, $S = \mathbb{E}[\tilde{X}\tilde{X}^T] = \overline{\Sigma_x}$, and $M = \overline{\Omega}$, and the overlines denote posterior means. Defining the Cholesky decompositions: $M = LL^T$ and

$S = QQ^T$, combining the matrix traces, completing the square with respect to $\boldsymbol{\gamma}$, and converting the trace to the vectorization operator, we obtain:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\gamma}) &= \text{tr}[M(\boldsymbol{\gamma} S \boldsymbol{\gamma}^T - 2M^{-1}A\boldsymbol{\gamma}^T] + \lambda \parallel \mathbf{vec}(\boldsymbol{\gamma})\parallel_1 \\
&\propto \text{tr}\left[M(\boldsymbol{\gamma} - M^{-1}AS^{-1})S(\boldsymbol{\gamma} - M^{-1}AS^{-1})^T\right] + \lambda \parallel \mathbf{vec}(\boldsymbol{\gamma})\parallel_1 \\
&= \text{tr}\left[LL^T(\boldsymbol{\gamma} - L^{-T}L^{-1}AS^{-1})S(\boldsymbol{\gamma} - L^{-T}L^{-1}AS^{-1})^T\right] + \lambda \parallel \mathbf{vec}(\boldsymbol{\gamma})\parallel_1 \\
&= \text{tr}\left[L^T(\boldsymbol{\gamma} - L^{-T}L^{-1}AS^{-1})S(\boldsymbol{\gamma} - L^{-T}L^{-1}AS^{-1})^T L\right] + \lambda \parallel \mathbf{vec}(\boldsymbol{\gamma})\parallel_1 \\
&= \text{tr}\left[(L^T\boldsymbol{\gamma} - L^{-1}AQ^{-T}Q^{-1})QQ^T((L^T\boldsymbol{\gamma} - L^{-1}AQ^{-T}Q^{-1})^T\right] + \lambda \parallel \mathbf{vec}(\boldsymbol{\gamma})\parallel_1 \\
&= \text{tr}\left[(L^T\boldsymbol{\gamma}Q - L^{-1}AQ^{-T})(L^T\boldsymbol{\gamma}Q - L^{-1}AQ^{-T})^T\right] + \lambda \parallel \mathbf{vec}(\boldsymbol{\gamma})\parallel_1 \\
&= \mathbf{vec}(L^T\boldsymbol{\gamma}Q - L^{-1}AQ^{-T})^T\mathbf{vec}(L^T\boldsymbol{\gamma}Q - L^{-1}AQ^{-T}) + \lambda \parallel \mathbf{vec}(\boldsymbol{\gamma})\parallel_1 .
\end{aligned}
$$

The proportionality in line 2 is up to an additive constant with respect to the action variable, $\boldsymbol{\gamma}$. We arrive at the final utility by distributing the vectorization and rewriting the inner product as a squared ℓ_2 norm.

$$
\mathcal{L}(\boldsymbol{\gamma}) = \left\| \left[Q^T \otimes L^T\right] \mathbf{vec}(\boldsymbol{\gamma}) - \mathbf{vec}(L^{-1}AQ^{-T}) \right\|_2^2 + \lambda \parallel \mathbf{vec}(\boldsymbol{\gamma})\parallel_1 .
$$

1.3 Derivation of the loss function under fixed predictors

We devote this section to deriving an analogous loss function for multivariate regression when the predictors are assumed fixed. Notice that this is essentially an extension of Hahn and Carvalho [2015] to the multiple response case and adds to the works of Brown et al. [1998] and Wang [2010] by providing a posterior summary strategy that relies on more than just marginal quantities like posterior inclusion probabilities.

Suppose we observe N realizations of the predictor vector defining the design matrix $\mathbf{X} \in \mathbb{R}^{Nxp}$. Future realizations $\tilde{\mathbf{Y}} \in \mathbb{R}^{Nxq}$ at this fixed set of predictors are generated from a matrix normal distribution:

$$\tilde{\mathbf{Y}} \sim \text{Matrix Normal}_{N,q}\left(\mathbf{X}\boldsymbol{\gamma}^T,\ \mathbb{I}_{NxN},\ \Psi_{qxq}\right). \tag{1.4}$$

In this case, the optimal posterior summary $\boldsymbol{\gamma}^*$ minimizes the expected loss $\mathcal{L}_\lambda(\boldsymbol{\gamma}) = \mathbb{E}[\mathcal{L}_\lambda(\tilde{\mathbf{Y}}, \Theta, \boldsymbol{\gamma})]$. Here, the expectation is taken over the joint space of the predictive and posterior distributions: $p(\tilde{\mathbf{Y}}, \Theta|\mathbf{Y}, \mathbf{X})$ *where* $\tilde{X}$ *is now absent since we are relegated to predicting at the observed covariate matrix* $\mathbf{X}$. We define the utility function using the negative kernel of distribution (1.4) where, as before, $\boldsymbol{\gamma}$ is the summary defining the sparsified linear predictor and $\Omega = \Psi^{-1}$:

$$\mathcal{L}_\lambda(\tilde{\mathbf{Y}}, \Theta, \boldsymbol{\gamma}) = \frac{1}{2}\text{tr}\left[\Omega(\tilde{\mathbf{Y}} - \mathbf{X}\boldsymbol{\gamma}^T)^T(\tilde{\mathbf{Y}} - \mathbf{X}\boldsymbol{\gamma}^T)\right] + \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1,$$

Expanding the inner product and dropping terms that do not involve $\boldsymbol{\gamma}$, we define the loss up to proportionality:

$$\mathcal{L}_\lambda(\tilde{\mathbf{Y}}, \Theta, \boldsymbol{\gamma}) \propto \text{tr}\left[\Omega(\boldsymbol{\gamma}\mathbf{X}^T\mathbf{X}\boldsymbol{\gamma}^T - 2\tilde{\mathbf{Y}}^T\mathbf{X}\boldsymbol{\gamma}^T)\right] + \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1.$$

Analogous to the stochastic predictors derivation, we integrate over $(\tilde{Y}, \Theta)$ to obtain our expected loss:

$$\begin{aligned}\mathcal{L}_\lambda(\boldsymbol{\gamma}) &= \mathbb{E}[\mathcal{L}_\lambda(\tilde{\mathbf{Y}}, \Theta, \boldsymbol{\gamma})] \\ &= \text{tr}[M\boldsymbol{\gamma}S_f\boldsymbol{\gamma}^T] - 2\text{tr}[A_f\boldsymbol{\gamma}^T] + \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1.\end{aligned}$$

where, similar to the random predictor case, $A_f = \mathbb{E}[\Omega \tilde{\mathbf{Y}}^T \mathbf{X}]$, $S_f = \mathbf{X}^T\mathbf{X}$, $M = \overline{\Omega}$, and the overlines denote posterior means. The subscript f is used to denote quantities calculated at *fixed* design points $\mathbf{X}$. Defining the Cholesky decompositions: $M = LL^T$ and $S_f = Q_f Q_f^T$, this expression can be formulated in the form of a standard penalized regression problem:

$$\mathcal{L}_\lambda(\boldsymbol{\gamma}) = \left\| \left[Q_f^T \otimes L^T \right] \mathbf{vec}(\boldsymbol{\gamma}) - \mathbf{vec}(L^{-1} A_f Q_f^{-T}) \right\|_2^2 + \lambda \left\| \mathbf{vec}(\boldsymbol{\gamma}) \right\|_1 \qquad (1.5)$$

with covariates $Q_f^T \otimes L^T$, "data" $L^{-1} A_f Q_f^{-T}$, and regression coefficients $\boldsymbol{\gamma}$. Accordingly, (1.5) can be optimized using existing software such as the `lars` R package of Efron et al. [2004a].

We use loss function (1.5) as a point of comparison to demonstrate how incorporating covariate uncertainty may impact the summarization procedure in our applications.

Bibliography

Lucy F Ackert and Yisong S Tian. Arbitrage, liquidity, and the valuation of exchange traded funds. *Financial markets, institutions & instruments*, 17 (5):331–362, 2008.

Anna Agapova. Conventional mutual index funds versus exchange-traded funds. *Journal of Financial Markets*, 14(2):323–343, 2011.

Weihua An. Bayesian propensity score estimators: incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, 40(1): 151–189, 2010.

Joshua Angrist and Victor Lavy. The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American economic review*, 99(4):1384–1414, 2009.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

Manuel Arellano. Computing robust standard errors for within-groups estimators. *Oxford bulletin of Economics and Statistics*, 49(4):431–434, 1987.

Susan Athey and Guido Imbens. Machine learning methods for estimating heterogeneous causal effects. *arXiv preprint arXiv:1504.01132*, 2015.

Brad M Barber and Terrance Odean. Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance*, pages 773–806, 2000.

Brad M Barber, Yi-Tsung Lee, Yu-Jane Liu, and Terrance Odean. Just how much do individual investors lose by trading? *Review of Financial studies*, 22(2):609–632, 2009.

Maria Maddalena Barbieri and James O Berger. Optimal predictive model selection. *Annals of Statistics*, pages 870–897, 2004.

M.J. Bayarri, J.O. Berger, A. Forte, and G. Garcia-Donato. Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577, 2012a.

MJ Bayarri, JO Berger, A Forte, G García-Donato, et al. Criteria for bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577, 2012b.

John E Beasley, Nigel Meade, and T-J Chang. An evolutionary heuristic for the index tracking problem. *European Journal of Operational Research*, 148 (3):621–643, 2003.

Robert M Bell and Daniel F McCaffrey. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2): 169–182, 2002.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

James O. Berger and German Molina. Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica*, 59(1):3–15, 2005.

Jonathan B Berk and Jules H Van Binsbergen. Measuring skill in the mutual fund industry. *Journal of Financial Economics*, 118(1):1–20, 2015.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275, 2004.

Michael J Best and Robert R Grauer. On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results. *Review of Financial Studies*, 4(2):315–342, 1991.

Swati Biswas and Shili Lin. Logistic bayesian lasso for identifying association with rare haplotypes and application to age-related macular degeneration. *Biometrics*, 68(2):587–597, 2012.

Mark Britten-Jones. The sampling error in estimates of mean-variance efficient portfolio weights. *The Journal of Finance*, 54(2):655–671, 1999.

Mark Broadie. Computing efficient frontiers using estimated parameters. *Annals of Operations Research*, 45(1):21–58, 1993.

Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

Lawrence D Brown. An ancillarity paradox which appears in multiple linear regression. *The Annals of Statistics*, pages 471–493, 1990.

Philip J Brown, Marina Vannucci, and Tom Fearn. Multivariate bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):627–641, 1998.

P.J. Brown and Marina Vannucci. Multivariate bayesian variable selection and prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 627–641, 1998.

A Colin Cameron, Jonah B Gelbach, and Douglas L Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008.

Nilgun A Canakgoz and John E Beasley. Mixed-integer programming approaches for index tracking and enhanced indexation. *European Journal of Operational Research*, 196(1):384–399, 2009.

Emmanuel Candes and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

Marine Carrasco and Nérée Noumon. Optimal portfolio selection using regularization. Technical report, Discussion paper, 2011.

Carlos M Carvalho, Mike West, et al. Dynamic matrix-variate graphical models. *Bayesian analysis*, 2(1):69–97, 2007.

Carlos M Carvalho, Jeffrey Chang, Joseph E Lucas, Joseph R Nevins, Quanli Wang, and Mike West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484), 2008.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. In *International Conference on Artificial Intelligence and Statistics*, pages 73–80, 2009.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, page asq017, 2010a.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 2010b.

Carlos M. Carvalho, Jared D. Fisher, and David Puelz. Monotonic effects of characteristics on returns. *preprint*, 2018.

Matias D Cattaneo, Richard K Crump, Max Farrell, and Ernst Schaumburg. Characteristic-sorted portfolios: estimation and inference. 2018.

Chen Chen and Roy H Kwon. Robust portfolio selection for index tracking. *Computers & Operations Research*, 39(4):829–837, 2012.

M. Clyde and E.I. George. Model uncertainty. *Statistical Science*, 19:81–94, 2004.

John H Cochrane. Presidential address: Discount rates. *The Journal of Finance*, 66(4):1047–1108, 2011.

Conceicao and Maechler. Deoptimr, 2015.

D Conniffe. Covariance analysis and seemingly unrelated regressions. *The American Statistician*, 36(3a):169–171, 1982.

Martijn Cremers, Miguel A. Ferreira, Pedro Matos, and Laura Starks. Indexing and active fund management: International evidence. *Journal of Financial Economics*, 120(3):539 – 560, 2016. ISSN 0304-405X. doi: http://dx.doi.org/10.1016/j.jfineco.2016.02.008. URL `//www.sciencedirect.com/science/article/pii/S0304405X16300083`.

CRSP. The center for research in security prices. Wharton Research Data Services, 1992-2016.

Kent Daniel, Mark Grinblatt, Sheridan Titman, and Russ Wermers. Measuring mutual fund performance with characteristic-based benchmarks. *Journal of Finance*, 52(3):1035–58, 1997.

A Philip Dawid. Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68(1):265–274, 1981.

A Philip Dawid and Steffen L Lauritzen. Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, pages 1272–1317, 1993.

Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies*, 22(5):1915–1953, 2007.

Victor DeMiguel, Lorenzo Garlappi, Francisco J Nogales, and Raman Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812, 2009a.

Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies*, 22(5):1915–1953, 2009b.

Arthur P Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

John P Dickinson. The reliability of estimation procedures in portfolio analysis. *Journal of Financial and Quantitative Analysis*, 9(03):447–462, 1974.

James A DiLellio and Keith Jakob. Etf trading strategies to enhance client wealth maximization. *Financial Services Review*, 20(2):145, 2011.

Adrian Dobra, Chris Hans, Beatrix Jones, Joseph R Nevins, Guang Yao, and Mike West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212, 2004.

John J Donohue III and Steven D Levitt. The impact of legalized abortion on crime. *Quarterly Journal of Economics*, pages 379–420, 2001.

Hitesh Doshi, Redouane Elkamhi, and Mikhail Simutin. Managerial activeness and mutual fund performance. *The Review of Asset Pricing Studies*, 5(2): 156, 2015. doi: 10.1093/rapstu/rav005. URL `+http://dx.doi.org/10.1093/rapstu/rav005`.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004a.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004b.

Johan Ericsson and Sune Karlsson. Choosing factors in a multifactor asset pricing model: A bayesian approach. Technical report, Stockholm School of Economics, 2004.

Ashkan Ertefaie, Masoud Asgharian, and David Stephens. Variable selection in causal inference using a simultaneous penalization method. *arXiv preprint arXiv:1511.08501*, 2015.

M. J. Bertin et al. *Pisot and Salem Numbers*. user Verlag, Berlin, 1992.

Eugene F Fama and Kenneth R French. The cross-section of expected stock returns. *the Journal of Finance*, 47(2):427–465, 1992.

Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993a.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3 – 56, 1993b. ISSN 0304-405X. doi: http://dx.doi.org/10.1016/0304-405X(93)90023-5. URL `//www.sciencedirect.com/science/article/pii/0304405X93900235`.

Eugene F. Fama and Kenneth R. French. Luck versus Skill in the Cross-Section of Mutual Fund Returns. *Journal of Finance*, 65(5):1915–1947, October 2010. URL `https://ideas.repec.org/a/bla/jfinan/v65y2010i5p1915-1947.html`.

Eugene F Fama and Kenneth R French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015.

Eugene F Fama and James D MacBeth. Risk, return, and equilibrium: Empirical tests. *The Journal of Political Economy*, pages 607–636, 1973.

Jianqing Fan, Jingjin Zhang, and Ke Yu. Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107 (498):592–606, 2012.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, 2013.

Jianqing Fan, Yuan Liao, and Han Liu. Estimating large covariance and precision matrices. *arXiv preprint arXiv:1504.02995*, 2015.

Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Available at SSRN 2324292*, 2013.

Björn Fastrich, Sandra Paterlini, and Peter Winker. Constructing optimal sparse portfolios using regularization methods. *Computational Management Science*, pages 1–18, 2013a.

Björn Fastrich, Sandra Paterlini, and Peter Winker. Constructing optimal sparse portfolios using regularization methods. *Computational Management Science*, pages 1–18, 2013b.

Björn Fastrich, Sandra Paterlini, and Peter Winker. Cardinality versus q-norm constraints for index tracking. *Quantitative Finance*, 14(11):2019–2032, 2014.

Marcelo Fernandes, Guilherme Rocha, and Thiago Souza. Regularized minimum-variance portfolios using asset group information, 2012.

Wayne E Ferson and Rudi W Schadt. Measuring Fund Strategy and Performance in Changing Economic Conditions. *Journal of Finance*, 51

(2):425–461, June 1996. URL https://ideas.repec.org/a/bla/jfinan/v51y1996i2p425-61.html.

George M Frankfurter, Herbert E Phillips, and John P Seagle. Portfolio selection: the effects of uncertain means, variances, and covariances. *Journal of Financial and Quantitative Analysis*, 6(05):1251–1262, 1971.

Joachim Freyberger, Andreas Neuhierl, and Michael Weber. Dissecting characteristics nonparametrically. Technical report, National Bureau of Economic Research, 2017.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. doi: 10.1093/biostatistics/kxm045. URL http://biostatistics.oxfordjournals.org/content/9/3/432.abstract.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010a.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010b.

Peter A Frost and James E Savarino. An empirical bayes approach to efficient portfolio selection. *Journal of Financial and Quantitative Analysis*, 21(03): 293–305, 1986.

Peter A Frost and James E Savarino. For better performance: Constrain portfolio weights. 1988.

G Garcia-Donato and MA Martinez-Beneito. On sampling strategies in bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association*, 108(501):340–352, 2013.

Lorenzo Garlappi, Raman Uppal, and Tan Wang. Portfolio selection with parameter and model uncertainty: A multi-prior approach. *Review of Financial Studies*, 20(1):41–81, 2007.

Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Debashis Ghosh, Yeying Zhu, and Donna L Coffman. Penalized regression procedures for variable selection in the potential outcomes framework. *Statistics in medicine*, 34(10):1645–1658, 2015.

Daniel Giamouridis and Sandra Paterlini. Regular (ized) hedge fund clones. *Journal of Financial Research*, 33(3):223–247, 2010.

Michael R Gibbons, Stephen A Ross, and Jay Shanken. A test of the efficiency of a given portfolio. *Econometrica: Journal of the Econometric Society*, pages 1121–1152, 1989.

DEA Giles and AC Rayner. The mean squared errors of the maximum likelihood and natural-conjugate bayes regression estimators. *Journal of Econometrics*, 11(2):319–334, 1979.

Jim E Griffin and Philip J Brown. Structuring shrinkage: some correlated priors for regression. *Biometrika*, page asr082, 2012.

Anne Gron, Bjørn N Jørgensen, and Nicholas G Polson. Optimal portfolio choice and stochastic volatility. *Applied Stochastic Models in Business and Industry*, 28(1):1–15, 2012.

Paul Gustafson and Sander Greenland. Curious phenomena in bayesian adjustment for exposure misclassification. *Statistics in medicine*, 25(1):87–103, 2006.

P Richard Hahn and Carlos M Carvalho. Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448, 2015.

P Richard Hahn, Carlos M Carvalho, and Sayan Mukherjee. Partial factor modeling: predictor-dependent shrinkage for linear regression. *Journal of the American Statistical Association*, 108(503):999–1008, 2013.

P. Richard Hahn, Jingyu He, and Hedibert Lopes. Bayesian factor model shrinkage for linear IV regression with many instruments. Technical report, University of Chicago Booth School of Business, 2015.

P Richard Hahn, Carlos M Carvalho, David Puelz, Jingyu He, et al. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 2016a.

P. Richard Hahn, Jingyu He, and Hedibert Lopes. Elliptical slice sampling for Bayesian shrinkage regression with applications to causal inference. Technical report, University of Chicago Booth School of Business, 2016b.

P Richard Hahn, Jared S Murray, and Carlos Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *arXiv preprint arXiv:1706.09523*, 2017.

P. Richard Hahn, Carlos M. Carvalho, David Puelz, and Jingyu He. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 13(1):163–182, 2018a.

P Richard Hahn, Jingyu He, and Hedibert F Lopes. Efficient sampling for gaussian linear regression with arbitrary priors. 2018b.

Chris Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.

Chris Hans, Adrian Dobra, and Mike West. Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102 (478):507–516, 2007.

Jeff Harrison and Mike West. *Bayesian Forecasting & Dynamic Models*. Springer, 1999.

Campbell R Harvey and Yan Liu. Lucky factors. *Available at SSRN 2528780*, 2015.

Campbell R Harvey, Yan Liu, and Heqing Zhu. ? and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68, 2016.

James J Heckman, Hedibert F Lopes, and Rémi Piatek. Treatment effects: A bayesian perspective. *Econometric reviews*, 33(1-4):36–67, 2014.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 2011.

Michael Ho, Zheng Sun, and Jack Xin. Weighted elastic net penalized mean-variance portfolio design and computation. *arXiv preprint arXiv:1502.01658*, 2015.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

An Hongzhi and Gu Lan. On the selection of regression variables. *Acta Mathematicae Applicatae Sinica (English Series)*, 2(1):27–36, 1985.

Mei-Yueh Huang and Jun-Biao Lin. Do etfs provide effective international diversification? *Research in International Business and Finance*, 25(3): 335–344, 2011.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Kaoru Irie and Mike West. Bayesian emulation for optimization in multi-step portfolio decisions. *arXiv preprint arXiv:1607.01631*, 2016.

Liana Jacobi, Helga Wagner, and Sylvia Frühwirth-Schnatter. Bayesian treatment effects models with variable selection for panel outcomes with an application to earnings effects of maternity leave. *Journal of Econometrics*, 193(1):234–250, 2016.

Eric Jacquier and Nicholas Polson. Bayesian econometrics in finance, 2010a.

Eric Jacquier and Nicholas Polson. Simulation-based-estimation in portfolio selection, 2010b.

Eric Jacquier and Nicholas G Polson. Asset allocation in finance: A bayesian perspective. *Hierarchinal models and MCMC: a Tribute to Adrian Smith*, 2012.

Ravi Jagannathan and Tongshu Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. Technical report, National Bureau of Economic Research, 2002.

William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

H Jeffreys. Theory of probability (3rd edt.) oxford university press. 1961.

Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, 48(1):65–91, 1993.

Michael C. Jensen. The Performance Of Mutual Funds In The Period 19451964. *Journal of Finance*, 23(2):389–416, 05 1968. doi: j.1540-6261.1968.tb00815. URL `https://ideas.repec.org/a/bla/jfinan/v23y1968i2p389-416.html`.

Michael C Jensen, Fischer Black, and Myron S Scholes. The capital asset pricing model: Some empirical tests. 1972.

J David Jobson and Bob Korkie. Estimation for markowitz efficient portfolios. *Journal of the American Statistical Association*, 75(371):544–554, 1980.

Michael Johannes, Arthur Korteweg, and Nicholas Polson. Sequential learning, predictability, and optimal portfolio returns. *The Journal of Finance*, 69(2): 611–644, 2014.

Beatrix Jones, Carlos Carvalho, Adrian Dobra, Chris Hans, Chris Carter, and Mike West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4):pp. 388–400, 2005a. URL `http://www.jstor.org/stable/20061200`.

Beatrix Jones, Carlos Carvalho, Adrian Dobra, Chris Hans, Chris Carter, and Mike West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, pages 388–400, 2005b.

Philippe Jorion. International portfolio diversification with estimation risk. *Journal of Business*, pages 259–278, 1985.

Philippe Jorion. Bayes-stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis*, 21(03):279–292, 1986.

Philippe Jorion and Eduardo Schwartz. Integration vs. segmentation in the canadian stock market. *Journal of Finance*, pages 603–614, 1986.

Raymond Kan, Cesare Robotti, and Jay Shanken. Pricing model performance and the two-pass cross-sectional regression methodology. *The Journal of Finance*, 68(6):2617–2649, 2013.

George Karabatsos. Fast marginal likelihood estimation of the ridge parameter (s) in ridge regression and generalized ridge regression for big data. *arXiv preprint arXiv:1409.2437*, 2014.

Göran Kauermann and Raymond J Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396, 2001.

John L Kelly Jr. A new interpretation of information rate. *Information Theory, IRE Transactions on*, 2(3):185–189, 1956.

Gabor Kezdi. Robust standard error estimation in fixed-effects panel models. 2003.

Donald K. Knuth. *The TEXbook*. Addison-Wesley, 1984.

Robert Kosowski, Allan Timmermann, Russ Wermers, and Hal White. Can Mutual Fund "Stars" Really Pick Stocks? New Evidence from a Bootstrap Analysis. *Journal of Finance*, 61(6):2551–2595, December 2006. URL `https://ideas.repec.org/a/bla/jfinan/v61y2006i6p2551-2595.html`.

Leonard Kostovetsky. Index mutual funds and exchange-traded funds. *ETF and Indexing*, 2005(1):88–99, 2005.

Daniel R Kowal, David S Matteson, and David Ruppert. A bayesian multivariate functional dynamic linear model. *Journal of the American Statistical Association*, 112(518):733–744, 2017.

Leslie Lamport. *LaTeX: A document preparation system*. Addison-Wesley, 2nd edition, 1994.

Edward E Leamer. *Specification searches: Ad hoc inference with nonexperimental data*, volume 53. John Wiley & Sons Incorporated, 1978.

Edward E Leamer. Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43, 1983.

Olivier Ledoit and Michael Wolf. Honey, i shrunk the sample covariance matrix. *UPF Economics and Business Working Paper*, (691), 2003a.

Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621, 2003b.

Geneviève Lefebvre, Juli Atherton, and Denis Talbot. The effect of the prior distribution in the bayesian adjustment for confounding algorithm. *Computational Statistics & Data Analysis*, 70:227–240, 2014.

Steven D Levitt and Stephen J Dubner. *Freakonomics*, volume 61. Sperling & Kupfer editori, 2010.

Mingliang Li and Justin L Tobias. Bayesian analysis of treatment effect models. *Bayesian inference in the social sciences*, pages 63–90, 2014.

Yanming Li, Bin Nan, and Ji Zhu. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 71(2):354–363, 2015.

F. Liang, R. Paulo, G. Molina, M. Clyde, and J. Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–423, 2008a.

Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481), 2008b.

Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 2008c.

John Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The review of economics and statistics*, pages 13–37, 1965.

Hedibert Freitas Lopes and Mike West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–68, 2004.

F Mittelbach M Goosens and A Samarin. *The LaTeXCompanion*. Addison-Wesley, 1994.

Harry Markowitz. Portfolio selection*. *The journal of finance*, 7(1):77–91, 1952.

Lawrence C McCandless, Paul Gustafson, and Peter C Austin. Bayesian propensity score analysis for observational data. *Statistics in Medicine*, 28 (1):94–112, 2009.

Daniel McCarthy, Shane T Jensen, et al. Power-weighted densities for time series data. *The Annals of Applied Statistics*, 10(1):305–334, 2016.

Robert McCulloch. Utility based model selection for bayesian nonparametric modeling using trees. SBIES 2015, 2015.

Robert C Merton. On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, 8(4):323–361, 1980.

Attilio Meucci. *Risk and asset allocation*. Springer Science & Business Media, 2009.

Abdolreza Mohammadi and Ernst C Wit. Bdgraph: An r package for bayesian structure learning in graphical models. *arXiv preprint arXiv:1501.05108*, 2015.

Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2014.

Brent R Moulton. Random group effects and the precision of regression estimates. *Journal of econometrics*, 32(3):385–397, 1986.

Brent R Moulton. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The review of Economics and Statistics*, pages 334–338, 1990.

Iain Murray, Ryan Prescott Adams, and David JC MacKay. Elliptical slice sampling. *arXiv preprint arXiv:1001.0175*, 2009.

Jared Murray. bfa, 2015.

Jared S Murray, David B Dunson, Lawrence Carin, and Joseph E Lucas. Bayesian gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665, 2013.

Lilian Ng. Tests of the capm with time-varying covariances: A multivariate garch approach. *The Journal of Finance*, 46(4):1507–1521, 1991.

Mee Young Park and Trevor Hastie. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50, 2008.

Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Lubos Pastor and Pietro Veronesi. Learning in financial markets. Technical report, National Bureau of Economic Research, 2009.

Luboš Pástor, Robert F. Stambaugh, and Lucian A. Taylor. Scale and skill in active management. *Journal of Financial Economics*, 116(1):23 – 45, 2015. ISSN 0304-405X. doi: http://dx.doi.org/10.1016/j.jfineco.2014.11.008. URL //www.sciencedirect.com/science/article/pii/S0304405X14002542.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jie Peng, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R Pollack, and Pei Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The annals of applied statistics*, 4(1):53, 2010.

Anita K Pennathur, Natalya Delcoure, and Dwight Anderson. Diversification benefits of iShares and closed-end country funds. *Journal of Financial Research*, 25(4):541–557, 2002.

Davide Pettenuzzo and Francesco Ravazzolo. Optimal portfolio choice under decision-based model combinations. *Journal of Applied Econometrics*, 2015.

N Polson and B Tew. Bayesian portfolio selection: An analysis of the s&p 500 index 1970-1996. *Journal of Business and Economic Statistics*, 18:164–173, 1999.

James M Poterba and John B Shoven. Exchange traded funds: A new investment option for taxable investors. Technical report, National Bureau of Economic Research, 2002.

David Puelz, Carlos M Carvalho, and P Richard Hahn. Optimal ETF selection for passive investing. *arXiv preprint arXiv:1510.03385*, 2015.

David Puelz, P. Richard Hahn, and Carlos M. Carvalho. Optimal ETF selection for passive investing. 2016.

David Puelz, P. Richard Hahn, and Carlos M. Carvalho. Variable selection in seemingly unrelated regressions with random predictors. *Bayesian Analysis*, 12(4):969–989, 2017.

David Puelz, P. Richard Hahn, and Carlos M. Carvalho. Regret-based selection for sparse dynamic portfolios. 2018.

Justin K Rising and Abraham J Wyner. Partial kelly portfolios and shrinkage estimators. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pages 1618–1622. IEEE, 2012.

James M Robins, Steven D Mark, and Whitney K Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, pages 479–495, 1992.

R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Stephen A Ross. The arbitrage theory of capital asset pricing. *Journal of economic theory*, 13(3):341–360, 1976.

Adam J Rothman, Elizaveta Levina, and Ji Zhu. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962, 2010.

J. Scott and J.O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136:2144–2162, 2006.

William F Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk*. *The journal of finance*, 19(3):425–442, 1964.

William F Sharpe. Mutual fund performance. *Journal of business*, pages 119–138, 1966.

Sangheon Shin and Gökçe Soydemir. Exchange-traded funds, persistence in tracking errors and information dissemination. *Journal of Multinational Financial Management*, 20(4):214–234, 2010.

Thomas S Shively, Thomas W Sager, and Stephen G Walker. A bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):159–175, 2009.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22 (2):231–245, 2013.

Agniva Som, Christopher M Hans, and Steven N MacEachern. Block hyper-g priors in bayesian regression. *arXiv preprint arXiv:1406.6419*, 2014.

TP Speed and HT Kiiveri. Gaussian markov distributions over finite graphs. *The Annals of Statistics*, pages 138–150, 1986.

Michael Spivak. *The joy of TEX*. American Mathematical Society, Providence, R.I., 2nd edition, 1990.

Jason L Stein, Xue Hua, Suh Lee, April J Ho, Alex D Leow, Arthur W Toga, Andrew J Saykin, Li Shen, Tatiana Foroud, Nathan Pankratz, et al. Voxel-wise genome-wide association study (vgwas). *neuroimage*, 53(3):1160–1174, 2010.

Akiko Takeda, Mahesan Niranjan, Jun-ya Gotoh, and Yoshinobu Kawahara. Simultaneous pursuit of out-of-sample performance and sparsity in index tracking portfolios. *Computational Management Science*, 10(1):21–49, 2013.

Denis Talbot, Geneviève Lefebvre, and Juli Atherton. The bayesian causal effect estimation algorithm. *Journal of Causal Inference*, 3(2):207–236, 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Berwin A Turlach, William N Venables, and Stephen J Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.

Alf J. van der Poorten. Some problems of recurrent interest. Technical Report 81-0037, School of Mathematics and Physics, Macquarie University, North Ryde, Australia 2113, August 1981.

Chi Wang, Giovanni Parmigiani, and Francesca Dominici. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68(3):661–671, 2012.

Chi Wang, Francesca Dominici, Giovanni Parmigiani, and Corwin Matthew Zigler. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics*, 2015.

Hao Wang. Sparse seemingly unrelated regression modelling: Applications in finance and econometrics. *Computational Statistics & Data Analysis*, 54 (11):2866–2877, 2010.

Hao Wang. Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Anal.*, 10(2):351–377, 06 2015. doi: 10.1214/14-BA916. URL http://dx.doi.org/10.1214/14-BA916.

Hao Wang and Mike West. Bayesian analysis of matrix normal graphical models. *Biometrika*, 96(4):821–834, 2009.

Hao Wang, Craig Reeson, and Carlos M. Carvalho. Dynamic financial index models: Modeling conditional dependencies via graphs. *Bayesian Anal.*, 6 (4):639–664, 12 2011a. doi: 10.1214/11-BA624. URL http://dx.doi.org/10.1214/11-BA624.

Hao Wang, Craig Reeson, Carlos M Carvalho, et al. Dynamic financial index models: Modeling conditional dependencies via graphs. *Bayesian Analysis*, 6(4):639–664, 2011b.

Herbert I Weisberg and Victor P Pontes. Post hoc subgroups in clinical trials: Anathema or analytics? *Clinical Trials*, page 1740774515588096, 2015.

Russ Wermers. Mutual Fund Performance: An Empirical Decomposition into Stock-Picking Talent, Style, Transactions Costs, and Expenses. *Journal of Finance*, 55(4):1655–1703, 08 2000. URL https://ideas.repec.org/a/bla/jfinan/v55y2000i4p1655-1703.html.

Mike West. On scale mixtures of normal distributions. *Biometrika*, 74(3): 646–648, 1987.

Halbert White. *Asymptotic theory for econometricians*. Academic press, 2014.

Ander Wilson and Brian J Reich. Confounder selection via penalized credible regions. *Biometrics*, 70(4):852–861, 2014.

Jesse Windle, Carlos M Carvalho, et al. A tractable state-space model for symmetric positive-definite matrices. *Bayesian Analysis*, 9(4):759–792, 2014.

Jeffrey Wooldridge. *Introductory econometrics: A modern approach*. Cengage Learning, 2012.

Lan Wu and Yuehan Yang. Nonnegative elastic net and application in index tracking. *Applied Mathematics and Computation*, 227:541–552, 2014.

Lan Wu, Yuehan Yang, and Hanzhong Liu. Nonnegative-lasso and application in index tracking. *Computational Statistics & Data Analysis*, 70:116–126, 2014.

Lingzhou Xue, Shiqian Ma, and Hui Zou. Positive-definite 1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107(500):1480–1491, 2012.

Yu-Min Yen. A note on sparse minimum variance portfolios and coordinate-wise descent algorithms. Technical report, 2013.

Yu-Min Yen and Tso-Jung Yen. Solving norm constrained portfolio optimization via coordinate-wise descent algorithms. *Computational Statistics & Data Analysis*, 76:737–759, 2014.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Ming Yuan, Ali Ekici, Zhaosong Lu, and Renato Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3): 329–346, 2007.

Leonid Zamdborg and Ping Ma. Discovery of protein–dna interactions by penalized multivariate regression. *Nucleic acids research*, page gkp554, 2009.

Arnold Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368, 1962.

Arnold Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243, 1986a.

Arnold Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243, 1986b.

Arnold Zellner and Tomohiro Ando. A direct monte carlo approach for bayesian analysis of the seemingly unrelated regression model. *Journal of Econometrics*, 159(1):33–45, 2010.

Arnold Zellner and A Siow. *Basic issues in econometrics*. University of Chicago Press Chicago, 1984.

Arnold Zellner and Aloysius Siow. Posterior odds ratios for selected regression hypotheses. *Trabajos de estadística y de investigación operativa*, 31(1):585–603, 1980.

Shu-Qin Zhang, Wai-Ki Ching, Nam-Kiu Tsing, Ho-Yin Leung, and Dianjing Guo. A new multiple regression approach for the construction of genetic regulatory networks. *Artificial intelligence in medicine*, 48(2):153–160, 2010.

Zoey Yi Zhao, Meng Xie, and Mike West. Dynamic dependence networks: Financial time series forecasting and portfolio decisions. *Applied Stochastic Models in Business and Industry*, 32(3):311–332, 2016.

Hua Zhou, Mary E Sehl, Janet S Sinsheimer, and Kenneth Lange. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, 26(19):2375–2382, 2010.

Xiaocong Zhou, Jouchi Nakajima, and Mike West. Bayesian forecasting and portfolio decisions using dynamic dependent sparse factor models. *International Journal of Forecasting*, 30(4):963–980, 2014.

Corwin M Zigler, Krista Watts, Robert W Yeh, Yun Wang, Brent A Coull, and Francesca Dominici. Model feedback in bayesian propensity score estimation. *Biometrics*, 69(1):263–273, 2013.

Corwin Matthew Zigler and Francesca Dominici. Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, 109 (505):95–107, 2014.

Vita

David Puelz was born in Athens, Georgia on February 19, 1989. He graduated from Wesleyan University in May 2011 with a BA in math and physics. After graduating from Wesleyan, he worked at Goldman Sachs in New York City for their investment management division. This sparked his interest in statistics and finance, and he enrolled in a statistics PhD program at the University of Texas at Austin. He graduated with an MS in 2015 and PhD in 2018.