Final Thesis Proposal --- Cover Sheet

Proposal Title: Economic Feature Selection and Return Forecasting: A Bayesian MCMC Approach for Monthly Returns on the S&P 500

Author Name: Omer Sen
Student ID: 13561693
Specialization Track: Quantitative Finance
Preferred Supervisor (optional):

Brief Summary of Feedback Received during Thesis Seminar:*

Feedback received during Thesis Seminar:

Research Question and Focus: Clarify and integrate the dual aspects of forecasting and feature selection.

Time Period and Data Context: Specify the period and address market condition variability.

Literature Review: Broaden the review to include BMCMC theoretical foundations and justify factor choices.

Methodology: Reorganize the methods section for clarity and include clear benchmarks and performance metrics.

Computational/Data Issues: Explicitly discuss convergence, computational strategies, and data quality concerns.

Explanation of how Final Proposal Incorporates Feedback:*

Adjusted my research question to integrate the dual aspect of my thesis clearly

Specified the time period that will be used for my research. As well as a clear description of what data and which factors I intend to include and why.

Completely reorganized my methodology to be much clearer. Explaining each formula and the intuition behind my methodology.

Addressed the concerns about computational issues and data quality problems. Completely scrapped off the idea to use alternative data such as weather, sentiment, social media. Also, specified to use of efficient algorithms to combat computational issues.

Broadened the review of BMCMC theoretical foundations in my literature review and justified factor choices using empirical research.

*It is not necessary to list all feedback received or changes made. Answer each item in one paragraph, or using a list that includes 4-6 bullet points. If you chose not to incorporate most of the feedback into your final proposal, then briefly explain why it was not helpful.

1. **Economic Feature Selection and Return Forecasting: A Bayesian MCMC Approach for Monthly Returns on the S&P 500**

Research Question: Can a hierarchical Bayesian MCMC framework that integrates automatic feature selection (via spike-and-slab priors) with return forecasting improve out-of-sample predictions of monthly dividend-adjusted returns for S&P 500 stocks—compared to traditional machine learning models—and identify the most significant stock-level predictors?

Key Point (Motivating Feature Selection):

My research does not treat forecasting accuracy and feature selection as separate problems. Instead, the Bayesian MCMC framework I plan to use integrates both aspects naturally. By employing spike-and-slab priors within the model, I can automatically determine which features have a meaningful impact on monthly return forecasts. This approach not only potentially improves predictive performance over traditional interpretable ML models but also provides valuable economic insights into which factors drive returns. Essentially, the process of feature selection is a core part of how the Bayesian model learns and makes forecasts, and better feature selection leads to improved forecasting performance. Therefore, evaluating feature importance and assessing forecasting accuracy are two complementary dimensions of the same overall research question. Good forecasting performance in this framework comes from correctly identifying and weighting the significant predictors.

Under-Explored Angle:

Although many studies focus on ML techniques for stock return forecasting, few have employed a Bayesian MCMC framework that simultaneously enhances forecast accuracy and performs interpretable feature selection. This thesis evaluates whether a Bayesian approach

(using spike-and-slab priors to automatically rank predictors) can improve forecasts of monthly returns relative to interpretable ML models. In addition, by referencing the factor selection in the work of Jensen, Kelly, and Pedersen (2023), which distills 153 stock-level characteristics into 13 economically interpretable themes (such as Accruals, Debt Issuance, Investment, Low Leverage, Low Risk, Momentum, Profit Growth, Profitability, Quality, Seasonality, Size, Short-Term Reversal, and Value), the study provides a robust empirical basis for selecting factors. The resulting credible intervals and posterior inclusion probabilities offer rich uncertainty estimates and economic insights that bridge advanced statistical modeling and practical asset pricing.

Related Literature

Recent research has increasingly applied machine learning (ML) techniques to asset pricing, revealing substantial improvements over traditional regression models. For example, Gu, Kelly, and Xiu (2020) demonstrate that tree-based models and neural networks capture nonlinear predictor interactions that standard methods often miss, highlighting the importance of flexible modeling approaches. Giglio, Kelly, and Xiu (2021) further emphasize the role of feature selection in high-dimensional settings, while Kelly and Xiu (2023) stress the necessity of incorporating economic principles to prevent overfitting and enhance interpretability.

Complementing these ML advances, Bayesian MCMC (BMCMC) methods have emerged as a powerful framework for statistical inference in finance. Foundational texts such as Koop (2003) introduce the basic tenets of Bayesian econometrics, where prior beliefs are combined with observed data through Bayes' theorem to form the full posterior distribution of model parameters. This approach allows researchers to not only obtain point estimates but also quantify uncertainty via credible intervals—a key advantage in financial applications where data are often noisy and models are high-dimensional. Moreover, works like Greyserman, Jones, and Strawderman (2006) illustrate how hierarchical Bayesian models enable partial pooling across related predictors, thereby improving the stability of estimates when dealing with many correlated factors. Takaishi (2013) further shows that efficient sampling algorithms, such as Hamiltonian Monte Carlo (HMC) and the No-U-Turn Sampler (NUTS), can greatly enhance computational efficiency by reducing autocorrelation in the posterior samples.

In the context of factor selection, Jensen, Kelly, and Pedersen (2023) provide an important empirical contribution by distilling 153 stock-level characteristics into 13 economically interpretable themes (e.g., Accruals, Debt Issuance, Investment, Low Leverage, Low Risk, Momentum, Profit Growth, Profitability, Quality, Seasonality, Size, Short-Term Reversal, and Value). This clustering not only mitigates the "factor zoo" problem but also aligns with the economic theory that similar predictors should have related effects. Integrating BMCMC methods with this factor taxonomy leverages the strengths of Bayesian inference—such as uncertainty quantification and automatic variable selection via spike-and-slab priors—while remaining firmly grounded in economic theory.

This thesis contributes to the growing body of research by using Bayesian MCMC methods with integrated feature selection strategies to improve return forecasting. By benchmarking Bayesian methods against traditional ML models, this study aims to assess both predictive accuracy and the economic significance of selected features.

2. Data Collection & Cleaning

Time Period: In-sample data will cover monthly returns (including dividends) from 1965 to 2020, while 2020–2024 will be used for out-of-sample evaluation

S&P 500 Participants:

Monthly Price Data: Collect historical monthly prices (including dividends) for the S&P 500 from CRSP with CRSP codes 10-12

Returns Calculation: Compute monthly log returns adjusted for dividends; ensure consistent ticker coverage over the sample period.

Factor Data: Monthly stock-level characteristics are available on WRDS. Representative factors for the 13 themes (Accruals, Debt Issuance, Investment, Low Leverage, Low Risk, Momentum, Profit Growth, Profitability, Quality, Seasonality, Size, Short-Term Reversal, Value) will be selected based on Jensen, Kelly, and Pedersen (2023).

Macro Indicators:

I**nterest Rates:**

Monthly or weekly yields on instruments like the 10-year Treasury note.

Inflation / CPI:

Monthly data, adjusted or lagged appropriately.

ETF Flows:

Sector-level or broad market ETF flows (e.g., SPY, sector SPDRs). To add a behavioural dimension.

Data Cleaning:

Synchronize time stamps (e.g., monthly close prices aligned with macro data release dates). Handle missing values or outliers (e.g., winsorization, or excluding days with incomplete data).

3. Methodology

1. Model Specification: Hierarchical Bayesian MCMC Framework

- **Response Variable:**
  - Monthly returns (including dividends) for individual S&P 500 stocks (CRSP codes 10–12).
- **Predictors:**
  - A set of macroeconomic and market variables, along with stock-level factors. For the stock-level factors, representative predictors will be chosen based on the 153 characteristics distilled into 13 economic themes (e.g., Accruals, Debt Issuance, Investment, Low Leverage, Low Risk, Momentum, Profit Growth, Profitability, Quality, Seasonality, Size, Short-Term Reversal, and Value) as detailed in Jensen, Kelly, and Pedersen (2023). Data will be obtained from WRDS.
- **Likelihood and Regression Model:**
  - The likelihood is a function that gives the probability of observing our monthly return data given the model parameters. In this thesis, we assume the data follows a standard linear regression model:
  - $r_t = X_t\beta + \epsilon_t,\ \epsilon_t \sim N(0, \sigma^2)$
  - Where Xt is the matrix of predictors, β is the vector of coefficients, and ϵt is the normally distributed error. The likelihood is then the product of the probability densities of the observed returns given these parameters.
- **Prior Distributions:**

- Hierarchical Structure**:** Coefficients belonging to the same economic theme are assumed to come from a common distribution. For example, for predictors within a theme:
- $\beta_i \sim N(\mu_{theme}, \tau^2)$,
- and the group mean μ_{theme} has its own prior, such as $\mu_{theme} \sim N(0, \sigma_\mu^2)$

- **Spike-and-Slab Priors:**
  - These are added at the individual predictor level to "turn off" predictors that do not significantly affect returns. The spike (concentrated at zero) represents the belief that many predictors have no effect, while the slab (a wider distribution) allows for significant deviations from zero.
- **From Prior to Posterior:**
  - Using Bayes' theorem, the posterior distribution combines our prior beliefs with the information from the data (via the likelihood):
  - $p(\beta, \sigma^2 \mid r, X) \propto p(r \mid \beta, \sigma^2, X) p(\beta, \sigma^2)$
  - Intuitively, this means we update our prior beliefs about the coefficients based on how likely the observed returns are given those coefficients. The resulting posterior distribution reflects both our initial assumptions and the evidence provided by the data.
- **Sampling Method:**
  - Efficient MCMC algorithms, such as Hamiltonian Monte Carlo (HMC) or the No-U-Turn Sampler (NUTS), will be used to sample from this posterior distribution.

2. Feature Selection and Interpretability

- **Bayesian Variable Selection:**
  - By incorporating spike-and-slab priors within the hierarchical model, the framework will automatically identify significant predictors. This is done by estimating the posterior inclusion probability for each feature, which tells us how likely it is that a predictor is important.

3.Baseline ML Models (Non-Black-Box)

- **Interpretable Models:**
  - Baseline models will include linear models with LASSO regularization and tree-based models with clear feature importance measures (e.g., using SHAP values).

- **Training & Validation:**
    - A rolling or walk-forward scheme will be used to preserve temporal order and prevent look-ahead bias.
- **Evaluation:**
    - Model performance will be evaluated using out-of-sample predictive metrics, such as mean squared error (MSE) and R^2.

4.Model Evaluation

- **Predictive Performance:**
    - In-sample data will cover 1965–2020, and out-of-sample performance will be assessed using data from 2020–2024.

5. Interpretation of Posterior Distributions

- **Credible Intervals:**
    - Credible intervals will be computed for each predictor's coefficient to assess the strength and direction of its impact on returns.
- **Feature Importance:**
    - Posterior inclusion probabilities from the spike-and-slab priors will be used to rank predictors, with comparisons drawn to importance measures from the baseline ML models.
- **Economic Discussion:**
    - The findings will be interpreted in economic terms, linking significant predictors to underlying market behaviors and investor sentiment.

Clarification of the Approach:

- Bayesian MCMC is a framework that can be applied to any model to estimate the full posterior distribution of its parameters. In my thesis, I will first construct traditional, interpretable ML models (such as LASSO or decision trees) and evaluate their out-of-sample predictive performance using metrics like MSE and R^2. Then, I will implement a hierarchical Bayesian MCMC framework to estimate the regression model coefficients for forecasting monthly returns. This framework uses priors—set by initial beliefs about the coefficients—and updates them with the likelihood (i.e., the probability of observing the data given these coefficients) to produce the posterior distribution. The inclusion of spike-and-slab priors within this hierarchical model will automatically select significant features. In essence, I am comparing a traditional ML approach with a BMCMC-enhanced model to determine if the Bayesian method improves forecasting accuracy and enhances economic interpretability.

References

Koop, G. (2003). Bayesian Econometrics. Wiley
https://scholars.cityu.edu.hk/en/theses/deep-learning-and-bayesian-model-in-empirical-asset-pricing(b21acdc3-eed7-4fe3-92e8-d3eca8da1cec).html

Bagnara, M. (2024). Asset Pricing and Machine Learning: A critical review. Journal of Economic Surveys, 38, 27–56. https://doi.org/10.1111/joes.12532

Haejung Na, Soonho Kim, Predicting stock prices based on informed traders' activities using deep neural networks, Economics Letters, Volume 204, 2021, 109917

Gu, Shihao, Bryan Kelly, and Dacheng Xiu. "Empirical asset pricing via machine learning." *The Review of Financial Studies* 33.5 (2020): 2223-2273.

Giglio, Stefano, Bryan Kelly, and Dacheng Xiu. "Factor models, machine learning, and asset pricing." *Annual Review of Financial Economics* 14.1 (2022): 337-368.

Kelly, Bryan, and Dacheng Xiu. "Financial machine learning." *Foundations and Trends® in Finance* 13.3-4 (2023): 205-363.

Greyserman, A., Jones, D. H., & Strawderman, W. E. (2006). Portfolio selection using hierarchical Bayesian analysis and MCMC methods. *Journal of Banking and Finance*, *30*(2), 669-678. https://doi.org/10.1016/j.jbankfin.2005.04.008

Jensen, Theis Ingerslev and Kelly, Bryan T. and Pedersen, Lasse Heje, Is There a Replication Crisis in Finance? (January 30, 2021). NYU Stern School of Business Forthcoming, Available at
SSRN: https://ssrn.com/abstract=3774514 or http://dx.doi.org/10.2139/ssrn.3774514

Takaishi, T. (2013). Markov Chain Monte Carlo versus Importance Sampling in Bayesian Inference of the GARCH Model. Procedia Computer Science, 22, 1056–1064. https://doi.org/10.1016/j.procs.2013.09.191