Witzany, Jiří

Article

A bayesian approach to measurement of backtest overfitting

Risks

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Witzany, Jiří (2021) : A bayesian approach to measurement of backtest overfitting, Risks, ISSN 2227-9091, MDPI, Basel, Vol. 9, Iss. 1, pp. 1-22, https://doi.org/10.3390/risks9010018

This Version is available at:
https://hdl.handle.net/10419/258108

risks

MDPI

Article

A Bayesian Approach to Measurement of Backtest Overfitting

Jiří Witzany

Department of Banking and Insurance, Faculty of Finance and Accounting, University of Economics, W. Churchill Sq. 4, 130 67 Prague, Czech Republic; jiri.witzany@vse.cz; Tel.: +420-224-095-174

Abstract: Quantitative investment strategies are often selected from a broad class of candidate models estimated and tested on historical data. Standard statistical techniques to prevent model overfitting such as out-sample backtesting turn out to be unreliable in situations when the selection is based on results of too many models tested on the holdout sample. There is an ongoing discussion of how to estimate the probability of backtest overfitting and adjust the expected performance indicators such as the Sharpe ratio in order to reflect properly the effect of multiple testing. We propose a consistent Bayesian approach that yields the desired robust estimates on the basis of a Markov chain Monte Carlo (MCMC) simulation. The approach is tested on a class of technical trading strategies where a seemingly profitable strategy can be selected in the naïve approach.

Keywords: multiple testing; backtest overfitting; investment strategy; MCMC

Citation: Witzany, Jiří. 2021. A Bayesian Approach to Measurement of Backtest Overfitting. *Risks* 9: 18. https://doi.org/10.3390/risks9010018

Received: 19 November 2020
Accepted: 1 January 2021
Published: 8 January 2021

1. Introduction

The problem of backtest overfitting appears mainly in two important econometric research areas: testing and selection of factors explaining asset returns (see, e.g., De Prado 2015; Harvey and Liu 2013, 2014, 2015, 2017, 2020) and selection of investment strategies (see, e.g., White 2000; Bailey et al. 2016; López de Prado and Lewis 2019). Our focus is the investment strategy selection problem arising when many strategies are developed and tested on historical data in order to find a "performing" one. The selection process can be realized by an individual researcher or institution or latently by a set of researchers investigating various strategies and publishing only the promising ones (see, e.g., Andrews and Kasy 2019; Chen and Zimmermann 2020). The latter approach is more common for theoretical research, while the former, easier to control, would be typical for a quantitative investment firm.

We formalize and investigate the problem of strategy selection on the basis of a large set of candidates. Consider investment strategies[1] $\mathcal{S}_1, \ldots, \mathcal{S}_K$ that are backtested and evaluated over a historical period with T_1 (e.g., daily) returns $r_{k,t}$, $k = 1, \ldots, K$, $t = 1, \ldots, T_1$. Note that the strategies could have been developed on another preceding training period and backtested or validated on the $\{1, \ldots, T_1\}$ period. Another possibility that we use in the empirical study is that one considers a number of expertly proposed, e.g., technical, strategies that are just evaluated on the backtest period. On the basis of the historical data, we estimate the (annualized) sample means m_k, standard deviations s_k, or Sharpe ratios SR_k, and, given a criterion, we select the "best" strategy $\mathcal{S}_b$. Of course, the key question is what can be realistically expected from the best strategy if implemented in the future period of length T_2 (see Figure 1).

[1] By an investment strategy, we mean a rule that dynamically determines a portfolio of assets that can be long or short according to information available at the beginning of the period t over which the investment is held. The return over the period is net of the cost of borrowings needed to set up the portfolio (i.e., the strategy is self-financing). For example, an investment strategy may just determine at the end of each trading day whether a long or short position is taken in a specific index futures contract for the next day.

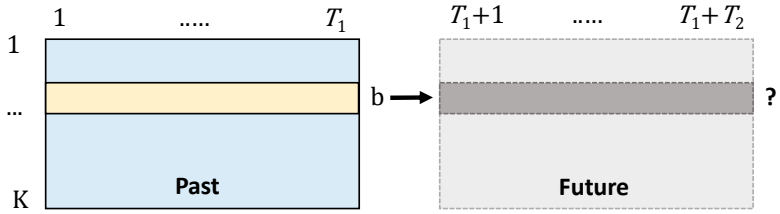

Figure 1. Future performance of the best strategy selected on the basis of past data.

Specifically, the questions usually asked include the following:

- First, is it sufficient to apply the standard single p-test to the best strategy?
- If not, how should we modify the test to incorporate the multiple test effects?
- What is the expected future, i.e., true out-of-sample (OOS) performance (return, SR, etc.) of the best strategy selected on the in-sample (IS), i.e., historical dataset?
- What is the haircut, i.e., the percentage reduction, of the expected OOS performance compared to the IS performance?
- What is the probability of loss, if the strategy is implemented over a future period?
- What is the expected OOS rank of the IS best strategy among the candidate strategies?
- What is the probability that the selected model will in fact underperform most of the candidate models?
- What is the probability that we have selected a false model (false discovery rate—FDR)?

We propose a Bayesian methodology that allows us to simulate many times the IS selection and OOS, i.e., future, realization process (Figure 1) in order to address the questions formulated above. We provide an overview of several methods proposed in the literature that are compared with our method in an empirical study on a set of technical strategies. Since the strategies are basically random and, in fact, we know their performance following the backtesting period, none should be selected. On the other hand, we also artificially modify one of the strategies making it a "true" discovery and analyze results of the strategy selection methods.

2. An Overview of the Existing Approaches

There are several relatively simple classical methods for adjusting p-values in order to accommodate the multiple test (see, e.g., Streiner and Norman 2011). More advanced and computationally demanding methods are based on various approaches to bootstrapping and simulation of the past and future data.

2.1. Classical Approaches

To test significance of a single strategy, for example, $\mathcal{S}_b$ with sample mean return m_b and standard deviation s_b observed over T periods, the classical approach is to calculate the t-ratio,

$$TR = \frac{m_b}{s_b/\sqrt{T}},$$

and the two-sided[2] p-value,

$$p^S = \Pr[|X| > TR], \tag{1}$$

where X is a random variable following the t-distribution with $T-1$ degrees of freedom. The implicit assumption is that the returns are i.i.d. normal. If the p-value p^S happens to be small enough, e.g., below 5% or 1%, then one tends to jump to the conclusion that a strategy with significantly positive returns has been discovered.

[2] A strategy with a significant negative t-ratio can be considered as a discovery as well since we can revert it in order to achieve systematic positive returns.

The problem with the process of selecting the best strategy or, alternatively, testing a number of strategies until we find a significant one is that the correct p-value should (Harvey and Liu 2015) reflect the fact that we selected the strategy with the best t-ratio TR out of K proposal strategies.

$$p^M = \Pr[\max\{|X_k|, k = 1, \ldots, K\} > TR],$$

where X_k are K random variables following the t-distribution with $T-1$ degrees of freedom (corresponding to the sample t-ratios of the K strategies). It can be noted (Harvey and Liu 2015) that, if the variables were independent, then we could find a simple relationship (also called Šidák's adjustment) between the single and multiple test p-values.

$$p^M = 1 - \prod_k \Pr[|X_k| \leq TR] = 1 - \left(1 - p^S\right)^K = Kp^S - \binom{K}{2}\left(p^S\right)^2 + \cdots.$$

Harvey and Liu (2015) provided an overview of simple adjustment methods, such as Bonferroni's adjustment $p^M = \min\{Kp^S, 1\}$, Holm's, or Benjamini, Hochberg, Zekutieli (BHY) adjustments, using the ordered sequence of the single-test p-values $p_1^S, \ldots, p_K^S$ for all the strategies. The weak point of all those methods is the assumption of independence since the tested strategies are often closely related (e.g., of the same type with varying parameters).

We also propose and test a numerically relatively simple and efficient method on the basis of an estimation of the covariance matrix Ω of the returns and numerically generate the distribution of $\max\{|X_k|, k = 1, \ldots, K\}$ conditional on the null hypothesis $m_k = 0$ for all k where X_k are K random variables following the t-distribution with $T-1$ degrees of freedom (or alternatively standard normal for a large T) and with covariances given by Ω.

Note that the classical (corresponding to the basic period, e.g., daily) Sharpe ratio can be easily calculated given the t-ratio and vice versa.

$$SR_k = \frac{m_k}{s_k} = \frac{TR_k}{\sqrt{T}}.$$

The Sharpe ratio is usually annualized as follows:

$$SR_k^a = \frac{m_k}{s_k}\sqrt{T_a} = TR_k\sqrt{\frac{T_a}{T}},$$

where T_a is the number of observation periods in a year, e.g., 252 in the case of daily returns. According to Equation (1), the maximal acceptable p-value level can be easily translated to a minimum required Sharpe ratio.

Generally, given a selected strategy with an in-sample (according to the backtest data) Sharpe ratio SR_{IS}, the question is what the expected (out-of-sample) Sharpe ratio $E_0[SR_{OOS}]$ over a future, e.g., 1-year period, is. Here, $E_0[.]$ denotes the expectation given all the information available today, in particular, given the in-sample performance such as SR_{IS}, the number of strategies from which the best one was selected, the relationship between the strategies, and the underlying asset return process properties. The Sharpe ratio haircut is then defined as the percentage we need to deduct from the in-sample Sharpe ratio to get a realistic estimate of the future performance,

$$HC = 1 - \frac{E_0[SR_{OOS}]}{SR_{IS}}. \tag{2}$$

Harvey and Liu (2015) noted that the rule-of-thumb haircut used by the investment industry is 50%, but that, according to their analysis, it significantly depends on the level of the in-sample Sharpe ratio and the number of strategies. They proposed to use the relationship between the single- and multiple-test p-values in order to derive the haircut

Sharpe ratio. Their estimate of the annualized expected Sharpe ratio ESR_{HL} is based on the idea that its corresponding single-test p-value should be equal to the adjusted multiple test p-value p^M, i.e.,

$$p^M = \Pr\left[|X| > ESR_{HL}\sqrt{\frac{T}{T_a}}\right], \text{ i.e.,}$$
$$ESR_{HL} = F^{-1}(p^M/2)\sqrt{\frac{T_a}{T}},$$

where X is a random variable following the t-distribution with $T-1$ degrees of freedom and F is its cumulative distribution function. The haircut is then calculated according to Equation (2). The haircut estimation, of course, depends on the p-value adjustment method as Bonferroni, Holm's, BHY, or the general one we suggested above. Although the estimation is obviously directionally correct, it is not obvious why this approach should yield a consistent estimate of the expected Sharpe ratio $E_0[SR_{OOS}]$ and of the corresponding haircut. We compare the different haircut estimates in the simulation study outlined below.

2.2. Stationary Bootstrap

In order to simulate the past and the future returns, we consider a bootstrapping and a cross-validation approach. The *stationary bootstrap* proposed and analyzed in White (2000), Sullivan et al. (1999), and Politis and Romano (1994) is applied to the underlying asset returns assumed to be strictly stationary and weakly dependent time-series to generate a pseudo time series that is again stationary (Figure 2). The tested strategies $\mathcal{S}_1, \ldots, \mathcal{S}_K$ are based only on the single underlying using its historical returns to take long, short, or possibly zero positions.

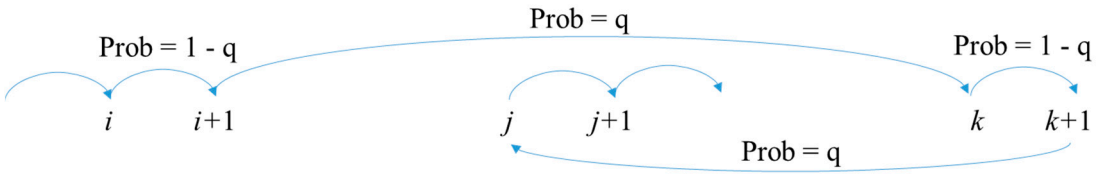

Figure 2. Stationary bootstrap process where the arrows indicate sampling of returns from the original series.

Formally, we generate new sequences of the underlying asset returns $\left\{u_{\Theta(i)}; i = 1, .., \widetilde{T}\right\}$ where $u_1, \ldots, u_{T_1}$ is the original series of observed returns and $\Theta(i) \in \{1, \ldots, T_1\}$. In order to implement the bootstrap, we need to select a smoothing parameter $0 < q = 1/b < 1$, where b corresponds to the mean length of the bootstrapped blocks, for example, $q = 0.1$, proposed by Sullivan et al. (1999). A bootstrapped sequence is obtained by drawing randomly $\Theta(1) \in \{1, \ldots, T\}$, and, for $i = 2, \ldots, \widetilde{T}$, setting $\Theta(i) = \Theta(i-1) + 1$ with probability $1-q$ or randomly drawing a new block starting position $\Theta(i) \in \{1, \ldots, T_1\}$ with probability q. If it happens that $\Theta(i) > T_1$ then we draw random $\Theta(i) \in \{1, \ldots, T_1\}$.

Next, given a bootstrapped sequence of the underlying asset returns, we need to apply strategies $\mathcal{S}_1, \ldots, \mathcal{S}_K$ to get the strategies' bootstrapped returns $\widetilde{r}_{k,t}$, $k = 1, \ldots, K$, $t = 1, \ldots, T_1$. Note that, since the strategies' decisions are built on the basis of past returns, we generally need to have a longer series of the bootstrapped asset returns, $\widetilde{T} > T_1$, where T_1 is the length of the in-sample period. Then, we evaluate our desired performance indicator values (mean return, Sharpe ratio, etc.) $\widetilde{f}_k$. Let f_k^* denote the performance indicators of the original series of returns. According to White (2000), under certain mild theoretical assumptions, the empirical distribution of the B bootstrapped values $\widetilde{V}_j = \max_{k=1,\ldots,K}\left(\widetilde{f}_k - f_k^*\right)$ for $j = 1, \ldots, B$ asymptotically converges to the distribution of the best strategy performance indicator under the null hypothesis H_0 that all the strategies have zero performance. That is, obtaining B bootstrapped values $\left\{\widetilde{V}_j; j = 1, \ldots, B\right\}$, we can test H_0 by calculating the empirical p-value $\Pr\left[\left|\widetilde{V}_j\right| > f_b^*\right]$, where b is the index of the best strategy and $f_b^* = \max_k f_k^*$.

The bootstrap technique can be also used to analyze the relationship between IS and OOS Sharpe ratio (or another indicator), generating series of strategy returns over a time period $1, \ldots, T_1$, selecting the best strategy $\mathcal{S}_b$ with an in-sample performance SR_{IS}, and then looking at its out-of-sample performance SR_{OOS} over the following period $T_1 + 1, \ldots, T_1 + T_2$. Note that, in this case, the original bootstrapping must be done over a period of length $\widetilde{T} > T_1 + T_2$. Then, we can compare the mean SR_{OOS} against the mean SR_{IS} or conditional on certain level of SR_{IS}. We may also bootstrap the OOS returns for the actually selected strategy $\mathcal{S}_b$ (according to the real dataset). However, in this case, it is especially obvious that even a truly positive strategy that is using medium-term or long-term trends to make good predictions does not have to work on the bootstrapped series of returns where the future and past returns of the original series are to a large extent mixed up. Therefore, the estimated conditional SR_{OOS} may easily lead to a false rejection of a positive strategy.

2.3. Combinatorial Symmetric Cross-Validation

Another disadvantage of the stationary bootstrap technique is that it cannot be applied if we are given only the strategy returns but not details on the strategies themselves. The stationary bootstrap is also problematic if the strategies are not technical ones and use a number of additional, possibly lagged, explanatory factors. This is not the case for the *combinatorial symmetric cross-validation* (CSCV) (Bailey et al. 2014, 2016) utilizing only the matrix of the strategies' returns $M = \{r_{k,t},\ k = 1, \ldots, K,\ t = 1, \ldots, T_1\}$. The idea is to split the time window of length $T_1 = SN$ into S blocks of length N, where S is even and draws combinations of $S/2$ blocks (Figure 3). The submatrix J formed by joining $T_1/2$ rows of M corresponding to the selected time indices in the original order then represents an in-sample dataset of returns where the best performing strategy can be selected while the complementary $K \times T_1/2$ submatrix $\bar{J}$ represents the out-of-sample returns. The sampling can be done with or without replacement. Since there are $\binom{S}{S/2}$ combinations, we can sufficiently form many different combinations with replacement as long as S is sufficiently large, e.g., at least 16.

Figure 3. An example of the *combinatorial symmetric cross-validation (*CSCV) combination for $S = 6$ where green blocks of returns are sampled as in-sample and blue blocks are sampled as out-of-sample.

Bailey et al. (2014, 2016) proposed using the technique to specifically estimate the probability of backtest overfitting (PBO) defined as the probability that the best IS selected strategy performs below the average OOS. More precisely, for K strategies $\mathcal{S}_1, \ldots, \mathcal{S}_K$,

$$PBO = \Pr[Rank_{OOS}(X) < K/2 | Rank_{IS}(X) = 1].$$

The PBO indicator, as well as the Sharpe ratio haircut, can be estimated using sufficient cross-validation pairs of the IS/OOS datasets $\langle J, \bar{J} \rangle$. However, it is obvious that the estimates are biased, introducing a negative drift into the OOS order of the strategies. For example, if all the strategies represented just pure noise with mean returns over the full time interval $\{1, \ldots, T_1\}$ close to zero, then, for an IS/OOS combination $\langle J, \bar{J} \rangle$, the best strategy IS return $\bar{r}_{b,J}$ implies that the complementary OOS return $\bar{r}_{b,\bar{J}} \approx -\bar{r}_{b,J}$ would probably be the worst on $\bar{J}$. We demonstrate the effect in the empirical section. The cross-validation technique also cannot be used, due to this property, to estimate the OOS Sharpe ratio or mean for a particular selected strategy. We can just estimate the overall PBO or Sharpe ratio haircut keeping in mind that the estimations incorporate a conservative bias. The cross-validation, as well as the bootstrapping approach, cannot be easily used to estimate the false discovery rate (FDR) since it is not clear how to identify true and false discoveries

given a CSCV simulation. This could be possibly done by testing the significance of OOS performance involving an ad hoc probability level.

We show that all the indicators of interest can be consistently estimated in the Bayesian setup outlined below.

3. Bayesian Simulation Approach

The Bayesian approach is based on the scheme given in Figure 4.

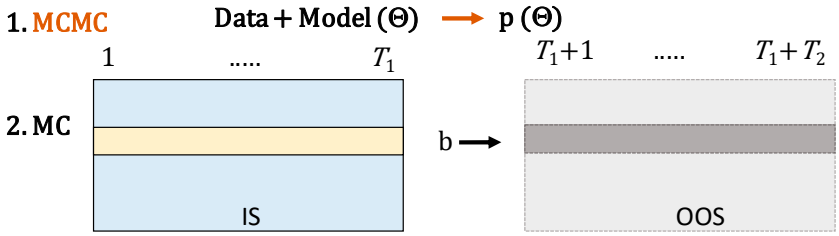

Figure 4. Two-step Bayesian simulation (Markov chain Monte Carlo (MCMC) parameter estimation and Monte Carlo (MC) data simulation).

First, of course, we need to specify a model defining the return generating process with unknown parameters Θ for the observed strategy returns $\{r_{k,t},\ k = 1,\ldots,K,\ t = 1,\ldots,T\}$, where $T = T_1$ or $T = T_1 + T_2$. Then, the plan is to use a Bayesian technique, specifically, the Markov Chain Monte Carlo (MCMC) simulation, in order to extract the posterior distribution of the model parameters Θ. Finally, we simulate the matrices of IS and OOS returns over desired time intervals $1,\ldots,T_1$ and $T_1+1,\ldots,T_1+T_2$. The Monte Carlo (MC) is done in two steps, always selecting the parameters Θ from the posterior distribution and then generating K series of $T_1 + T_2$ returns according to the model. The simulated IS returns can be used to select the best strategy and the OOS returns to measure its future performance. The average haircut or average relative rank can be easily estimated as in the case of the stationary bootstrap (see Figure 4).

We consider two models; the simple one assumes that the returns are multivariate normal with unknown covariance matrix and means, while the second incorporates unknown indicators of truly profitable strategies allowing us to consistently estimate the false discovery rate (FDR). The second model follows an idea of Scott and Berger (2006), also mentioned in Harvey et al. (2016); nevertheless, in both cases, the model was formulated only for observed mean returns and without considering a correlation structure of returns. It should be emphasized that our focus is to analyze the impact of backtest overfitting assuming that the strategies' cross-sectional returns behave in a relatively simple and stable way over time similarly to the classical, bootstrapping, or cross-validation approaches. One could certainly come up with state-of-the-art models incorporating jumps, switching regimes, stochastic variances, or even dynamic correlations. These improvements would make the methodology computationally difficult to manage with results probably even more conservative compared to the approaches we consider below.

3.1. The Naïve Model 1

To set up the naïve model, we assume that the cross-sectional strategy returns are multivariate normal,

$$\boldsymbol{r}_t = \langle r_{1,t},\ldots,r_{K,t}\rangle \ \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma}),$$

and that the observations over time are independent.

Given data $= \langle r_t\rangle$, i.e., the matrix of back test returns, and possibly some priors for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we can find the posterior distribution $p(\boldsymbol{\mu},\boldsymbol{\Sigma}|\text{data})$ using the standard Gibbs MCMC sampler.

Specifically, the iterative sampling is given by

$$p(\boldsymbol{\mu}|\boldsymbol{\Sigma}, \text{data}) = \varphi\left(\boldsymbol{\mu}; \frac{1}{\mathrm{T}}\sum_{t=1}^{T} \boldsymbol{r}_t, \frac{1}{T}\boldsymbol{\Sigma}\right), \text{ and}$$
$$p(\boldsymbol{\Sigma}|\boldsymbol{\mu}, \text{data}) = IW(\boldsymbol{\Sigma}; T, S),$$

where $S = \sum_{t=1}^{T} (\boldsymbol{r}_t - \boldsymbol{\mu})'(\boldsymbol{r}_t - \boldsymbol{\mu})$ is the scale matrix (i.e., the covariance matrix times T), and IW is the inverse Wishart distribution. For example, Matlab allows sampling from the distributions, and the posterior distribution may be obtained quite efficiently (e.g., 10,000 runs of the sampler).

Remark 1. *The sampler above assumes the noninformative prior on the means,* $p(\boldsymbol{\mu}) \propto 1$, *and the standard improper prior on the covariance matrix*,

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{K+1}{2}}.$$

Given the extracted posterior distribution $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\text{data})$, the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ can be now easily sampled in order to get the empirical distribution of the selected strategy performance. However, in the process of selecting the best strategy, we do not know $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ but only a time series of the backtested returns with cross-sections from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. On the basis of the time series, the "best" strategy $\mathcal{S}_b$ is selected. Our key question is about its expected forward-looking performance, e.g., μ_b or SR_b. Therefore, we need to run the following Monte Carlo simulation in order to faithfully sample the empirical distribution of the performance indicators:

1. Sample $\langle \boldsymbol{\mu}, \boldsymbol{\Sigma} \rangle$ from $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\text{data})$.
2. Independently sample $T_1 + T_2$ cross-sections $R_t \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
3. Determine the index of the best strategy b on the basis of the backtest statistics calculated from the matrix of backtested returns $\boldsymbol{R} = \langle \boldsymbol{R}_t \rangle$ for $t = 1, .., T_1$.
4. Calculate and store the performance indicators, $\hat{\mu}_b$, $S\hat{R}_b$, in the OOS period $T_1 + 1, \ldots, T_2$. Alternatively, store the selected strategy "true" performance indicators, i.e., μ_b, SR_b.

The MCMC estimation of the multivariate normal distribution parameters is known to converge relatively fast (see e.g., Lynch 2007); nevertheless, we apply a burn-out period according to simple diagnostics, e.g., the average of the vector of mean returns $\boldsymbol{\mu}$. The simulated posterior distribution of the desired performance indicators (after removing the burn-out period) then tells us the mean, median, confidence intervals, or Bayesian probabilities where the true performance is positive or above any given minimum threshold. The ratio between the ex post and ex ante performance indicators also gives us an estimate of the "backtest overfitting haircut".

3.2. Model 2—Bimodal Means Distribution

In order to capture the situation when most strategies are random and only some positive (nonzero), assume that there are, in addition, latent indicators $\gamma_i \in \{0, 1\}$ so that the mean of strategy i is $\mu_i^* = \gamma_i \mu_i$. Therefore, the row vector of returns has the distribution

$$\boldsymbol{r}_t = [r_{1,t}, \ldots, r_{K,t}] \sim N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}).$$

Here, we need to assume a prior distribution for $\gamma_i \sim Bern(1 - p_0)$ and $\mu_i \sim N(m_0, V_0)$. It means that the Bayesian distribution of the means is bimodal with a large probability mass on 0, and the other mode is normal with prior mean $m_0 > 0$ and variance V_0. The Gibb's sampler can be modified as follows:

1. Given $\boldsymbol{\mu}, \boldsymbol{\gamma}$, set $\mu_i^* = \gamma_i \mu_i$, and estimate $\boldsymbol{\Sigma}$ as above, i.e.,

$$p(\boldsymbol{\Sigma}|\boldsymbol{\mu}, \boldsymbol{\gamma}, \text{ data}) = IW(\boldsymbol{\Sigma}; T, S), \text{ where } S = \sum_{t=1}^{T} (\boldsymbol{r}_t - \boldsymbol{\mu}^*)'(\boldsymbol{r}_t - \boldsymbol{\mu}^*).$$

2. Given $\boldsymbol{\Sigma}$, $\boldsymbol{\gamma}$, estimate $\boldsymbol{\mu}$. Set $A = \frac{1}{T}\boldsymbol{\Sigma}$, $\Gamma = diag(\boldsymbol{\gamma})$, $\boldsymbol{m} = \frac{1}{T}\sum_{t=1}^{t} \boldsymbol{r}_t$, $\boldsymbol{m}_0 = [m_0, \ldots, m_0]$, and $D = diag([V_0, \ldots, V_0])$, where *diag* creates a matrix with diagonal elements given by the vector in the argument, and sample

$$p(\boldsymbol{\mu}|\boldsymbol{\Sigma}, \boldsymbol{\gamma}, \text{ data}) = \varphi\Big(\boldsymbol{\mu}; \Big(\Gamma A^{-1}\Gamma + \mathrm{D}^{-1}\Big)\Big(\Gamma \mathrm{A}^{-1}\boldsymbol{m} + D^{-1}m_0\Big), \Big(\Gamma A^{-1}\Gamma + \mathrm{D}^{-1}\Big)^{-1}\Big).$$

3. Given $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$, estimate $\boldsymbol{\gamma}$. For $i = 1, \ldots, K$ set Γ_0 equal to Γ with the exception of the diagonal element $\Gamma_0(i,i) = 0$, and Γ_1 setting $\Gamma_1(i,i) = 1$. Let

$$L_0 = \exp\Big(\frac{-1}{2}\Big((\Gamma_0\boldsymbol{\mu} - \boldsymbol{m})'A^{-1}(\Gamma_0\boldsymbol{\mu} - \boldsymbol{m})\Big)\Big)(1 - p_0),$$

$$L_1 = \exp\Big(\frac{-1}{2}\Big((\Gamma_1\boldsymbol{\mu} - \boldsymbol{m})'A^{-1}(\Gamma_1\boldsymbol{\mu} - \boldsymbol{m})\Big)\Big)p_0,$$

$$\widetilde{p} = \frac{L_1}{L_0 + L_1}, \text{ and finally, sample } \gamma_i \sim Bern(\widetilde{p}).$$

Proofs of the formulas used in steps 2 and 3 can be found in Appendix A. Again, we apply a burn-out period according to a convergence diagnostic of the averages of $\boldsymbol{\gamma}$ and of $\boldsymbol{\mu}$.

Remark 2. *There are certain possible extensions.*

- *We may allow $\gamma_i \in \{-1, 0, 1\}$ encoding negative significant mean return, zero return, or significant positive return. In this case, the mean parameter of the prior distribution $\mu_i \sim N(m_0, V_0)$ must be strictly positive. In the Gibb's sampler above, we just need to modify step 3 in a straightforward manner.*
- *The hyper-parameters p_0, m_0, V_0 for $\gamma_i \sim Bern(1 - p_0)$ and $\mu_i \sim N(m_0, V_0)$ might be estimated within the MCMC procedure. In this case, the Gibb's sampler can be extended as follows:*

  4. *Sample p_0 given $\boldsymbol{\gamma}$:*

$$p(p_0|\boldsymbol{\gamma}) \propto p_0^{n_1}(1 - p_0)^{1-n_1} p(p_0) \propto Beta(p_0; n_1 + k1 + 1, K - n_1 + k2 + 1),$$

  where $n_1 = \#\{i; \gamma_i = 1\}$ and $p(p_0) = Beta(p_0; k1 + 1, k2 + 1)$ is a conjugate prior distribution (e.g., $k_1 = k_2 = 1$).

  5 *Sample m_0, V_0 given $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$. Here, we just use the means $\{\mu_i|\gamma_i = 1\}$ where the signal is positive and the normal Gibb's sampler. Since the set may be empty, we need to use proper conjugate priors, e.g., $p(m_0) = \varphi(m_0; 0, m_P)$ and*

$$(V_0) \propto IG\Big(V_0, \frac{k_0}{2}, \frac{k_0 V_P}{2}\Big).$$

  For $\widetilde{K} = \#\{\mu_i|\gamma_i = 1\} \neq 0$, set $\widetilde{\mu} = \sum\{\mu_i|\gamma_i = 1\}/\widetilde{K}$ and $\widetilde{V} = \sum\Big\{(\mu_i - m_0)^2\Big|\gamma_i = 1\Big\}/\widetilde{K}$. Then,

$$p(m_0|\boldsymbol{\mu}, \boldsymbol{\gamma}, V_0) \propto \varphi(m_0; \widetilde{\mu}, V_0/\widetilde{K})\ \varphi(m_0; 0, V_P) \propto \varphi\Big(m_0; \frac{\widetilde{\mu}\widetilde{K}V_P}{\widetilde{K}V_P + V_0}, \frac{V_P V_0}{\widetilde{K}V_P + V_0}\Big) \text{ and}$$
$$p(V_0|\boldsymbol{\mu}, \boldsymbol{\gamma}, m_0) \propto IG\Big(0; \frac{\widetilde{K}}{2}, \frac{\widetilde{K}\widetilde{V}}{2}\Big)\ IG\Big(V_0, \frac{k_0}{2}, \frac{k_0 V_P}{2}\Big) \propto IG\Big(V_0, \frac{\widetilde{K} + k_0 + 1}{2}, \frac{\widetilde{K}\widetilde{V} + k_0 V_P}{2}\Big).$$

  If $\widetilde{K} = 0$. then we have to sample on the basis of conjugate priors $p(m_0)$ and $p(V_0)$ only.

4. Numerical Study

Following Sullivan et al. (1999) and other studies, we compare and illustrate the proposed Bayesian methods on a set of technical strategies' returns. We also artificially modify the mean returns of the strategies in order to test the methods, on one hand, if there is a clearly extraordinary strategy or, on the other hand, if the returns of the returns of all the strategies are very low.

4.1. Technical Strategies Selection

We used 1000 daily S&P 500 values and returns for the period 5 June 2009–24 May 2013 (1000 trading days). The period was selected with the purpose of finding at least one strategy with a higher mean return. As in Sullivan et al. (1999), we applied the filter, moving average, support, and resistance rules with varying parameters. We randomly selected 200 strategies with the condition that the daily return series are not collinear (it may even happen that the series are identical if the parameters do not differ too much).

The means and Sharpe ratios of the individual strategies' return series and their densities over the period 5 June 2009–24 May 2013 are shown in Figure 5. It should not be surprising that the strategies' returns are mostly positively correlated with the average pairwise correlation 23.32%. Note that strategy 7 is apparently the best with the annualized ($n_y = 252$) mean return over 21% p.a. and Sharpe ratio approximately 1.2 (it is a filter strategy with $x = y = 1\%$, i.e., a long or short position is taken if the previous daily return is over 1% or below -1%, respectively, and the minimum number of days to stay in a position is 20). The strategy returns look attractive; nevertheless, looking forward, it turns out that the realized mean return of the strategy in the 1000 days following 24 May 2013 is negative (-5.21%).

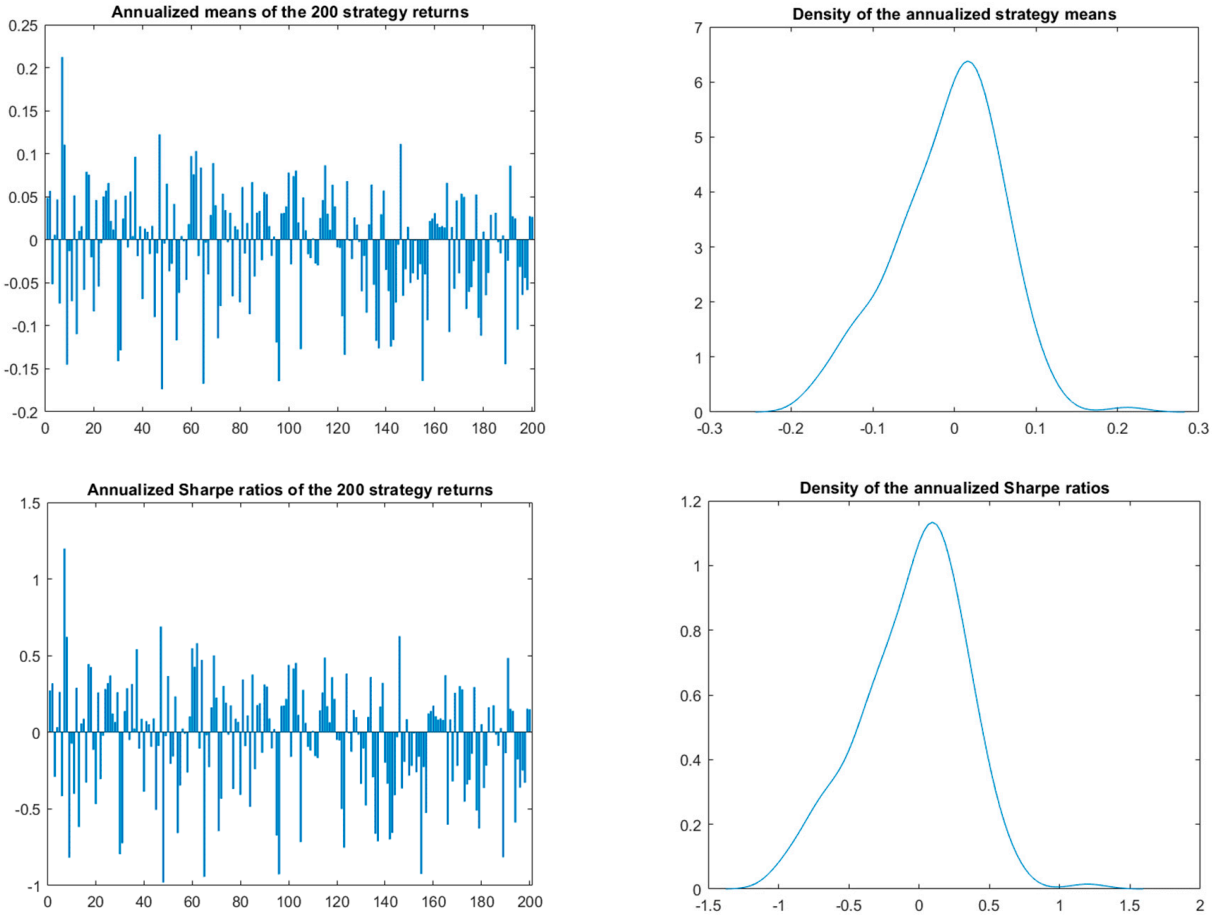

Figure 5. Annualized means of the selected 200 strategies and the Sharpe ratios' probability distributions.

4.2. Single and Multiple p-Value Testing

The single-test annualized t-ratio and the p-value of the best strategy 7 with $SR = 1.1987$ are

$$TR = SR\sqrt{\frac{n_{obs}}{n_y}} = 1.1987\sqrt{1000/252} = 2.3878 \text{ and } p^S = 0.0171.$$

The multiple test p-value after Bonferroni adjustment is simply

$$p^M = \min\{200 \times 0.171, 1\} = 1.$$

Thus, the adjusted expected Sharpe ratio is 0 and the haircut is 100%. Šidák's adjustment yields only a slightly more optimistic result with $p^M = 0.9685$, adjusted expected $SR = 0.02$, and haircut 98.3%. Similar adjustments can be obtained by the Holm or Benjamini, Hochberg, Yekutieli (BHY) methods using, for example, the package provided by Harvey and Liu (2015). The simple adjustment methods allow to easily estimate the minimum return of the best strategy (keeping the same covariance structure) in order to get the multiple test p-value at most 5%. The estimated minimum return using the same package is around 36%.

4.3. Multivariate Normal Simulation and the Stationary Bootstrap According to the Null Hypothesis

Another relatively simple possibility is to estimate the return covariance matrix and simulate the future multivariate returns on the basis of the return covariance matrix and conditional on zero means. Figure 6 below shows the density of 1000 simulated annualized SR as a function of a 1000 day period. The adjusted p-value of the best strategy with $SR = 1.1987$ is then relatively optimistic 0.352 and the adjusted expected SR is 0.4683 (i.e., the implied haircut is just 61%).

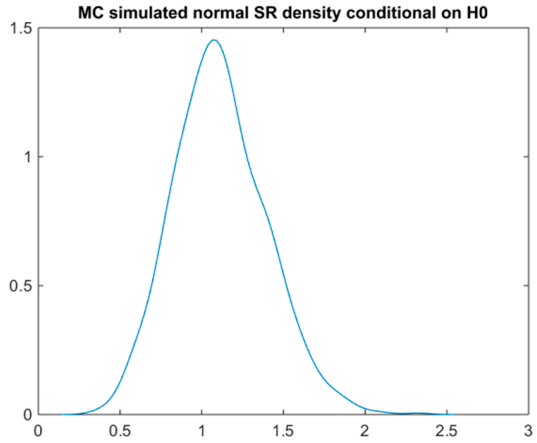

Figure 6. Sharpe ratio's probability density according to the null hypothesis and the multivariate normal MC simulation.

An analogous distribution (Figure 7) can be obtained via the much more computationally demanding stationary bootstrap (White's reality check). The p-value according to 1000 bootstrap simulations for a 1000-day time period and with $q = 0.1$ turns out to be 0.728, the corresponding adjusted expected SR is 0.175, and the SR haircut is 85.4%.

4.4. Stationary Bootstrap Two-Stage Simulation

The stationary bootstrap method can also be used to simulate the backtest period of length $T_1 = 1000$, as well as the future period with $T_2 = 1000$. The number of stationary bootstrap iterations is again 1000 according to the 5 June 2009–24 May 2013 window of S&P returns, and the parameter is set to $q = 0.1$ with the corresponding average length

of the bootstrapped blocks being 10. Figure 8 shows the typical strong shift of the ex ante performance density to the left-hand side and the wider ex post performance density. The results show that the best IS selected strategy performs poorly OOS with 32.8% probability of loss, PBO around 0.44 (see also Figure 9), and SR haircut over 73%. For detailed results including the ex ante and ex post SR or mean return values, see the summary in Table 1. Note that the row "stationary bootstrap" shows values obtained via the two-stage simulation except the *p*-value estimated by White's reality check.

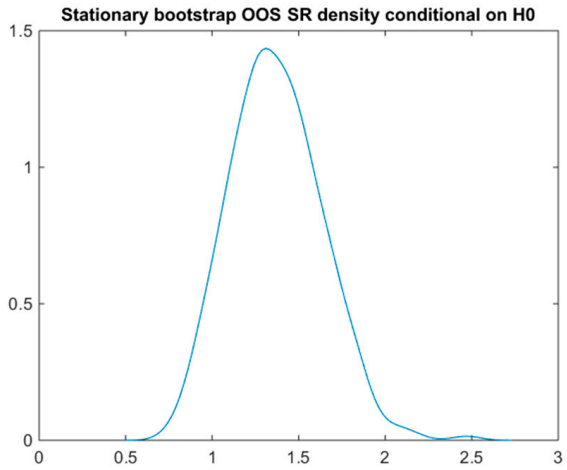

Figure 7. Sharpe ratio's probability density according to White´s reality check.

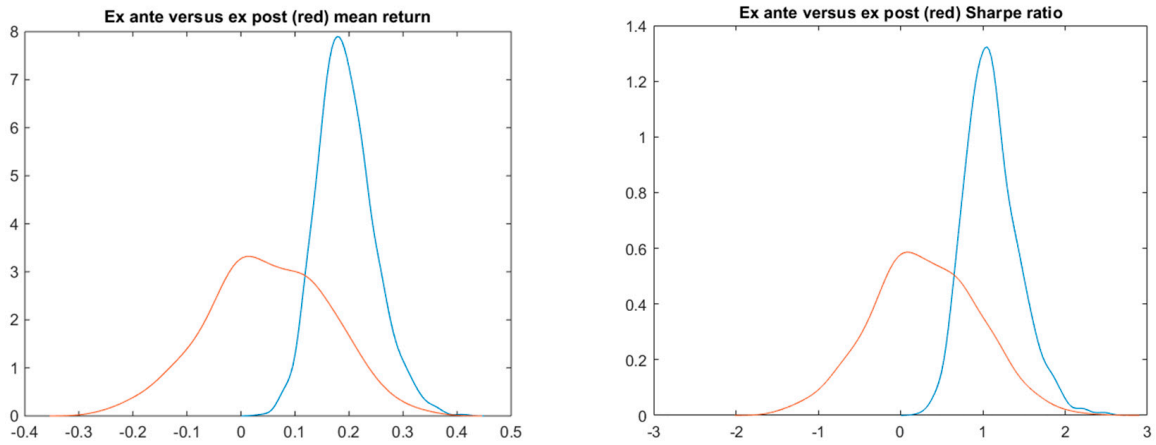

Figure 8. Stationary bootstrap simulation Sharpe ratio and the mean return ex ante (blue) and ex post (red) probability densities.

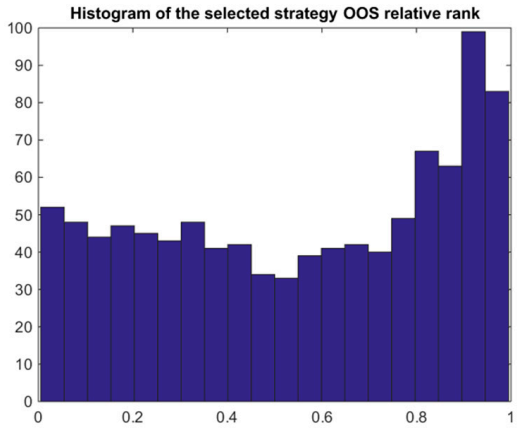

Figure 9. Stationary bootstrap simulation histogram of the OOS relative rank of the best IS strategy.

Table 1. Summary of the backtest overfitting tests' results.

	Adjusted *p*-Value (FDR)	Ex Ante av. SR/Mean	Adjusted Expected SR/Mean	SR/Mean Hair Cut	Probability of Loss	Mean OOS Rank	PBO
Boferroni method	1.00	1.199	0	100%	-	-	
Šidák's correction	0.968	1.199	0.02	98.3%	-	-	
Mult. norm. MC adj.	0.352	1.199	0.4668	61%	-	-	
Stationary bootstrap	0.728	1.110/0.194	0.297/0.051	73.2%/74%	0.328	55%	0.444
CSCV	-	1.382/0.244	0.336/0.058	75.7%/76.4%	0.371	66.8%	0.323
Bayes mod. 1	-	2.102/0.371	1.014/0.180	51.8%/51.5%	0.171	75.7%	0.168
Bayes mod. 2	0.549	1.201/0.213	0.211/0.037	82.5%/82.5%	0.380	60%	0.395

4.5. Combinatorial Symmetric Cross-Validation

In order implement the CSCV algorithm, we chose the number of blocks 20 corresponding to

$$\begin{pmatrix} 20 \\ 10 \end{pmatrix} = 184,756$$

combinations of 10 blocks of length 50. However, we sampled only 1000 combinations. In this case, we always split the 1000-day time into the IS and OOS parts on the same length, i.e., $T_1 = 500$ and $T_2 = 500$. The results shown in Table 1 are quite like the stationary bootstrap only with PBO being slightly lower (0.323). The slightly better performance is reflected in the bimodal ex post densities in Figure 10 where the right-hand side positive mode corresponds to the selected strategy that performs well IS, as well as OOS. Figure 11 also indicates that, in this case, compared to Figure 9, the best IS strategy remains the best OOS quite often.

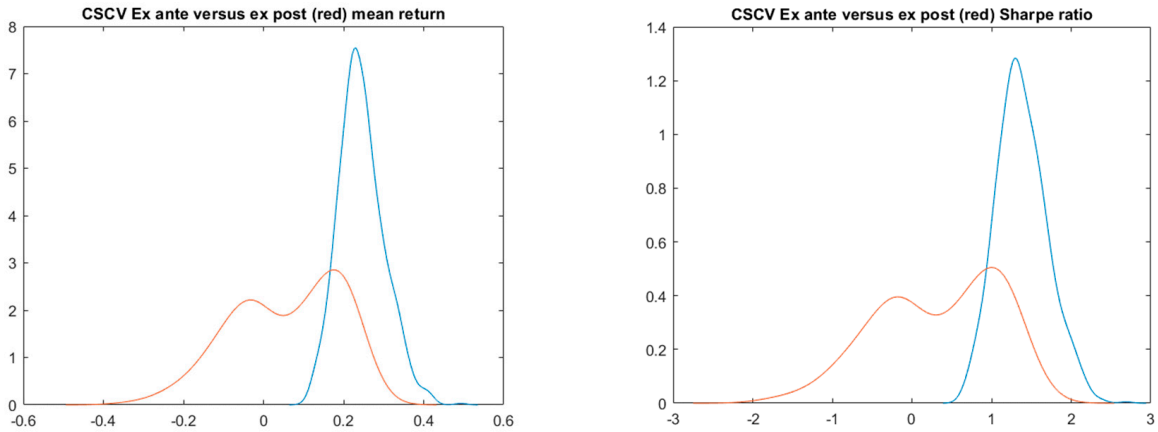

Figure 10. CSCV simulation of the ex ante (blue) and ex post (red) probability densities.

4.6. The Naïve Bayes Model 1

In the Bayesian approach, we firstly extract the multivariate normal model means and covariance given the observed data. This can be done in 1500 iterations using the standard Gibbs sampler including 500 burn-out simulations, i.e., only the last 1000 iterations are used to estimate the posterior parameters (see Figure A1, Appendix A, for the fast MCMC convergence). In the MC simulation, we can choose the length of the backtest (IS) period $T_1 = 1000$ and the OOS forward looking period $T_2 = 1000$. We then generated

1000 scenarios (after removing the burned-out simulations) sampling the parameters and the cross-sectional returns, selecting the best IS strategy, and measuring its OOS performance. We have to keep in mind that the sampled posterior means may differ quite significantly from the observed mean returns due to the high return volatility[3] and, thus, the observed best strategy 7 may look weak in the simulations, while other strategies are selected as the best. The detailed results given in Table 1 and Figure 12 confirm that the naïve multivariate normal model indeed appears too optimistic. Indeed, the best IS strategy remains the best quiet often as shown in Figure 13 and the PBO is quite low (0.168). This can be explained by the simple multivariate normal model and relatively long IS window allowing to identify the truly positive strategy.

In reality, if we generate a large number of models and the best model performance is still poor, then we are probably not going to enroll it for real trading. Therefore, we might also consider a minimum hurdle at which we choose or reject the best selected model. This is quite easy given the simulation outputs. For example, if we set the minimum SR to 1.2, then the condition will be satisfied in 96.6% of the simulations with average ex ante SR 2.14 and ex post SR 1.00, i.e., again with the haircut slightly over 53%. It is interesting that the haircut is not much sensitive to the hurdle, e.g., if the minimum SR was 2, then the corresponding average haircut would be even higher (37%). Nevertheless, the probability of loss can be reduced by setting the minimum SR higher, e.g., if we set the hurdle to 2, then the conditional probability of loss would decline to 16.5% (conditional on the strategy selection) and the unconditional probability would decline to 9.1% since 44.9% of the proposed models would be rejected in the simulation.

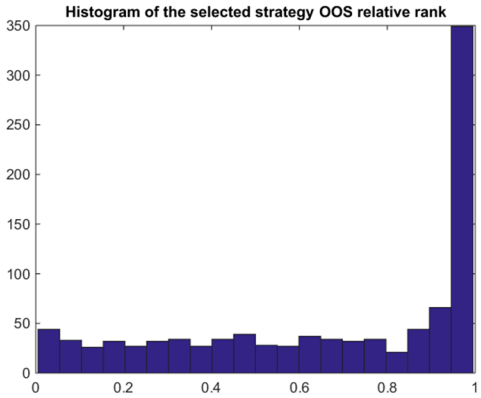

Figure 11. CSCV histogram of the out-of-sample relative rank of the best IS strategy.

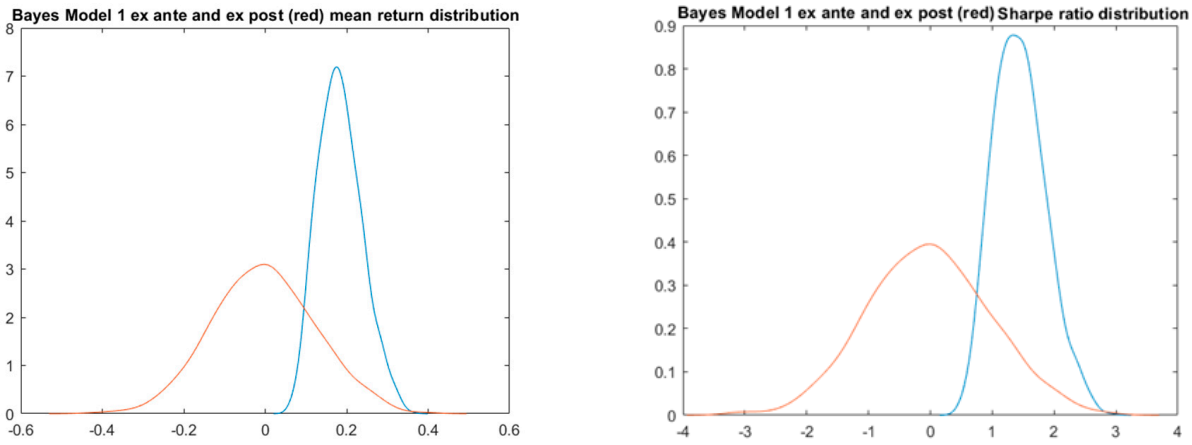

Figure 12. Bayes model 1 simulation of the ex ante and ex post probability densities.

[3] For example, 18% annualized volatility of returns is translated to $18\%\sqrt{252/100} \cong 9\%$ volatility of the posterior annualized mean return.

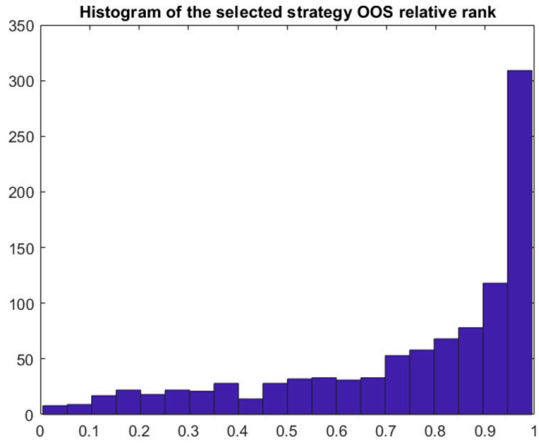

Figure 13. Bayes model 1 histogram of the out-of-sample relative rank of the best IS strategy.

4.7. Bayes Bimodal Mean Returns Model 2

In this case, in addition to the multivariate normal distribution with unknown parameters, we also consider latent indicators of zero and nonzero models. In order to extract the posterior distribution of the parameters and the latent indicators, we again run 1500 iterations and use only the last 1000 iterations to estimate the parameters (removing first 500 burn-out iterations—see Figure A1, Appendix A, for the relative MCMC convergence) of the Gibbs sampler outlined in Section 3.2. Analogously to Bayes Model 1, we then run 1000 Monte Carle simulations with $T_1 = 1000$ and $T_2 = 1000$. Since, in this case, the Bayesian model incorporates the uncertainty whether the model is a true discovery or not, the results should be more conservative compared to the naïve model. Indeed, the PBO turns out to be 0.395, substantially lower compared to model 1, the SR haircut is 82.5%, and the probability of loss is 38% (see Table 1, Figures 14 and 15).

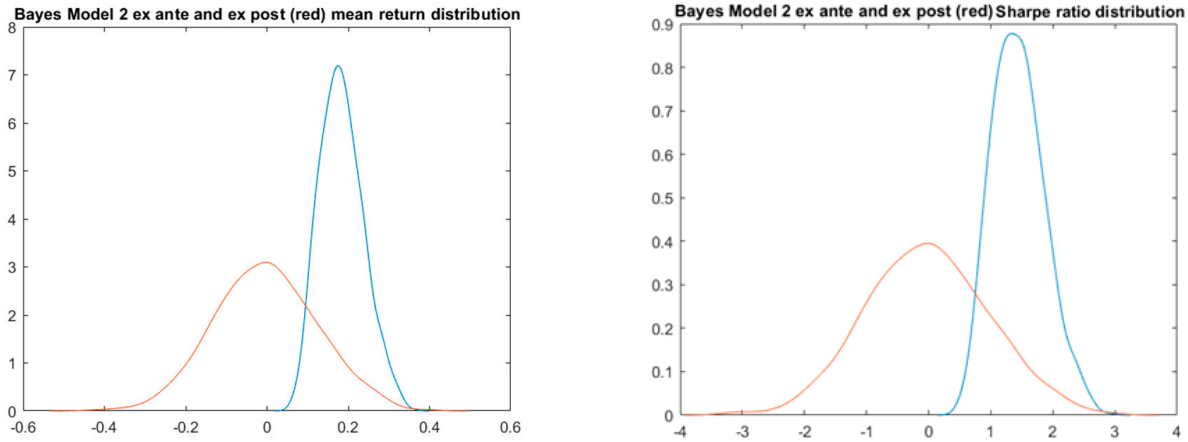

Figure 14. Bayes model 2 simulation of the ex ante (blue) and ex post (red) probability densities.

The model also provides posterior averages of γ_i for each individual model i (see Figure 16). The averages can be interpreted as Bayesian probabilities that the models are true discoveries. There are a few models with the averages over 80%, including the model 7 with the value over 86%. The complements of these Bayesian probabilities to 100% can be in a certain sense compared to the frequentist single-test p-values. However, the Bayesian model also allows us to answer the key question we are asking: Given the observed data and the general model assumptions, what is the probability that the best strategy b selected on the basis of observed data is a true discovery, i.e., $\gamma_b = 1$? This can be estimated as the mean of γ_b which turns out to be only 0.459. This means that, applying the selection

process, we identify the true discovery only in 45.1% of cases and we make a false discovery in 54.9% of cases, i.e., *FDR* = 54.9% can be shown instead of the adjusted *p*-value in Table 1.

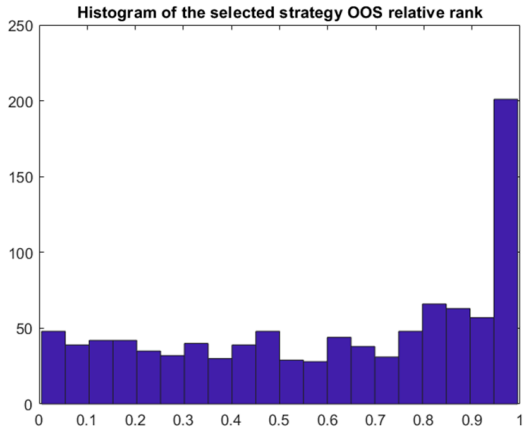

Figure 15. Bayes model 2 histogram of the out-of-sample relative rank of the best IS strategy.

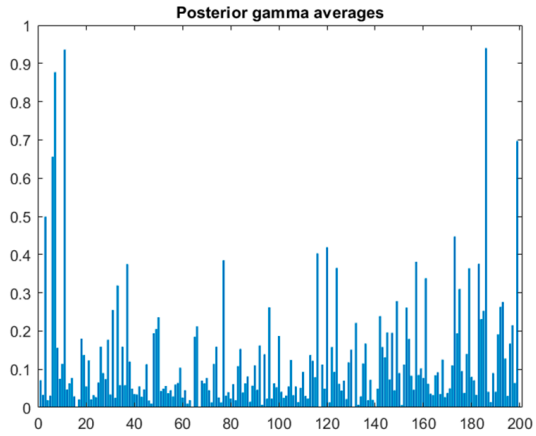

Figure 16. Posterior gamma average values for the 200 strategies.

We may again test whether a higher SR hurdle reduces the high SR haircut. The results are similar for the naïve model 1, i.e., the SR haircut stays around 82% more or less independently of the hurdle. The probability of loss can be reduced only slightly, e.g., for the hurdle of 1.5, the conditional probability of loss declines to 32.1%, but the unconditional probability of loss goes significantly down to 5.2% as 83.8% of the best strategies are rejected in the simulation.

It is also interesting to look at the dependence of the average posterior gamma depending on the number of strategies tested, e.g., the first 10, 20, . . . , 200 (with the number of simulations again being 1000 after removing 500 burn-out iterations, $T_1 = 1000$, $T_2 = 1000$). Note that the best observed strategy is included in the first ten strategies; however, as expected, the posterior expected gamma, mean, or SR of the best strategy declines with the number of strategies tested (Figure 17). The result of the simulation is expected, i.e., with an increasing number of strategies from which the best one is selected, the expected out-of-sample performance decreases.

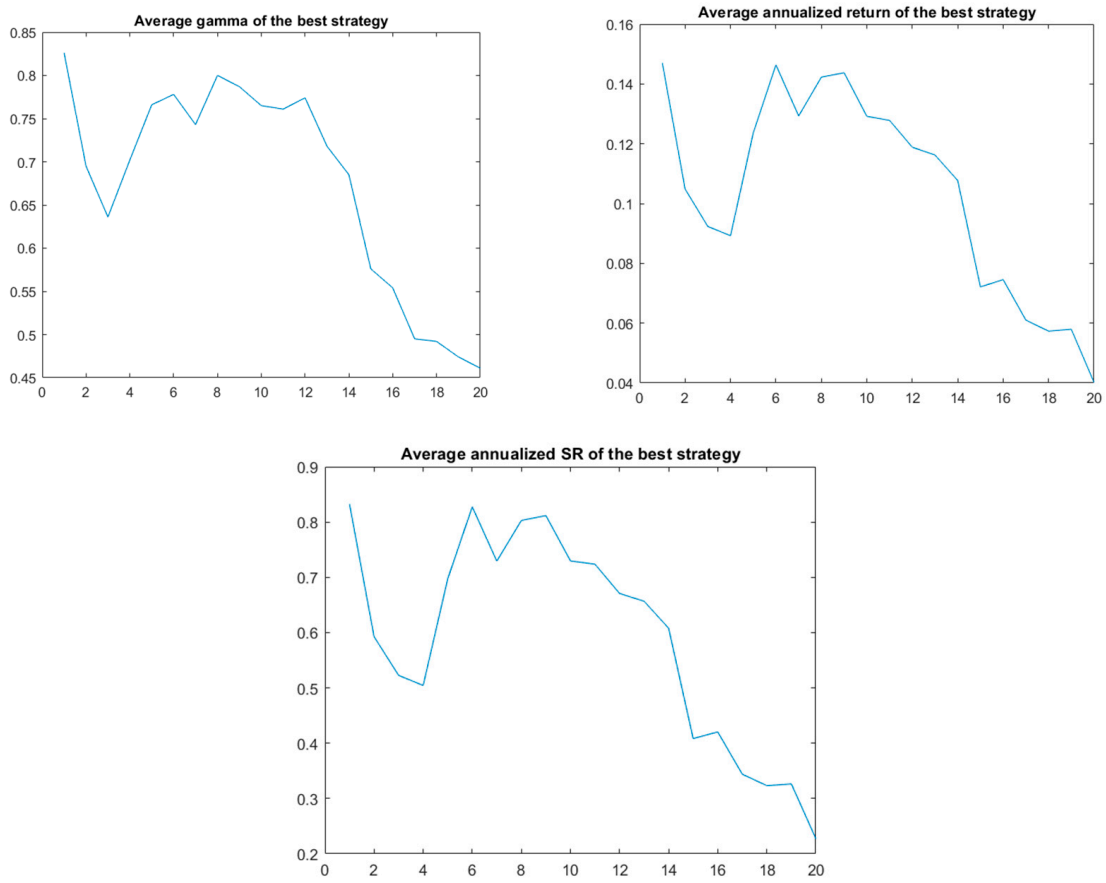

Figure 17. Estimated values of the average ex post gamma, mean return, and SR depending on the number of strategies tested.

4.8. Testing with Modified Mean Returns

In order to better compare the methods, we modify the vector of returns of the strategies while keeping the "natural" correlation structure. Firstly, we increase the strategy 7 return by 19% p.a. while keeping the other returns unchanged so that the strategy 7 with mean over 40% and SR 2.27 stands out among the others (Figure 18), and one expects that it should be identifiable as significant using the various methods.

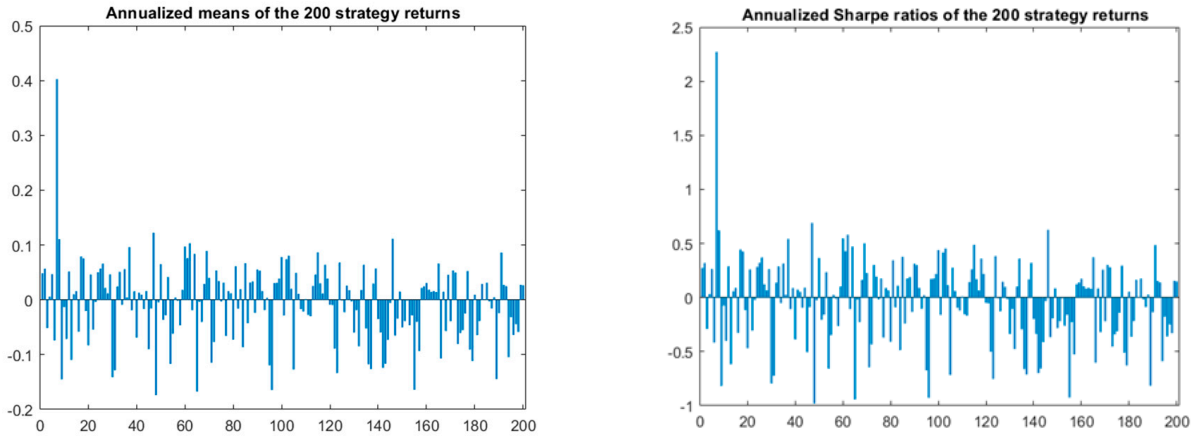

Figure 18. The modified annualized mean returns and Sharpe ratios of the 200 strategies.

Table 2 shows the results (for simplicity, focusing only on SR values). Note that, in this case, we are not able to implement the stationary bootstrap since there is no real strategy behind the modified returns of "strategy" 7 and, thus, the row is missing.

Table 2. Summary of the test results.

	Adjusted *p*-Value (FDR)	Ex Ante av. SR	Adjusted Expected SR	Hair Cut	Probability of Loss	Mean Rank	PBO
Boferroni method	0.0014	2.2707	1.6121	29%	-	-	
Šidák's correction	0.0014	2.2707	1.6122	29%	-	-	
Multivariate norm. MC adj.	0.0004	2.2707	1.783	21.5%	-	-	
CSCV	-	2.257	2.196	2.7%	0.014	98.5%	0.01
Bayes mod. 1	-	2.549	1.887	26%	0.092	87.1%	0.087
Bayes mod. 2	0.11	1.776	1.442	18.8%	0.067	92.1%	0.059

All methods confirm that a positive strategy can be selected with CSCV being the most optimistic in terms of SR haircut or probability of loss. Figure 19 indicates that, in this case, there is a fairly good coincidence between the ex ante and ex post SR distributions for all the methods with CSCV again looking the best. Bayes model 2 provides a reasonable estimate of the haircut and the probability of loss, but the estimated "*p*-value", i.e., the probability that the selected model is a false discovery is surprisingly high (11%). Nevertheless, it should be noted that, in the MC simulations based on Bayesian posterior parameters, the SR of the strategy 7 might be quite lower than the "observed" value of 40% due to the high return volatility as already mentioned above.

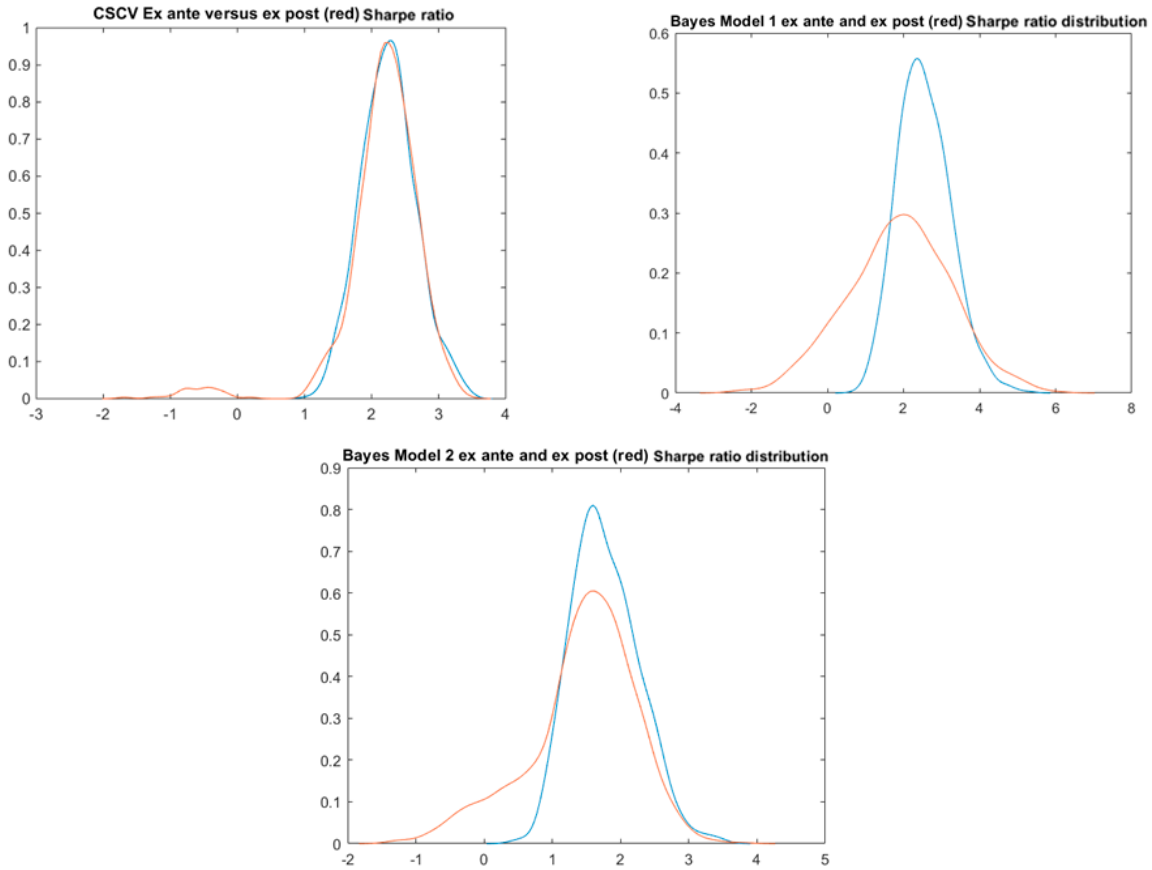

Figure 19. The simulations of the ex ante (blue) and ex post (red) SR probability densities.

Lastly, we modify the returns of the strategies by deducting the observed mean returns (from the daily strategy returns) and adding random noise means with standard deviation 1% p.a. (Figure 20). Therefore, in this case, we expect the methods to reject the existence of a positive strategy.

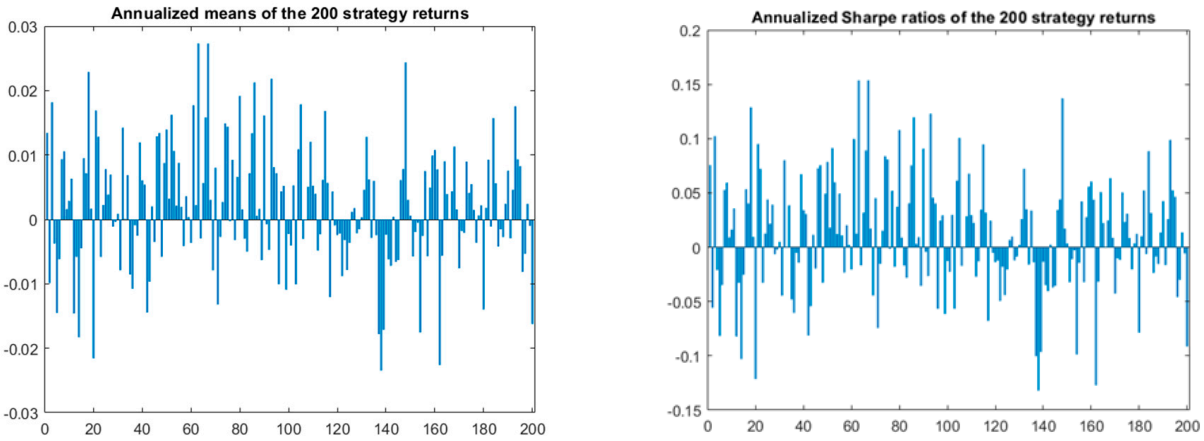

Figure 20. The modified annualized mean returns and Sharpe ratios of the 200 strategies.

All the methods, except for the Bayes Model 1, clearly refute the existence of a positive strategy (Table 3). The surprisingly optimistic results of Bayes Model 1 can be again explained by the volatility incorporated into the Bayes parameter MCMC estimation leading to sampling of models with higher positive means in the MC part of the simulations. The first graph in Figure 21 also clearly demonstrates the strong negative bias of the CSCV method where the best IS model tends to the worst OOS not because of the models but due to the design of the method. See also the IS/OOS scatter plots in Figure 22.

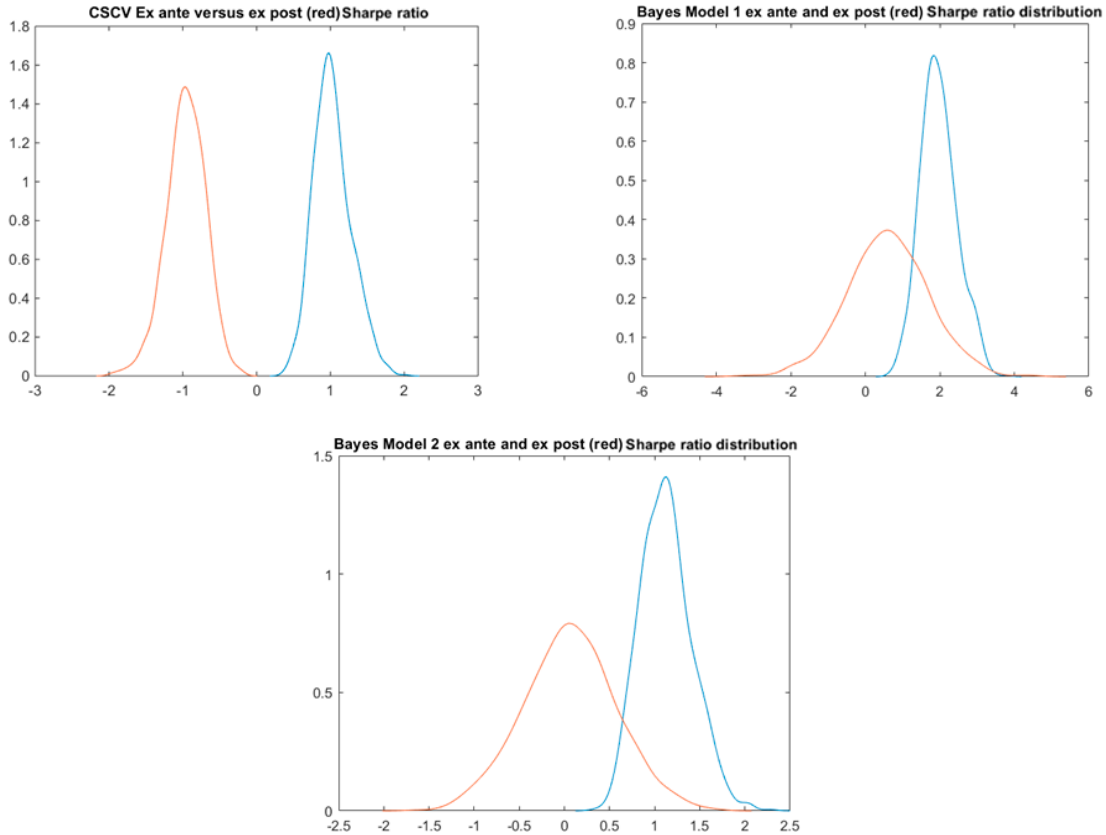

Figure 21. The simulations of the ex ante and ex post SR probability densities.

Table 3. Summary of the test results.

	Adjusted *p*-Value (FDR)	Ex Ante av. SR	Adjusted Expected SR	Hair Cut	Probability of Loss	Mean Rank	PBO
Boferroni method	1	0.1536	0	100%	-	-	
Šidák's correction	1	0.1536	0	100%	-	-	
Multivariate norm. MC adj.	0.997	0.1536	0.002	98.7%	-	-	
CSCV	-	1.027	−0.960	193.5%	1.00	1.2%	1
Bayes mod. 1	-	1.975	0.607	69.3%	0.286	67.1%	0.265
Bayes mod. 2	0.887	1.127	0.062	94.5%	0.447	53.2%	0.450

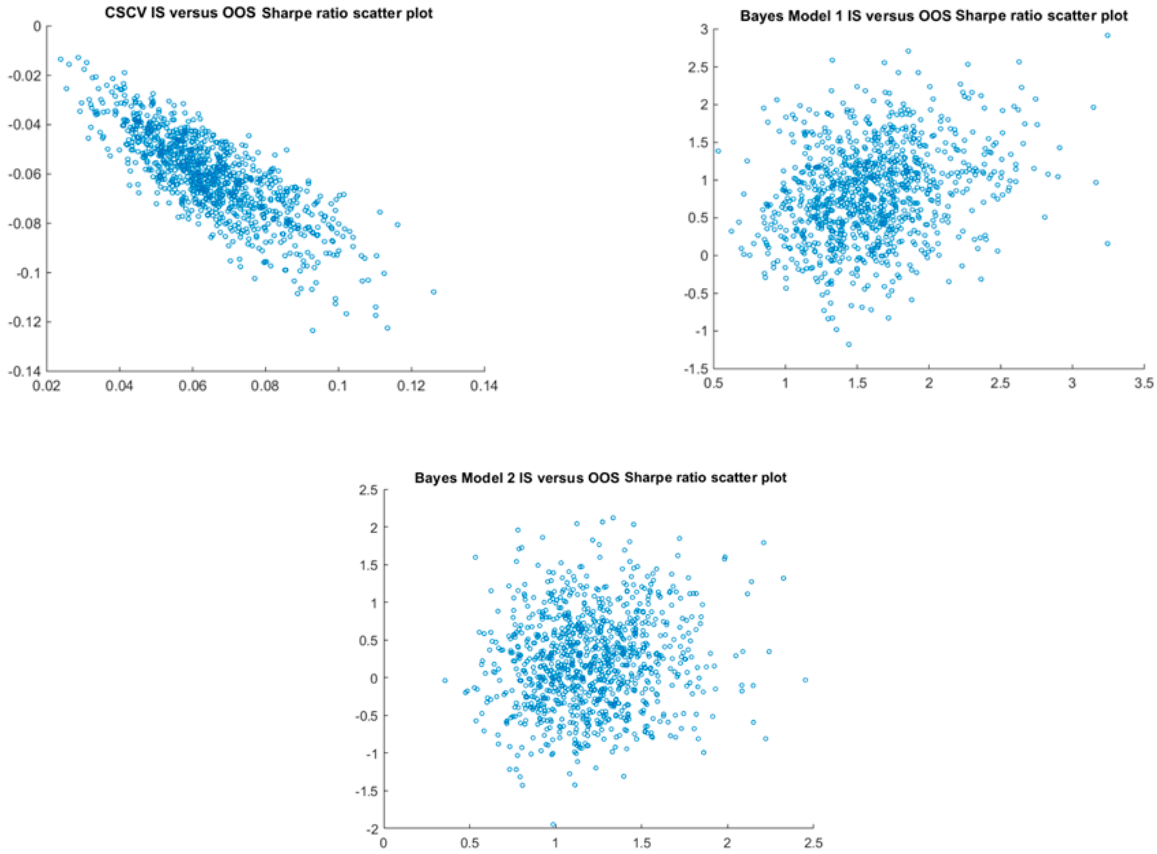

Figure 22. Scatter plots of IS versus OOS Sharpe ratios generated by the three models.

5. Conclusions

The classical methods to adjust single-test *p*-values for the effect of multiple testing when selecting a trading strategy out of many possibilities such as Bonferroni, Holms, or BHY work relatively well, but provide very conservative estimations due to their approximate nature. Certain improvement can be achieved by applying the independence-based multiple-test *p*-value (Šidák's) adjustment or the proposed multivariate normal MC simulation method. The derived expected SR and the related haircut proposed by Harvey and Liu (2015) are rather heuristic and, in our view, not well theoretically founded. The stationary bootstrap method proposed by Sullivan et al. (1999) provides a consistent *p*-value adjustment. However, if used in a two-stage simulation, it may damage functionality of a positive strategy depending on medium/long-term trends due to the mixing bootstrap algorithm. It also turns out to be the most computationally demanding, since all strategies

must be replicated for each sequence of bootstrapped asset prices. Moreover, it cannot be used if the strategies are not known or depend on other economic series. The CSCV method (Bailey et al. 2016) is relatively computationally efficient and provides good results if the mean returns of the strategies are well diversified. However, if the strategies' mean returns are all close to zero, then the method gives negatively biased results. On the other hand, it appears overoptimistic if one strategy stands high above the others.

Lastly, we proposed and investigated two Bayesian methods, the naïve one based on the simple assumption that the returns are multivariate normal, and the second extended with latent variables indicating zero and nonzero mean return strategies. While the naïve model gives mixed results, the second provides, according to our empirical study, the most consistent results, and it is a useful tool to properly analyze the issue of backtest overfitting. In addition to the probability of loss and backtest overfitting (PBO), it estimates the posterior probabilities of whether each individual strategy is a true discovery and, at the same time, the probability of making a true discovery (and the complementary FDR) and selecting the best one. We believe that the proposed method (Bayesian model 2) provides an efficient way to analyze the effect of backtest overfitting, keeping relatively parsimonious assumptions on the underlying data-generating model. More advanced multivariate stochastic models of the underlying returns might be considered in further research.

Funding: This research was supported by the Czech Science Foundation Grant 18-05244S and by the VSE institutional grant IP 100040.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The publicly available data (daily S&P 500 index values and returns for the period 5 June 2009–24 May 2013) were obtained from https://finance.yahoo.com.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

The following are proofs of the formulas used for Model 2 in Section 3:

Proof of Step 2.

$$p(\boldsymbol{\mu}|\boldsymbol{\Sigma},\boldsymbol{\gamma},\text{ data}) \propto \varphi(\Gamma\boldsymbol{\mu};m,A)\varphi(\mu;m_0,D) \propto$$

$$\propto \exp\left(\frac{-1}{2}\left(\boldsymbol{\mu}'\left(\Gamma A^{-1}\Gamma + D^{-1}\right)\boldsymbol{\mu} - 2\boldsymbol{\mu}'\left(\Gamma A^{-1}\boldsymbol{m} + D^{-1}\boldsymbol{m}_0\right)\right)\right)$$

$$\propto \varphi\left(\boldsymbol{\mu};\left(\Gamma A^{-1}\Gamma + \mathrm{D}^{-1}\right)\left(\Gamma \mathrm{A}^{-1}m + D^{-1}\boldsymbol{m}_0\right),\ \left(\Gamma A^{-1}\Gamma + \mathrm{D}^{-1}\right)^{-1}\right).$$

□

Proof of Step 3. Again

$$p(\boldsymbol{\gamma}|\boldsymbol{\Sigma},\boldsymbol{\mu},\text{ data}) \propto \varphi(\Gamma\boldsymbol{\mu};m,A)p(\boldsymbol{\gamma}),$$

where $\Gamma = diag(\boldsymbol{\gamma})$ and $p(\boldsymbol{\gamma}) = \prod_i {p_0}^{\gamma_i}(1-p_0)^{1-\gamma_i}$.

Since we can sample $\gamma_i \in \{0,1\}$ step by step given γ_j, for $j \neq i$, it is enough to calculate

$$p(\gamma_i = 0|\ldots) \propto \varphi(\Gamma_0\boldsymbol{\mu};m,A)p(\boldsymbol{\gamma}) \propto \exp\left(\frac{-1}{2}\left((\Gamma_0\boldsymbol{\mu}-\boldsymbol{m})'A^{-1}(\Gamma_0\boldsymbol{\mu}-\boldsymbol{m})\right)\right)(1-p_0),$$

and similarly for $p(\gamma_i = 1|\ldots)$. We prefer the expression on the right-hand side of the relation above in order to avoid a numerical underflow problem that appears for a higher dimension if the full multivariate density function is used. □

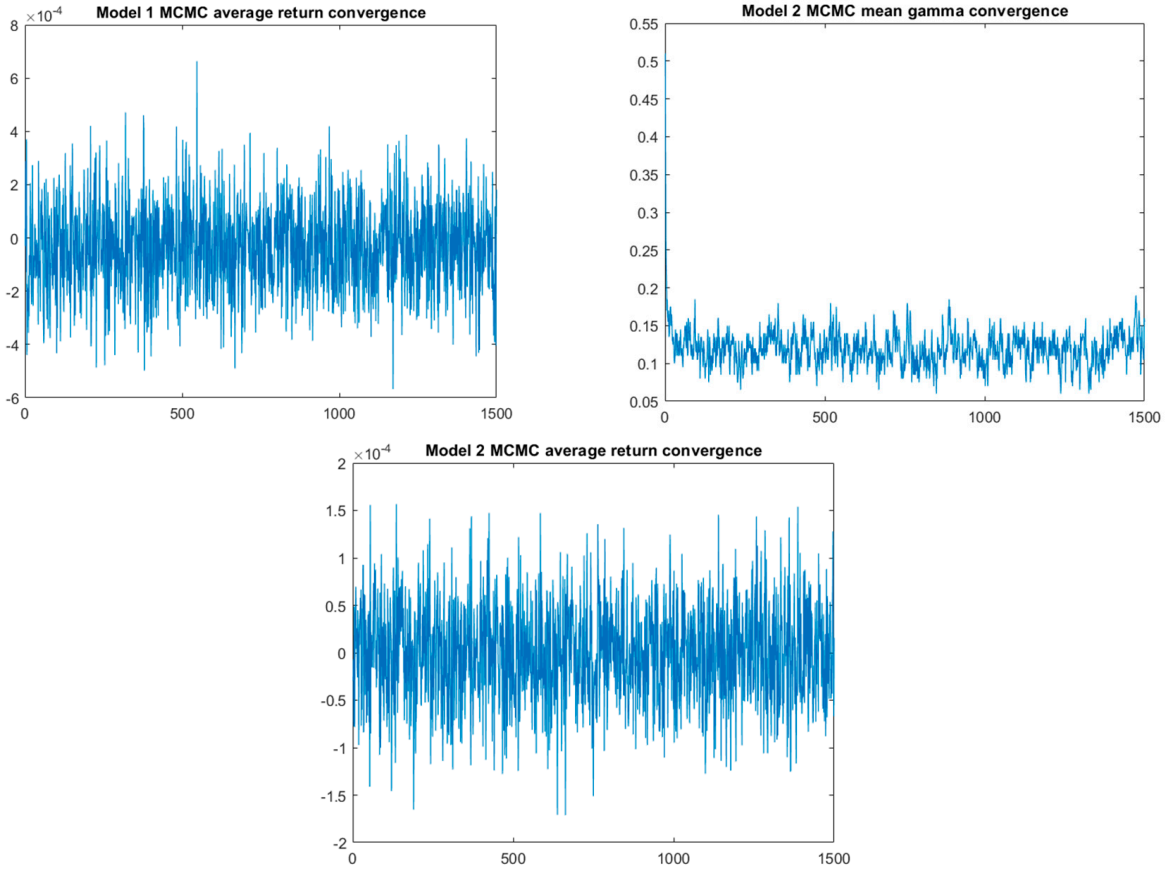

Figure A1. MCMC convergence diagnostics for the Models 1 and 2.

References

Andrews, Isaiah, and Maximilian Kasy. 2019. Identification of and correction for publication bias. *American Economic Review* 109: 2766–94.

Bailey, David H., Jonathan Borwein, Marcos Lopez de Prado, and Qiji Jim Zhu. 2014. Pseudomathematics and financial charlatanism: The effects of backtest over fitting on out-of-sample performance. *Notices of the AMS* 61: 458–71. [CrossRef]

Bailey, David H., Jonathan Borwein, Marcos Lopez de Prado, and Qiji Jim Zhu. 2016. *The Probability of Backtest Overfitting*. Available online: https://ssrn.com/abstract=2840838 (accessed on 21 June 2016).

Chen, Andrew Y., and Tom Zimmermann. 2020. Publication bias and the cross-section of stock returns. *The Review of Asset Pricing Studies* 10: 249–89. [CrossRef]

De Prado, Marcos Lopez. 2015. The future of empirical finance. *The Journal of Portfolio Management* 41: 140–44. [CrossRef]

Harvey, Campbell R., and Yan Liu. 2013. *Multiple Testing in Economics*. Available online: https://ssrn.com/abstract=2358214 (accessed on 28 April 2017).

Harvey, Campbell R., and Yan Liu. 2014. Evaluating trading strategies. *The Journal of Portfolio Management* 40: 108–18. [CrossRef]

Harvey, Campbell R., and Yan Liu. 2015. Backtesting. *The Journal of Portfolio Management* 42: 13–28. [CrossRef]

Harvey, Campbell R., and Yan Liu. 2017. Lucky Factors. Available online: https://ssrn.com/abstract=2528780 (accessed on 28 June 2017). [CrossRef]

Harvey, Campbell R., and Yan Liu. 2020. False (and missed) discoveries in financial economics. *The Journal of Finance* 75: 2503–53. [CrossRef]

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

López de Prado, Marcos, and Michael J. Lewis. 2019. Detection of false investment strategies using unsupervised learning methods. *Quantitative Finance* 19: 1555–65. [CrossRef]

Lynch, Scott M. 2007. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Berlin/Heidelberg: Springer, p. 359.

Politis, Dimitris N., and Joseph P. Romano. 1994. The stationary bootstrap. *Journal of the American Statistical Association* 89: 1303–13.
Scott, James G., and James O. Berger. 2006. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference* 136: 2144–62. [CrossRef]
Streiner, David L., and Geoffrey R. Norman. 2011. Correction for multiple testing: Is there a resolution? *Chest* 140: 16–18. [CrossRef] [PubMed]
Sullivan, Ryan, Allan Timmermann, and Halbert White. 1999. Data-snooping, technical trading rule performance, and the bootstrap. *The Journal of Finance* 54: 1647–91. [CrossRef]
White, Halbert. 2000. A reality check for data snooping. *Econometrica* 68: 1097–126. [CrossRef]