Bayesian Regression for Financial Forecasting

Daily and Intraday Applications to S&P 500 Returns

CN: 39270, ST451

May 7, 2025

Abstract

We investigate the use of Bayesian regression models to forecast short-term financial returns using both daily and intraday S&P 500 data. A rolling-window Bayesian linear model with shrinkage priors and Student-t likelihood is implemented using PyMC. We evaluate predictive accuracy and use the posterior distribution to guide trading decisions. The model is benchmarked against classical baselines (OLS, Ridge, Lasso). We find that predictive accuracy is only slighly higher than chance, for daily data, and performs suboptimally for intra-day data. However, the research also finds that the Bayesian model does quantify uncertainty well.

Contents

1 Introduction

Problem Statement

Forecasting financial asset returns is a longstanding challenge in quantitative finance and portfolio management. While markets are widely believed to be efficient in the short term, small but exploitable predictive signals may exist, particularly when uncertainty is properly quantified. Other common approaches to financial forecasting or trading algorithms include Autoregressive-Moving Average (ARMA), Generalised Autoregressive Conditional Heteroskedasticity (GARCH), or potentially machine learning methods such as reinforcement learning. This research attempts to use Bayesian regression techniques within machine learning to develop a trading algorithm for longer term and high-frequency trading.

Motivation and Relevance

This project investigates the use of Bayesian linear regression models to forecast short-term returns on the S&P 500 index, using both daily and intraday data. The Bayesian framework allows us to incorporate prior beliefs, apply shrinkage through regularising priors, and obtain full posterior distributions rather than just point forecasts. This enables uncertainty-aware decision-making, which is particularly relevant in financial applications where risk control is as important as predictive accuracy.

We implement a rolling-window Bayesian regression model with Laplace priors on coefficients and a Student-t likelihood to account for heavy-tailed return distributions. Inference is performed via automatic differentiation variational inference (ADVI) using PyMC. The predictive distribution is then used to make directional trading decisions under simple threshold-based strategies. To benchmark performance, we compare against classical linear models such as ordinary least squares (OLS), Ridge regression, and Lasso. In the financial simulations we assume no transactional costs, for the sake of simplicity.

Structure of the Report

The remainder of the report is organised as follows. Section 2 describes the datasets and feature construction. Section 3 presents the Bayesian modelling framework and inference procedure. Section 4 outlines the baseline models. Section 5 reports empirical results, including forecasting accuracy and trading performance. Section 6 concludes.

2 Data Description

2.1 Daily SPY Data

We source daily SPY (S&P 500 ETF) prices and volumes from Yahoo Finance via the `yfinance` API. Our main dataset spans January 1, 2014 through December 31, 2024, which gives approximately 2,700 trading days. For each trading day t we record:

- **Close price** P_t and **Volume** V_t
- **Log return** $r_t = \ln(P_t/P_{t-1})$
- **Lagged returns** $r_{t-1}, \ldots, r_{t-5}$
- **Simple moving averages** SMA_5 and SMA_{10} (computed on price and shifted by one day)
- **5-day volatility** $\sigma_5 = \text{std}(r_{t-4:t})$
- **20-day volume z-score** $z_t = (V_t - \overline{V}_{20})/\text{sd}(V_{20})$, shifted by one day

After computing these features we drop any rows with missing values from rolling or lagged operations, leaving $\approx 2\,700$ complete observations. The volume z-score is simply to standardise the trading volume relative to recent history.

2.2 Intraday Sample

To assess high-frequency robustness, we draw one-minute SPY data over the most recent 30 calendar days, then randomly select ten trading days for intraday backtests. For each selected day we use minute-level:

- **Price** and **Volume** between 09:30–16:00 ET
- **Minute return** $\ln(P_t/P_{t-1})$
- **Lag-1 return**, 5-minute rolling mean/volatility
- **5- and 10-minute SMAs** of price
- **20-minute volume z-score**

We then apply a 60-minute rolling window for feature computation and Bayesian forecasting.

2.3 Summary Statistics

We start the analysis by calculating the styalised statistics for the returns data.

	Mean return	SD return
Daily (2014–2024)	0.000493	0.010864

Table 1: Empirical mean and standard deviation of daily SPY log-returns.

As is standard for finanical returns we see a mean return close to zero, in this case the returns are positive which is expected for a 10 year period of SPY data where overall we expect growth.

Test	Statistic	p-value
Shapiro–Wilk (normality)	0.8841	< 0.001
Jarque–Bera (normality)	22513.38	< 0.001
Ljung–Box on squared returns (lag 10)	3637.85	< 0.001

Table 2: Normality and volatility-clustering diagnostics for daily SPY returns.

Next we investigate the normality conditions with the Shapiro-Wilk test and Jarque-Bera, then the Ljung-Box test to analyse volatility clustering.

These results show the typical findings for financial data, returns do not appear to be normally distributed. The results from the Ljung-Box test do show the prescence of volatility clustering over the period of two trading weeks.

The autocorrelation function (ACF) plot shows that there is no significant prescence of day-to-day correlation between the returns 5.

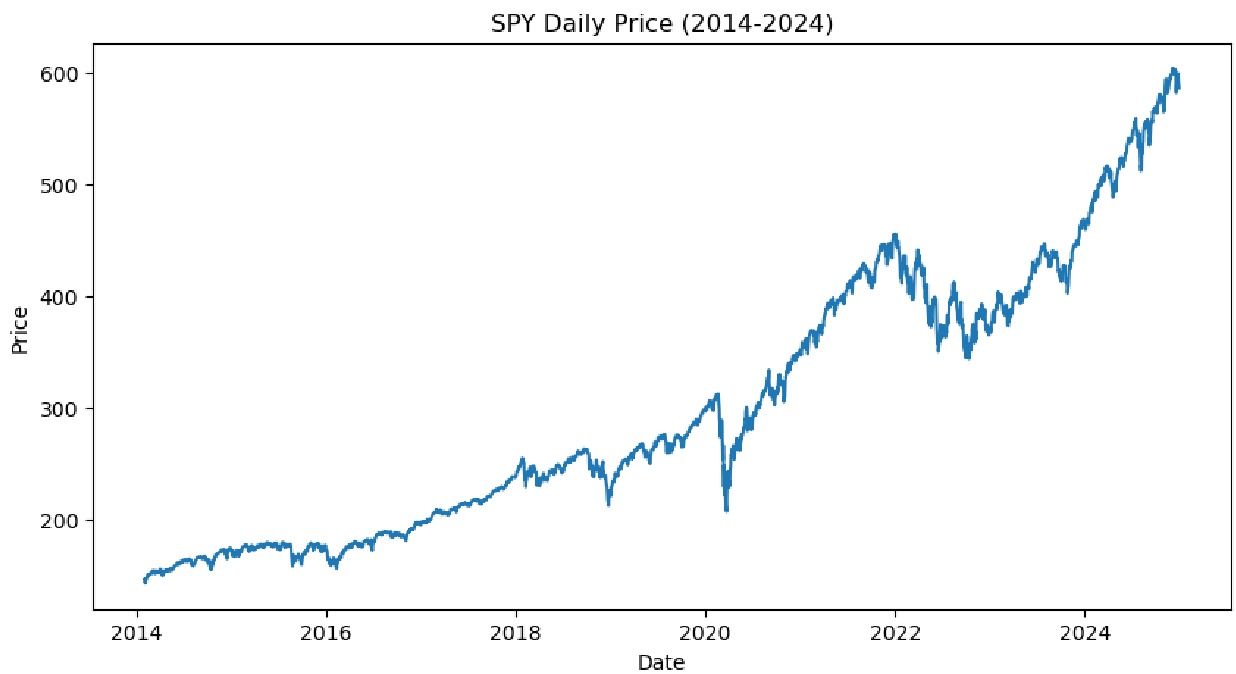

Figure 1: SPY daily closing price (2014–2024).

From this price plot we can see an overall growth in the value of the S&P-500. Also visible are the crashes from COVID-19 in early 2020 and the interest rate/high inflation crisis during 2022.

3 Methodology

In this section we describe our Bayesian regression framework, the sequential rolling-window evaluation, and our choices of priors and inference algorithm.

3.1 Bayesian Linear Regression Model

Let $\mathbf{x}_t \in \mathbb{R}^p$ denote the p-dimensional feature vector at time t (five lagged log-returns, two moving averages, rolling volatility, volume z-score), and let y_t be the log return on the next trading day. We use a linear model with Student-t observation noise:

$$y_t \mid \mathbf{x}_t, \beta, \sigma, \nu \sim \text{Student-}t\big(\nu,\ \mu_t = \mathbf{x}_t^\top \beta,\ \sigma\big)\,.$$

Here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ are the regression coefficients, $\sigma > 0$ is the scale of the noise, and $\nu > 2$ is the degrees-of-freedom which impacts tail thickness. Student-t has the advantage over the Normal distribution for modelling financial returns due to their heavy tailed nature.

3.2 Priors on Parameters

To impose shrinkage on coefficients and guard against over-fitting in small samples, we choose independent Laplace (double-exponential) priors on each β_j,

$$\beta_j \ \sim\ \text{Laplace}\big(\mu = 0,\ b\big), \quad j = 1, \ldots, p,$$

where $b > 0$ controls the strength of ℓ_1 shrinkage. For the noise parameters we now use weakly-informative priors that encourage moderate dispersion but allow heavy tails if supported by the data:

$$\sigma \ \sim\ \text{HalfNormal}\big(\sigma = 1\big), \qquad \nu \ \sim\ \text{Gamma}\big(\alpha = 2,\ \beta = 0.1\big),$$

where the Half-Normal prior on σ centers the scale around unity (on the standardised y–scale) and the Gamma prior on ν places negligible mass at $\nu \to 0$ while allowing for moderately heavy-tailed Student-t errors.

The use of a flexible Student-t likelihood with these priors is based off the STAN documentation[1] and a paper titled Model-Based Clustering of Non-Gaussian Panel Data Based on Skew-t Distributions [2].

3.3 Hyperparameter Selection

Laplace prior scale b. To choose the shrinkage strength b in our Laplace prior $\beta_j \sim \text{Laplace}(0, b)$, we performed a fast one-step-ahead cross-validation over 20 equally spaced test time points. For each candidate

$$b \in \{0.001,\ 0.01,\ 0.1,\ 0.15,\ 0.2,\ 0.5,\ 1.0\},$$

we refit the Bayesian model on the prior 252 trading days, forecast the next day, and computed the squared prediction error. Averaging these 20 errors gave a CV MSE, and we selected

$$b = 0.001 \quad (\text{CV MSE} = 1.8245 \times 10^{-4}).$$

See Table 5 in the Appendix for the full results.

Baseline model penalties. For our Ridge and Lasso benchmarks we instead used a proper time-series cross-validation (no shuffling) via `TimeSeriesSplit(n_splits=5)`. We scanned

$$\alpha_{\text{Ridge}} \in \{0.001,\, 0.01,\, 0.1,\, 1.0,\, 10.0\}, \quad \alpha_{\text{Lasso}} \in \{10^{-4},\, 10^{-3},\, 10^{-2},\, 10^{-1}\},$$

refitting on each growing train-set and evaluating RMSE on the hold-out block. We then picked the α minimising the average CV RMSE. These selected values were used in all downstream baseline forecasts.

3.4 Sequential Rolling-Window Inference

We evaluate predictive performance in a realistic "walk-forward" scheme. Let T be the total number of observations. For each $t = W, W+1, \ldots, T-1$, where $W = 252$ (one trading year), we:

1. Extract the training set $\{(\mathbf{x}_i, y_i)\}_{i=t-W}^{t-1}$.
2. Standardise each feature by its training-set mean and standard deviation.
3. Build the Bayesian model on the scaled training data.
4. Fit the variational approximation to the posterior via ADVI (10,000 iterations).
5. Draw $N = 500$ posterior samples of $(\boldsymbol{\beta}, \sigma, \nu)$.
6. Forecast y_t by passing $\mathbf{x}_t$ (scaled by the same training transform) through the posterior predictive Student-t.
7. Store the predictive mean and standard deviation for evaluation.

By refitting each day on only past data and standardising within each window, we avoid any look-ahead leakage. The result is a sequence of fully Bayesian one-step-ahead forecasts that incorporate parameter uncertainty and heavy-tailed noise.

3.5 Variational Inference

Because our model uses a heavy-tailed Student-t likelihood and shrinkage (Laplace) priors, we can't solve for the exact posterior in closed form. Instead, we lean on PyMC's Automatic Differentiation Variational Inference (ADVI)routine to quickly and efficiently approximate it. ADVI approximates the joint posterior $p(\boldsymbol{\beta}, \sigma, \nu \mid \mathcal{D})$ with a mean-field normal variational family,

$$q(\boldsymbol{\theta}) \;=\; \prod_j \mathcal{N}(\theta_j \mid \mu_j, \rho_j^2), \quad \boldsymbol{\theta} = (\beta_1, \ldots, \beta_p,\ \log\sigma,\ \log\nu),$$

by maximising the evidence lower bound (ELBO). We run ADVI for 10,000 iterations per window, monitoring the ELBO trace to ensure convergence, and then draw 500 samples from $q(\boldsymbol{\theta})$ for posterior predictive checks.

3.6 ELBO Convergence Diagnostics

During each ADVI fit we record the evidence lower bound (ELBO) at every iteration, which acts as a surrogate for the true log-marginal likelihood. A well-behaved VI run will show a monotonically increasing ELBO (or equivalently decreasing −ELBO loss) that plateaus before the chosen iteration limit. We perform 10,000 ADVI steps per window and plot the final ELBO trace (Figure 2) to verify that the approximation has stabilised. This diagnostic ensures that our 500 posterior draws are sampled from a variational distribution that has in fact converged.

3.7 Intraday Trading

To test how the Bayesian model performs at a finer timescale, we conduct a separate intraday experiment on ten randomly selected trading days within the most recent 30-day window. For each chosen day we:

1. Download one-minute SPY prices from 09:30 to 16:00.
2. Compute minute-by-minute returns and the same set of features (lag-1 return, 5-minute rolling average and volatility, 5- and 10-minute SMA, volume z-score).
3. Use a 60-minute rolling window to fit the Bayesian model via ADVI (10,000 iterations, 200 posterior draws) and forecast the next minute's return.
4. Convert the predictive mean into a long/short position and measure P&L, and number of trades for that day.

This intraday trading is to see how the model performs in a high frequency trading scenario. A key weakness in this is that any asset price movements are effectively a random-walk/noise. We expect the noise, and the limited sampling, to be too significant for the model to exploit any signal and make a consistent profit.

4 Baseline Models

In order to gauge the incremental value of our Bayesian regression, we first establish three classical linear benchmarks. All models are refit each day on a rolling window of the previous 252 trading days and used to forecast the next-day return. Our chosen models differ in how, or whether, they regularise the coefficient estimates and balance bias–variance trade-offs.

4.1 Ordinary Least Squares (OLS)

$$\hat{\beta}_{\text{OLS}} = \arg\min_{\beta} \|y - X\beta\|^2 \tag{1}$$

Ordinary Least Squares fits a linear relationship by minimising the sum of squared errors between the predicted and actual returns. It imposes no explicit constraint on coefficient size, which makes it susceptible to over-fitting when predictors are noisy or highly correlated. Nonetheless, OLS remains the standard unbiased estimator in the absence of regularisation, and serves as a *no-frills* baseline.

4.2 Ridge Regression

$$\hat{\beta}_{\text{Ridge}} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|^2 \tag{2}$$

Ridge Regression augments OLS with an ℓ_2 penalty on the coefficient vector, effectively shrinking all coefficients toward zero. The penalty weight, α, is selected by time-series cross-validation to optimise out-of-sample mean-squared error. Ridge is particularly useful when many small signals are spread across features or when multicollinearity inflates variance in the OLS estimates.

4.3 Lasso Regression

$$\hat{\beta}_{\text{Lasso}} = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|_1 \tag{3}$$

Lasso applies an ℓ_1 penalty, which can drive some coefficients exactly to zero and perform implicit variable selection. This sparsity enhances interpretability by isolating the most predictive features and guards against over-fitting when certain predictors carry negligible signal. Like Ridge, the regularisation parameter is chosen via rolling cross-validation.

4.4 Comparison of Bayesian and Frequentist Linear Models

Although both the Bayesian and the classical (frequentist) benchmarks share the same linear predictor

$$\mu_t = \mathbf{x}_t^{\top}\boldsymbol{\beta},$$

their estimation methods differ:

- **Frequentist (OLS, Ridge, Lasso)** Fit $\boldsymbol{\beta}$ by optimising a loss function on the data, least-squares for OLS, plus an ℓ_2 or ℓ_1 penalty for Ridge/Lasso. These penalties can be seen as MAP estimates under Gaussian or Laplace priors, but the frequentist stops at the point estimate and does not provide a distribution over β.

- **Bayesian** Places explicit priors on $\boldsymbol{\beta}, \sigma, \nu$ and infers the full posterior $p(\boldsymbol{\beta}, \sigma, \nu \mid \mathcal{D})$ via variational inference. This yields not only a central estimate (the posterior mean) but also an uncertainty quantification (posterior variance) for every parameter and every one-step-ahead forecast.

In practice, with weakly-informative priors the Bayesian posterior mean often closely matches the penalised-regression point estimates, but the Bayesian approach adds the benefit of more robust uncertainty estimates.

Buy-and-Hold Benchmark In our trading performance comparison (Section 5) we also include a simple "buy-and-hold" strategy that is always holding a long position in SPY to show how an entirely passive investment would have performed over the same period. This passive baseline contextualises any gains from forecasting, not just acting as a comparison against the asset price, simply holding S&P-500 is a standard investment due to its relatively consistent long-term growth.

5 Empirical Results and Discussion

5.1 Daily-frequency Performance

Model	RMSE	MAE	Dir. Acc. (%)	Sharpe
OLS	0.0118	0.0078	48.56	−0.30
Ridge	0.0116	0.0077	49.00	−0.10
Lasso	0.0112	0.0073	54.01	0.21
Bayesian	0.0112	0.0073	53.53	0.14

Table 3: One-step-ahead forecast performance and annualised strategy Sharpe for all models.

Figure 3 (in Appendix A) shows the cumulative wealth paths for the Bayesian strategy and the market benchmark (buy-and-hold), both starting at 1.

Among the four approaches, Lasso and the Bayesian model achieve the lowest RMSE (0.0112), whereas Lasso shows the highest directional accuracy (54.01%), with a positive annualised Sharpe of 0.21. The Bayesian model is a close second with MAE = 0.0073 and directional hit rate (53.53%), producing a Sharpe of 0.14. In contrast, unregularised OLS and Ridge deliver poorer forecast accuracy and negative threshold-trading Sharpe ratios, indicating that simple least-squares fits are prone to overfitting in this noisy daily setting.

Notably, all four strategies underperform a passive buy-and-hold benchmark, which achieves an annualised Sharpe of 0.71. This gap underscores two key points: first, regularisation (via ℓ_1 or Bayesian shrinkage) yields tangible improvements in short-horizon forecasts compared to OLS; second, however, the residual noise in daily returns and the crude fixed-threshold trading rule limit any meaningful outperformance. In practice, even small transaction costs would likely erase the slim gains observed here.

5.2 Intraday Performance

Date	Sharpe	Accuracy	P&L	Trades
2025-04-14	0.0816	0.5224	0.00403	10
2025-04-24	0.0205	0.4531	0.00064	9
2025-04-11	0.1261	0.5538	0.00937	10
2025-04-16	0.0241	0.4925	0.00083	5
2025-04-25	−0.0873	0.5303	−0.00255	11
2025-04-22	−0.0133	0.5000	−0.00050	5
2025-04-21	0.2212	0.5942	0.00659	1
2025-04-15	−0.1014	0.4091	−0.00442	12
2025-04-17	−0.1378	0.3913	−0.00756	4
2025-04-28	0.1694	0.5294	0.00561	3
Average	0.030314	0.4976	0.00120	7

Table 4: Intraday backtest results over 10 sampled trading days: per-day Sharpe ratio, directional accuracy, P&L, and number of trades, with average accuracy and P&L.

Performance is highly variable day-to-day. While 2025-04-21 achieves a Sharpe above 0.2, several sessions (e.g. 2025-04-17) register negative risk-adjusted returns. This dispersion indicates that minute-level noise dominates any predictive signal from our simple features. In a realistic setting with transaction costs and slippage, such a strategy would likely struggle to break even.

Despite an average directional accuracy close to coin-flip (49.8%), the strategy still generates a small positive P&L of 0.12% per day on average. This suggests that even marginally correct signals, when weighted by position size, can produce modest gains in aggregate. However, the narrow gap above 50% accuracy, combined with low per-trade edge, means that any realistic transaction costs or slippage would likely eliminate these profits.

6 Conclusion

We have implemented a rolling-window Bayesian linear regression with Laplace priors (scale $b = 0.001$) and a Student-t likelihood, fit via 10 000-step ADVI with diagnostic ELBO traces (Figure 2). On daily SPY returns (2014–2024), the Bayesian strategy achieved an RMSE of 0.0112, MAE of 0.0073, 53.5% directional accuracy and an annualised Sharpe of 0.14. Classical benchmarks showed that Lasso (with $\alpha = 0.01$) slightly outperforms Bayesian in Sharpe (0.21), while OLS and Ridge delivered poorer forecasts and negative Sharpe. All four methods underperformed passive buy-and-hold (Sharpe 0.71), underscoring the dominance of noise in daily returns and the limits of fixed-threshold trading. Aditionally the chosen dataset exhibits high growth so holding the asset from the start may have always been the most optimal strategy.

In our intraday robustness check ten one-minute sessions sampled from the most recent 30 days, the Bayesian model averaged 49.8% accuracy, 0.12% daily P&L, 0.03 Sharpe, and 7 trades per day. Performance varied widely (Sharpe from –0.14 to 0.22), indicating that minute-level noise quickly erodes any signal and that transaction costs would likely eliminate these slim gains.

Overall, while Bayesian shrinkage yields comparable forecast accuracy to the best penalised frequentist methods and provides full predictive distributions, the practical improvement in risk-adjusted returns is modest. This reflects the inherently low signal-to-noise ratio of short-horizon financial data. Moreover, our 2014–2024 sample spans the COVID-19 crash of early 2020 and its rapid rebound, an extreme volatility regime that likely exacerbated forecasting difficulty and hampered any simple threshold-based strategy in capturing transient market dynamics.

Future work could explore more dynamic priors and parameter selection and a more rigourous trading strategy. Additionally instead of choosing just one asset if the model had access to many different assets and was more focused on portfolio management this could return better results.

References

[1] Stan Development Team. Student-t prior choice recommendations. `https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations#prior-for-degrees-of-freedom-in-students-t-distribution`, 2025. Accessed: 2025-05-07.

[2] M. A. Juárez and M. F. J. Steel. Model-based clustering of non-gaussian panel data based on skew-t distributions. *Journal of Business & Economic Statistics*, 28(1):52–66, 2010.

A Supplementary Plots

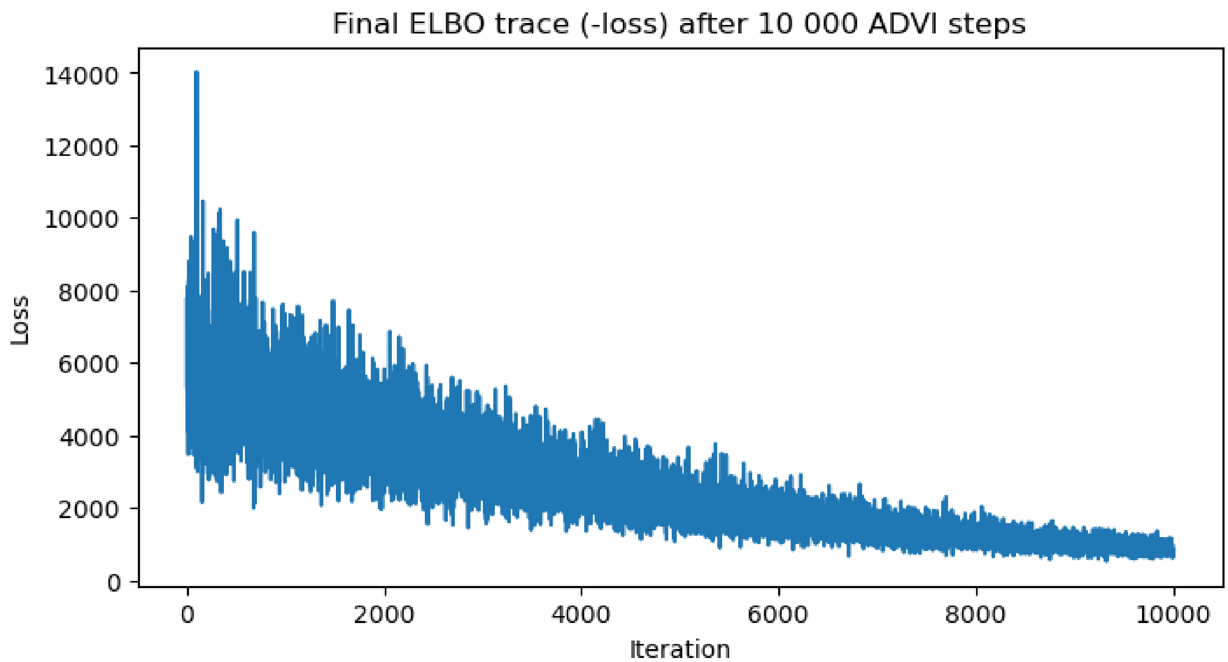

Figure 2: ELBO decay shows that approximation stabilises within the chosen 10,000 ADVI steps

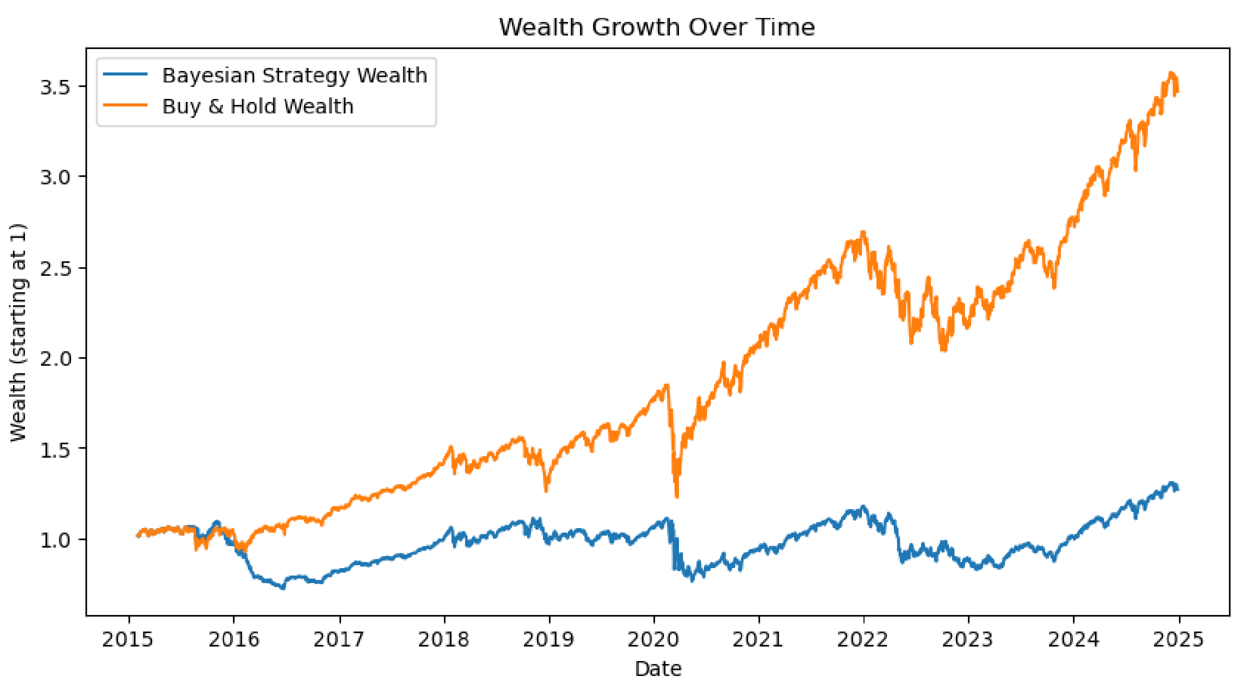

Figure 3: Cumulative wealth growth (starting at 1) for the Bayesian strategy versus Buy & Hold (Market).

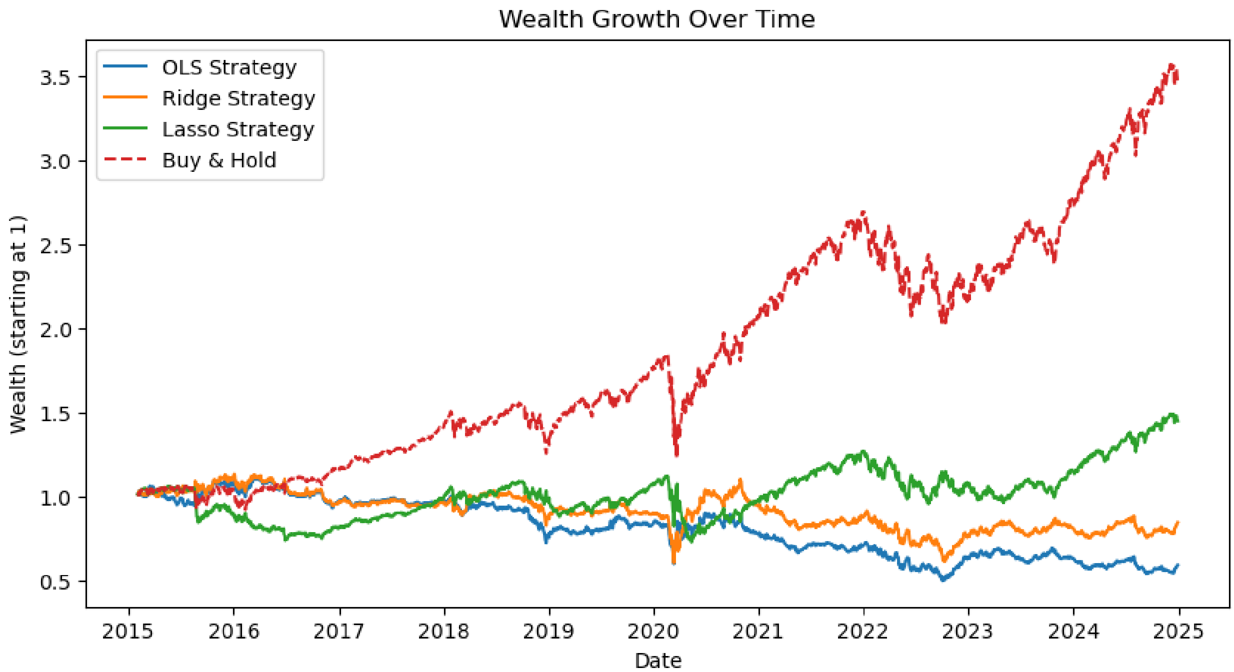

Figure 4: Cumulative wealth growth (starting at 1) for the OLS, Ridge and Lasso strategies versus Buy & Hold.

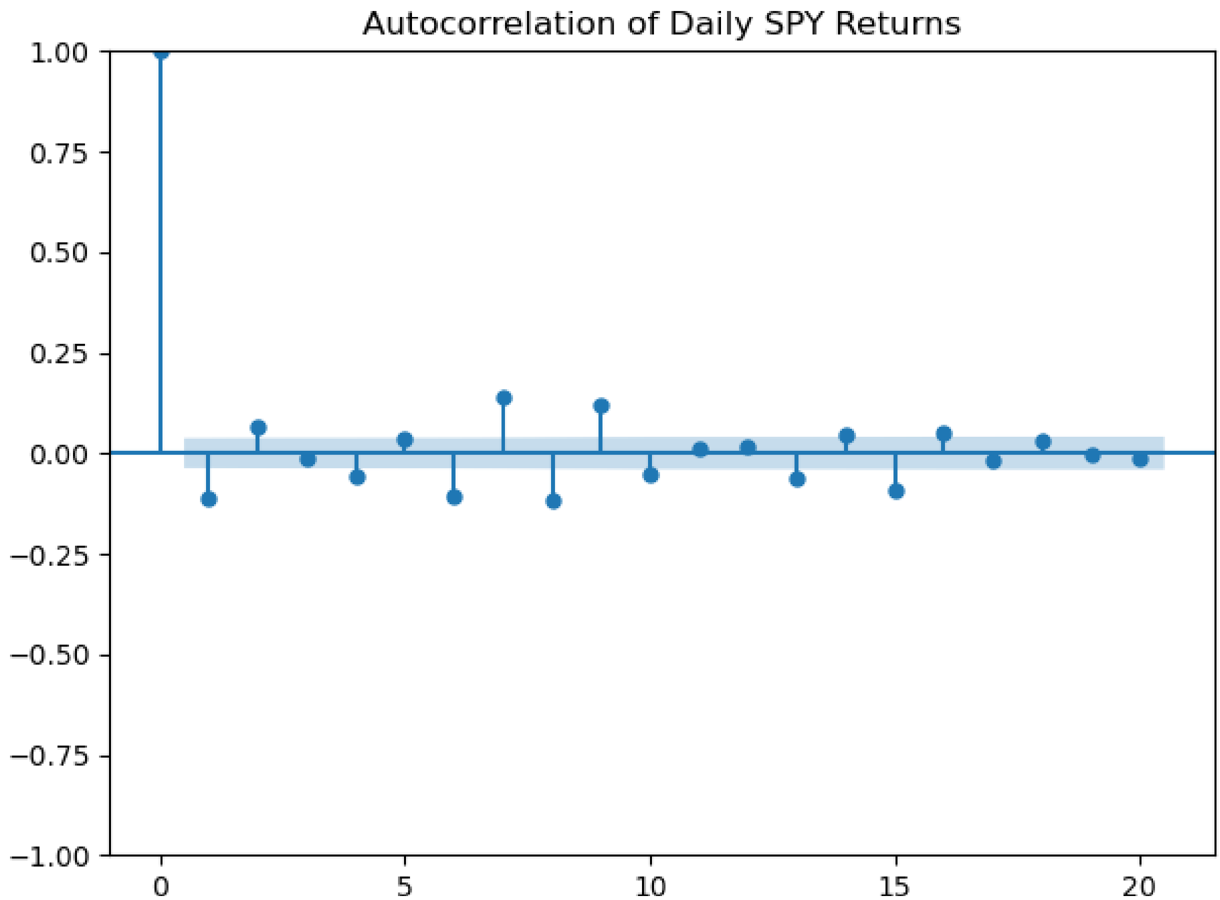

Figure 5: Autocorrelation of daily SPY log-returns (lags 1–20).

B Supplementary Tables

b	CV MSE ($\times 10^{-4}$)
0.001	**1.8245**
0.01	2.1045
0.1	2.0867
0.15	2.1806
0.2	2.1029
0.5	2.0772
1.0	2.1204

Table 5: Cross-validation results for the Laplace prior scale b; the selected value is shown in bold.

Model	α	CV MSE
Ridge Regression		
0.001		$1.099522_{\text{e-02}}$
0.01		$1.099515_{\text{e-02}}$
0.1		$1.099476_{\text{e-02}}$
1.0		$1.099568_{\text{e-02}}$
10.0		$1.099394_{\text{e-02}}$
Selected	**10.0**	$\mathbf{1.099394_{\text{e-02}}}$
Lasso Regression		
0.0001		$1.093164_{\text{e-02}}$
0.001		$1.072067_{\text{e-02}}$
0.01		$1.070511_{\text{e-02}}$
0.1		$1.070511_{\text{e-02}}$
Selected	**0.01**	$\mathbf{1.070511_{\text{e-02}}}$

Table 6: Cross-validation results for the Ridge and Lasso penalty parameters. The selected α (minimising CV MSE) is set in bold.