A Bayesian Approach to Backtest Overfitting

Jiří Witzany[1]

Abstract

Quantitative investment strategies are often selected from a broad class of candidate models estimated and tested on historical data. Standard statistical technique to prevent model overfitting such as out-sample back-testing turns out to be unreliable in the situation when selection is based on results of too many models tested on the holdout sample. There is an ongoing discussion how to estimate the probability of back-test overfitting and adjust the expected performance indicators like Sharpe ratio in order to reflect properly the effect of multiple testing. We propose a consistent Bayesian approach that consistently yields the desired robust estimates based on an 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.

JEL Classification: G1, G2, C5, G24, C11, C12, C52.

Keywords: Backtest, multiple testing, bootstrapping, cross-validation, probability of backtest overfitting, investment strategy, optimization, Sharpe ratio, Bayesian probability, MCMC

1. Introduction

According to Prado (2015) empirical finance is in crisis. The former president of the American Finance Association claims (Harvey et al., 2016) that "most claimed research findings in financial economics are likely false." The heart of the problem is systematic or latent multiple testing and datamining. The issues appears mainly in two important areas: testing and selection of factors explaining asset returns (see e.g. Harvey and Liu, 2015 or 2017) and selection of investment strategies (see e.g. White, 2000 or Bailey et al., 2016). 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. The latter approach is more common for theoretical research while the former, easier to control, would be typical for a quantitative investment firm.

We are going to formalize and investigate the problem of strategy selection based on a large set of candidates. Consider self-financing strategies $\mathcal{S}_1, \dots, \mathcal{S}_K$ that are backtested and

[1] University of Economics, Faculty of Finance and Accounting, Department of Banking and Insurance, W. Churchill Sq. 4, 130 67, Prague, Czech Republic, e-mail: jiri.witzany@vse.cz
This Research has been supported by the Czech Science Foundation Grant 15-00036S "Credit Risk Modeling for Financial and Commodity Assets Portfolios".
Working paper: June 2017

evaluated over a historical period with (e.g. daily) returns $r_{k,t}, k = 1, \dots, K, t = 1, \dots, T$. Note that the strategies could have been developed on a preceding training period and backtested or validated on the $\{1, \dots, T\}$ 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 evaluated on the backtest period. Based on the historical data we estimate the (annualized) sample means m_k, standard deviations s_k, or Sharpe ratios SR_k, and given a criterion 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 (see Figure 1)

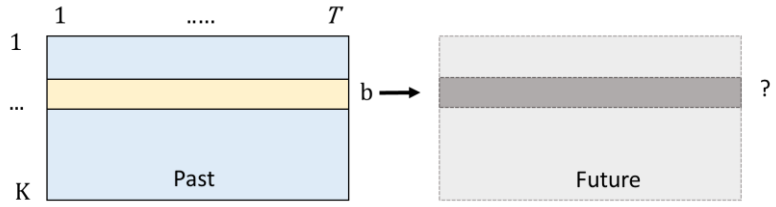


Figure 1. Future performance of the best strategy selected based on the past data

Specifically, the questions usually asked include:

- 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, that is 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 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 (FDR)?

We are going propose a Bayesian methodology that allows us to simulate many times the IS selection and OOS realization process (Figure 1) in order to address the questions formulated above. We will provide an overview of several methods proposed in literature that will be compared with our proposed strategy in an empirical study.

2. An Overview of the Existing Approaches

There are several relatively simple classical methods how to adjust p-values in order to accommodate the multiple test. More advanced and computationally demanding methods are based on various approaches to bootstrapping and simulation of the past and future data.

Classical approaches

To test significance of a single strategy (for example $\mathcal{S}_b$), the classical approach is to calculate the t-ratio

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

and the one-sided or two-sided[2] p-value

$$p^S = \Pr[|X| > TR] \quad (1)$$

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 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.

The problem of 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 in fact should be (Harvey, Liu, 2015) reflecting the fact that we are selecting the strategy with the best t-ratio:

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

where X_k are k random variables following the t-distribution with $T-1$ degrees of freedom. It is noted (Harvey, 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:

$$p^M = 1 - \prod\nolimits_k \Pr[|r_k| \le TR] = 1 - (1-p^S)^K = Kp^S - \binom{K}{2}(p^S)^2 + \cdots$$

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

We are also proposing and will test a numerically relatively simple and efficient method based on an estimation of the covariance matrix Ω of the returns and numerically generating the distribution of $\max\{|X_k|, k = 1, \dots, K\}$ 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) Sharp ratio can be easily calculated given the t-ratio and vice-versa

$$SR = \frac{m_b}{s_b} = \frac{TR}{\sqrt{T}}.$$

The ratio is usually annualized as follows

[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.

$$SR_a = \frac{m_b}{s_b}\sqrt{T_a} = TR\sqrt{\frac{T_a}{T}}$$

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

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

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

Harvey and Liu (2015) note 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 Sharp ratio and the number of strategies. They propose to use the relationship between the single and multiple test p-values in order to estimate the haircut Sharp ratio. Their estimate of the annualized expected Sharp 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.

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

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 by (2). The haircut estimation, of course depends, on the p-value adjustment method as Bonferroni, Holm's, BHY, or the general one we have suggested above. Although the estimation is obviously directionally correct, it is not obvious why this approach should yield a consistent estimate of $E_0[.]$ and the corresponding haircut. We are going to compare the different haircut estimates in the simulation study outlined below.

Stationary Bootstrap

In order to simulate the past and future returns we are going to consider a bootstrapping and a cross-validation approach. The *stationary bootstrap* proposed and analyzed in White (2000), Sullivan et al. (1999) and Politis, 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).

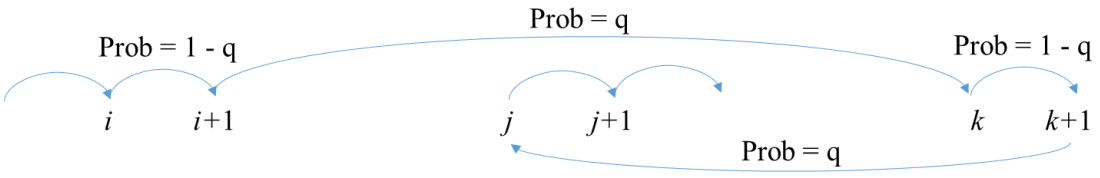


Figure 2. Stationary Bootstrap Process

Formally, we generate new sequences of the underlying asset returns $\{u_{\Theta(i)}; i = 1, .., \tilde{T}\}$ where $u_1, \dots, u_T$ is the original series of returns and $\Theta(i) \in \{1, \dots, T\}$. 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 $\Theta(1) \in \{1, \dots, T\}$, and for $i = 2, \dots, \tilde{T}$ setting $\Theta(i) = \Theta(i-1) + 1$ with probability $1 - q$ or randomly drawing a new block starting position $\Theta(i) \in \{1, \dots, T\}$ with probability q. If it happens that $\Theta(i) > T$ then we draw random $\Theta(i) \in \{1, \dots, T\}$.

Next, given a bootstrapped sequence of the underlying asset returns we need to apply strategies $\mathcal{S}_1, \dots, \mathcal{S}_K$ to get the strategies' bootstrapped returns $\tilde{r}_{k,t}, k = 1, \dots, K, t = 1, \dots, T_2$. Note that since the strategies' decision are often built based on the past we generally need to have a longer series of the bootstrapped asset returns, $\tilde{T} > T_2$. Then we evaluate our desired performance indicator values (mean, Sharp ratio, etc.) $\tilde{f}_k$. Let f_k^* denote the performance indicators of the original series of returns. According to White (2000), under certain mild theoretical assumptions, the bootstrapped values $\tilde{V} = \max_{k=1,\dots,K}(f_k^* - \tilde{f}_k)$ asymptotically converge to the distribution of the best strategy performance indicator under the null hypothesis H_0 that all the strategies have zero performance. I.e., obtaining B bootstrapped values $\{\tilde{V}_j; j = 1, \dots, B\}$ we can test H_0 by calculating the empirical p-value $\Pr[|\tilde{V}| > f_k^*]$.

The bootstrapping 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 $1, \dots, T_1$ selecting the best strategy $\mathcal{S}_b$ with in-sample performance SR_{IS} and then looking on its out-of-sample performance SR_{OOS} over the following period $T_1 + 1, \dots, T_1 + T_2$. Note that the original bootstrapping has to be done over a period of length $\tilde{T} > T_1 + T_2$. We can then 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 $\mathcal{S}_b$ (based on the real dataset). However, particularly in this case, it is 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 large extent mixed up. Therefore, the estimated conditional SR_{OOS} may easily lead to a false rejection of a good strategy.

Combinatorial Symmetric Cross-Validation

A 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 of the *combinatorial symmetric cross-validation* (CSCV) *(*Bailley et al., 2014, 2016) utilizing only the matrix of the strategies' returns $M = \{r_{k,t}, k = 1, \dots, K, t = 1, \dots, T\}$. The idea is to split the time window of length $T = SN$ into S blocks of length N where S is even and draw combinations of $S/2$ blocks (Figure 3). The submatrix J formed by joining $T/2$ 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 $K \times T/2$ 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 form sufficiently many different combinations with replacement as long as S is sufficiently large, e.g. at least 16.

Figure 3. An example of the CSCV combination for $S = 6$

Bailley et al. (2014, 2016) propose to use the technique to estimate specifically 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 $\mathcal{S}_1, \dots, \mathcal{S}_K$,

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

The PBO indicator as well as the Sharp ratio haircut can be estimated using sufficiently many 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 $\{1, \dots, T\}$ 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 be probably the worst on $\bar{J}$. We will demonstrate the effect in the empirical part. The cross-validation technique also cannot be used, due to this property, to estimate the OOS Sharp ratio or mean for a particular selected strategy. We can just estimate the overall PBO or Sharp 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 (or equivalently FWER if the best IS strategy is automatically declared to be a discovery) since it is not clear how to identify true and false discoveries given a CSCV simulation. This could be possibly done by testing significance of OOS performance involving an ad hoc probability level. We are going to show that all the indicators of interest can be consistently estimated in a Bayesian set up we are going to outline below.

3. Bayesian Simulation Approach

The Bayesian approach will be based on the following the scheme given in Figure 4.

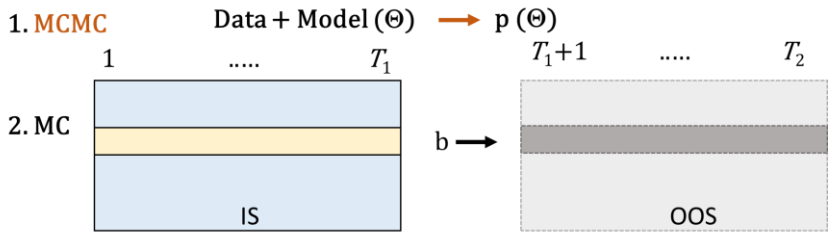


Figure 4. Two step Bayesian simulation (MCMC parameter estimation and MC data simulation)

First, a model defining the return generating process with unknown parameters Θ for the observed strategy returns $\{r_{k,t}, k = 1, \dots, K, t = 1, \dots, T\}$ needs to be specified. Then the plan is to use a Bayesian technique, in particular the Markov Chain Monte Carlo (MCMC) simulation in order to extract the posterior distribution of the model parameters Θ. Finally, simulate matrices of IS and OOS returns over desired time intervals $1, \dots, T_1$ and $T_1 + 1, \dots, T_1 + T_2$. 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 case of the stationary bootstrap.

We are going to 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 estimate consistently the false discovery rate (FDR) etc. The second model follows an idea of Scott and Berger (2006), also mentioned in Harvey (2016), nevertheless, in both cases the model is 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 are going to consider below.

The Naïve Model 1

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

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

and that the observations over time are independent.

Given data $= \langle \boldsymbol{r}_t \rangle$, 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 to sample from the distributions and the posterior distribution may be obtained quite efficiently (e.g. 10 000 runs of the sampler).

Remark: The sampler above assumes the non-informative 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 back tested returns with cross sections from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Based on 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 sample faithfully the empirical distribution of the performance indicators:

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

The simulated posterior distribution of the desired performance indicators then tells us what are the mean, median, confidence intervals, or Bayesian probabilities that the true performance is positive or above any given minimum threshold. The ratio between the ex post and ex ante performance indicators also give us an estimate of the "backtest overfitting haircut."

Model 2 – Bimodal Means Distribution

In order to capture the situation when most strategies are random and only some positive (non-zero) 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 being 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. $p(\boldsymbol{\Sigma}|\boldsymbol{\mu}, \boldsymbol{\gamma}, \text{data}) = IW(\boldsymbol{\Sigma}; T, S)$, 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 $A = \frac{1}{T}\boldsymbol{\Sigma}$, $\Gamma = diag(\boldsymbol{\gamma})$, $\boldsymbol{m} = \frac{1}{T}\sum_{t=1}^{t} \boldsymbol{r}_t$, $\boldsymbol{m}_0 = [m_0, \dots, m_0]$, $D = diag([V_0, \dots, V_0])$, where $diag$ creates a matrix with diagonal elements given by the vector in the argument, and sample $p(\boldsymbol{\mu}|\boldsymbol{\Sigma}, \boldsymbol{\gamma}, \text{data}) = \varphi(\boldsymbol{\mu}; (\Gamma A^{-1}\Gamma + \mathrm{D}^{-1})(\Gamma \mathrm{A}^{-1} m + D^{-1}\boldsymbol{m}_0), (\Gamma A^{-1}\Gamma + \mathrm{D}^{-1})^{-1})$.

3. Given $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$, estimate $\boldsymbol{\gamma}$. For $i = 1, \dots, K$ 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

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

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

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

Proof of step 2:

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

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

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

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

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

and similarly for $p(\gamma_i = 1| \dots)$. 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. □

Remark: 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 $(V_0) \propto IG\left(V_0, \frac{k_0}{2}, \frac{k_0 V_P}{2}\right)$.
   For $\widetilde{K} = \#\{\mu_i|\gamma_i = 1\} \neq 0$ set $\tilde{\mu} = \sum\{\mu_i|\gamma_i = 1\}/\widetilde{K}$ and $\tilde{V} = \sum\{(\mu_i - m_0)^2|\gamma_i = 1\}/\widetilde{K}$. Then
   $p(m_0|\boldsymbol{\mu}, \boldsymbol{\gamma}, \mathrm{V_0}) \propto \varphi(m_0; \tilde{\mu}, V_0/\widetilde{K})\,\varphi(m_0; 0, V_P) \propto \varphi\left(m_0; \frac{\tilde{\mu}\widetilde{K}V_P}{\widetilde{K}V_P + V_0}, \frac{V_P V_0}{\widetilde{K}V_P + V_0}\right)$ and
   $p(V_0|\boldsymbol{\mu}, \boldsymbol{\gamma}, \mathrm{m_0}) \propto IG\left(0; \frac{\widetilde{K}}{2}, \frac{\widetilde{K}\tilde{V}}{2}\right) IG\left(V_0, \frac{k_0}{2}, \frac{k_0 V_P}{2}\right) \propto IG\left(V_0, \frac{\widetilde{K}+k_0+1}{2}, \frac{\widetilde{K}\tilde{V}+k_0 V_P}{2}\right)$.
   If $\widetilde{K} = 0$ then we have to sample based on the conjugate priors $p(m_0)$ and $p(V_0)$ only.

4. Numerical Study

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

Technical Strategies Selection

We have used 1000 daily S&P 500 values and returns for the period 5.6.2009 – 24.5.2013. The period has been selected with the purpose to find at least one strategy with a higher mean return. As in Sullivan et al. (1999) we have applied the filter, moving average, support and resistance rules with varying parameters. We have selected randomly 200 strategies with the condition that the daily profit series are not collinear (it may even happen that the series are identical if the parameters do not differ too much).

The means and Sharp ratios of the individual strategies and their densities 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 the strategy 7 is apparently the best with annualized ($n_y = 252$) mean return over 21% p.a. and Sharp ratio approximately 1.2 (it is a filter strategy with x=y=1%, i.e. long or short position is taken if the previous daily return is over 1% or below -1%, respectively, minimum number of days to stay in a position = 20). The strategy returns look attractive, nevertheless, looking forward, it turns out that the mean return of the strategy in the following 1000 days period is negative (-5.21%).

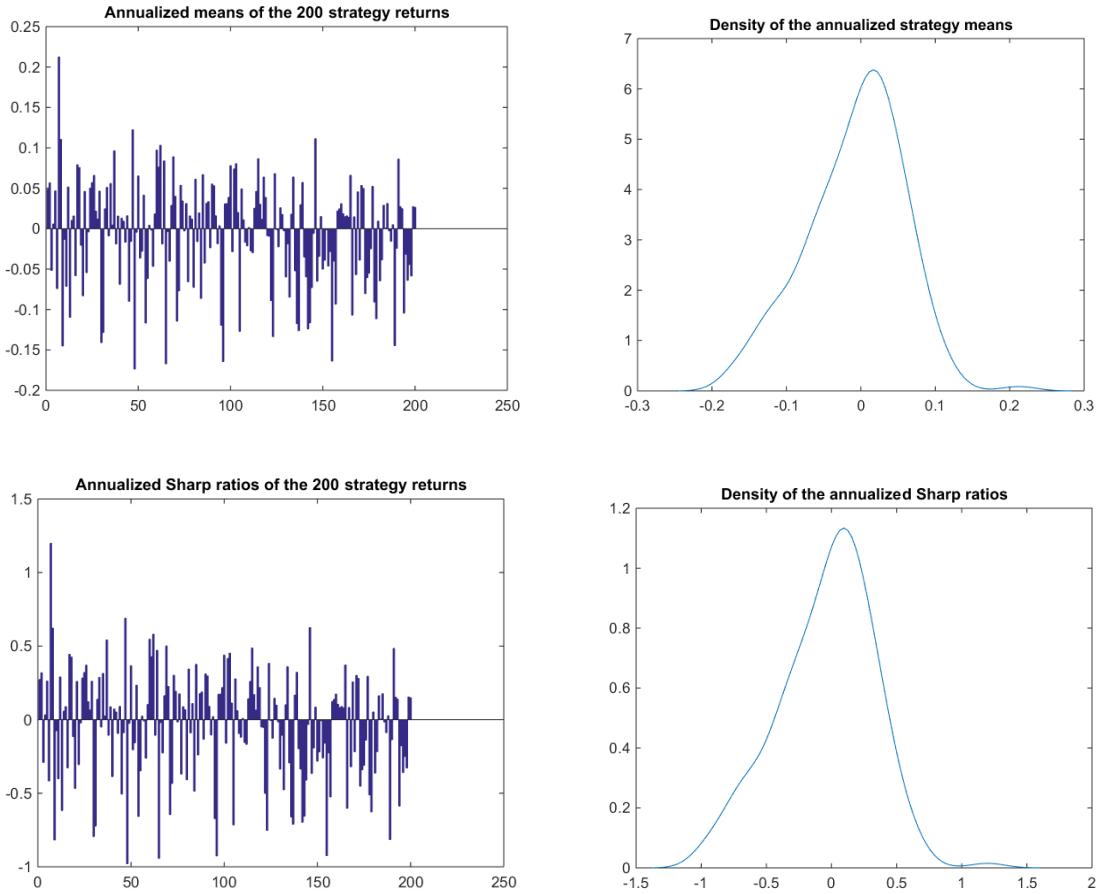


Figure 5. Annualized means and Sharp ratios of the selected strategies

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

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

The multiple test p-value after Bonferroni adjustment is simply $p^M = \min\{200 \times 0.0.171{,}1\} = 1$, and so the adjusted expected Sharp ratio is 0 and the haircut 100%. Šidák's adjustment yields only slightly more optimistic result with $p^M = 0.9685$, adjusted expected $SR = 0.02$ and the 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, Liu (2015). The simple adjustment methods allow to estimate easily 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%.

Multivariate Normal Simulation and the Stationary Bootstrap Based on the Null Hypothesis

Another relatively simple possibility is to estimate the return covariance matrix and simulate the future multivariate returns based the return covariance matrix and conditional on zero means. The figure below shows the density of 1000 simulated annualized SR based on 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%).

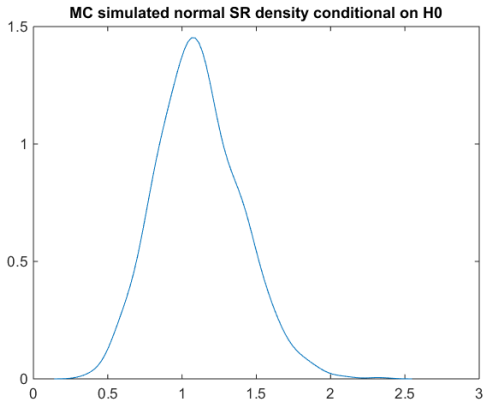


Figure 6. Sharp ratio density based on the null hypothesis and the multivariate normal MC simulation

Analogous distribution can be obtained by the much more computationally demanding stationary bootstrap (White´s reality check). The p-value based on 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 at 0.175, and the SR haircut 85.4%.

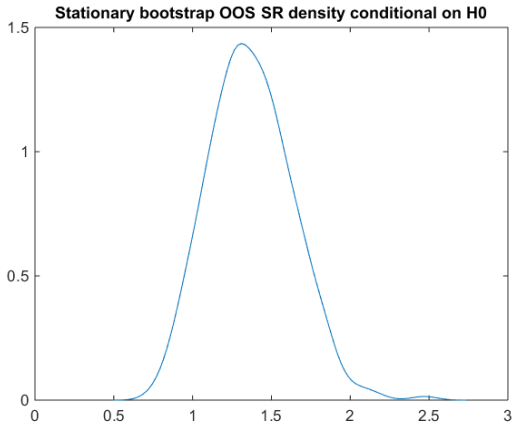


Figure 7. Sharp ratio density based on the White´s reality check

Stationary Bootstrap two-stage simulation

The stationary bootstrap method can be also used to simulate the backtest period of length $T_1 = 1000$ as well as the future period with $T_2 = 1000$. Number of stationary bootstrap iterations will be again 1000 based on the 5.6.2009 – 24.5.2013 window of S&P returns, and the parameter is set to $q = 0.1$ with the corresponding average length of the bootstrapped blocks 10. The results however 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 the SR haircut over 73%. For detailed results including the ex ante and ex post SR or mean return values see the summary Table 1. Note that the row "Stationary bootstrap" shows values obtained by the two-stage simulation except the p-value estimated by the White's reality check. Figure 9 shows the typical strong shift of the ex ante performance density to the left hand side and the wider ex post performance density.

	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	-	1.771 / 0.314	1.142 / 0.203	35.5% / 35.4%	0.052	89.2%	0.036
Bayes mod. 2	0.589	1.213 / 0.215	0.212 / 0.038	82.5% / 82.5%	0.397	61%	0.36

Table 1. Summary of the backtest overfitting tests

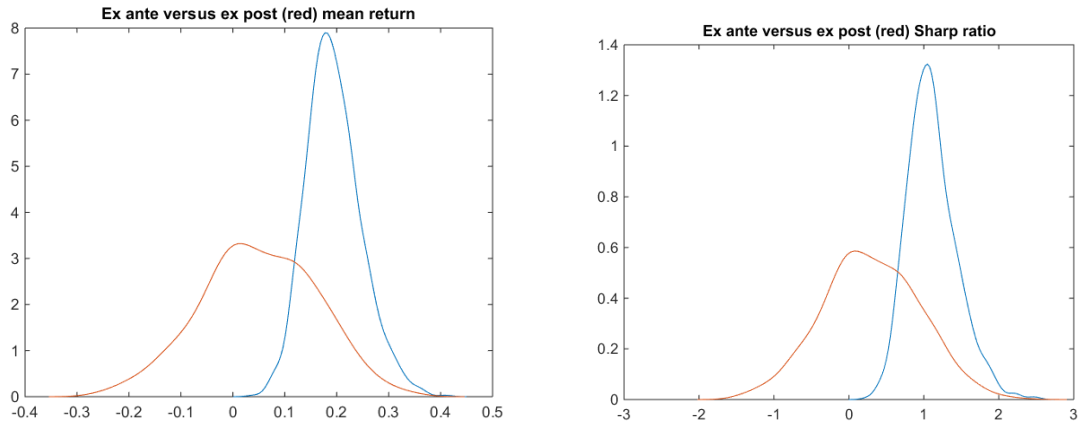


Figure 8. Sharp ratio and the mean return ex ante and ex post densities

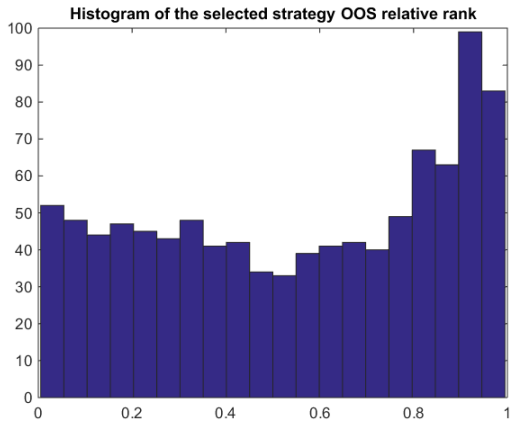


Figure 9. Histogram of the out-of-sample relative rank of the best IS strategy

Combinatorial Symmetric Cross-Validation

In order implement the CSCV algorithm we have chosen the number of blocks 20 corresponding up to $\binom{20}{10} = 184756$ combinations of 10 blocks of length 50. However, we have 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 similar to the Stationary bootstrap only with PBO being slightly lower 0.323. Figure 11 indicates that in case, compared to Figure 9, the best IS strategy remains the best OOS quite often. This is also 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.

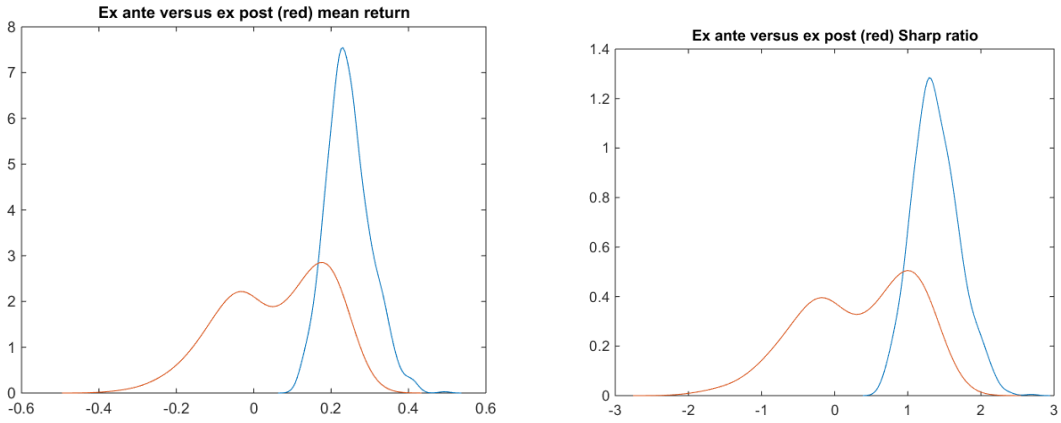


Figure 10. CSCV simulation of the ex-ante and ex post densities

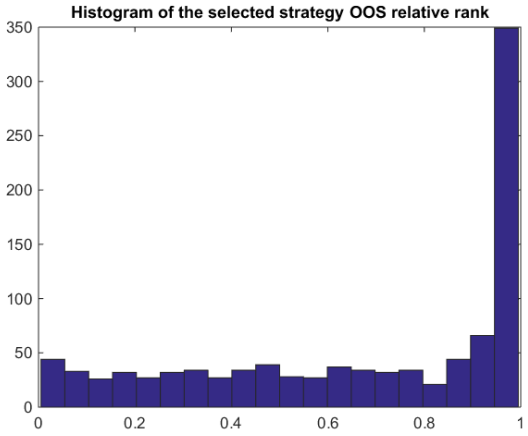


Figure 11. CSCV histogram of the out-of-sample relative rank of the best IS strategy

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 1000 iterations using of the standard Gibbs sampler. 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 have then generated 1000 scenarios sampling the parameters, 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 so the observed best strategy 7 may look weak in the simulations while other strategies are selected as the best. However, the best IS strategy will remain the best quiet often as shown in Figure 13 and the low PBO = 0.036. This can be explained by the simple multivariate normal model and relatively long IS window allowing to identify the truly positive strategy. The detailed results given in Table 1 and Figure 12 confirm that the naïve multivariate normal model indeed appears too optimistic.

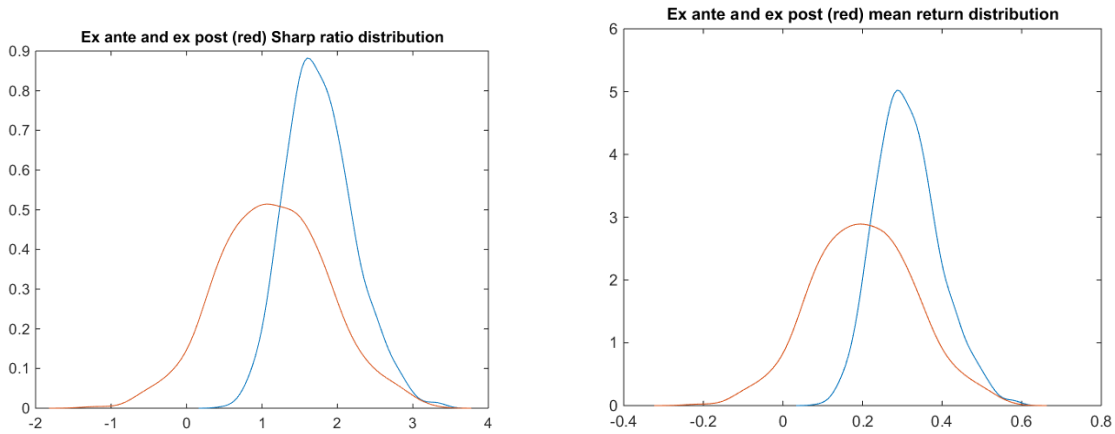


Figure 12. Bayes model 1 simulation of the ex-ante and ex post densities

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

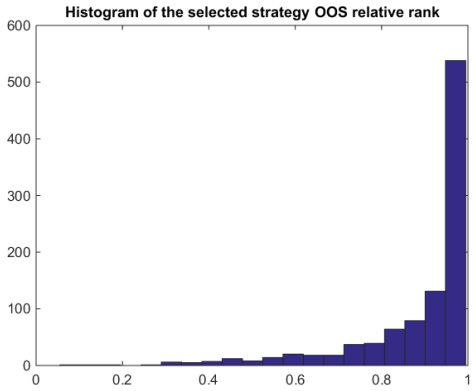


Figure 13. Bayes model 1 histogram of the out-of-sample relative rank of the best IS strategy

In reality, if we generate a large number of models and the best model performance is still poor, e.g. negative or close to zero, then, probably, we are 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 91.3% of the simulations with average ex ante SR 1.84 and ex post SR 1.18, i.e. again with the haircut slightly over 35%. 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 set the hurdle to 2 then the conditional probability of loss would decline to 1.4% (conditional on model selection) and the unconditional to 0.4% since 71.5% of the proposed models would be rejected in the simulation.

Bayes Bimodal Mean Returns Model 2

In this case, besides the multivariate normal distribution with unknown parameters, we also consider latent indicators of zero and nonzero model. In order to extract the posterior distribution of the parameters and the latent indicators we gain run 1000 iterations of the Gibbs sampler outlined in Section 3 and as well as 1000 the MC 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 0.36, much lower compared to model 1, the SR haircut 82.5% or the probability of loss 39.7% (see Table 1, Figure 14, and Figure 15).

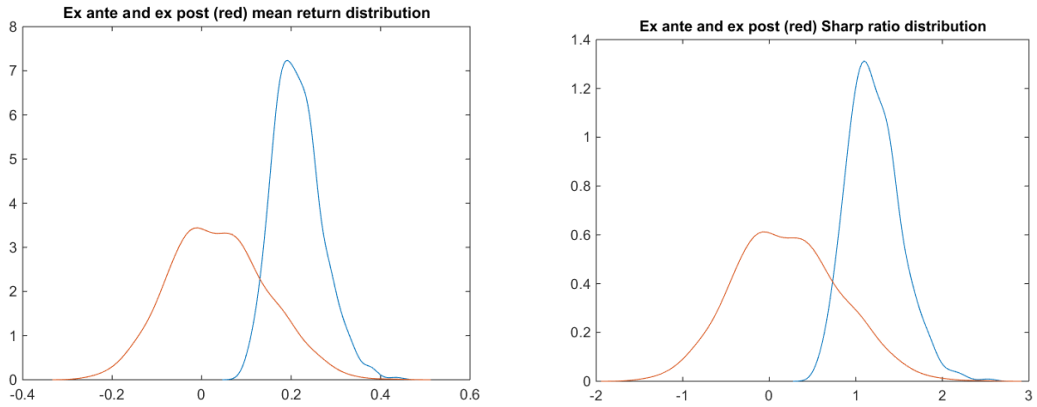


Figure 14. Bayes model 2 simulation of the ex-ante and ex post densities

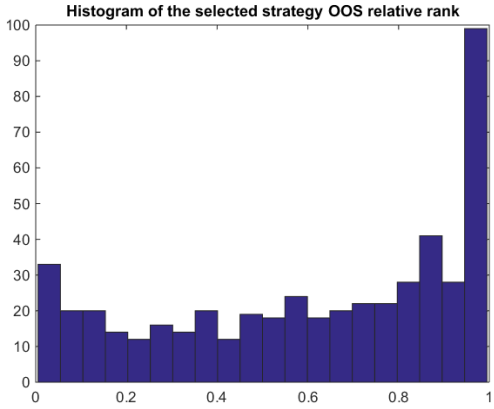


Figure 15. Bayes model 2 histogram of the out-of-sample relative rank of the best IS strategy

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 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: giving the observed data and the general model assumptions what is the probability that the best model b selected based on the 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 0.419. It means that, applying the selection process, only in 41.9% of cases we identify the true discovery and in 58.1% we make a false discovery, i.e. $FDR = 58.1\%$ shown instead the adjusted p-value in Table 1.

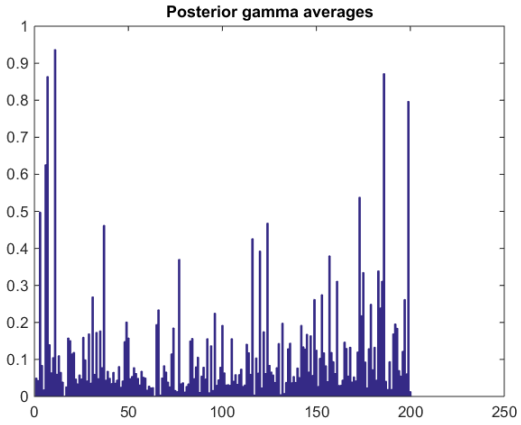


Figure 16. Posterior gamma average values for the 200 strategies

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 turns to stay around 82% more-or-less independently on the hurdle. The unconditional probability of loss can be reduced only slightly, e.g. for the hurdle of 1.5, the conditional probability of loss declines to 37%, but the unconditional goes significantly down to 6.1% as 83.5% of best models 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. 10, 20,…,200 (with $N = 500, T_1 = 1000,\ T_2 = 1000$). Note that the best observed strategy is include in the first ten but, as expected, the posterior expected gamma, mean, or SR do decline with the number of strategies tested (Figure 17).

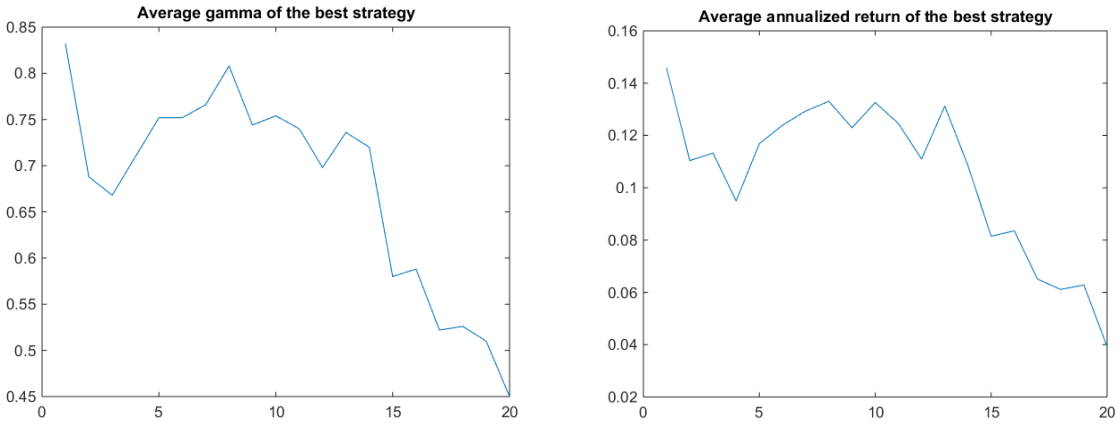


Figure 17. Estimated values the average ex post gamma, mean return and SR depending on the number of strategies tested

Testing with Modified Mean Returns

In order to better compare the methods we are going to modify the vector of returns of the strategies while keeping the "natural" correlation structure. Firstly, we increase the strategy 7 return by 19% p.s. 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 by the various methods.

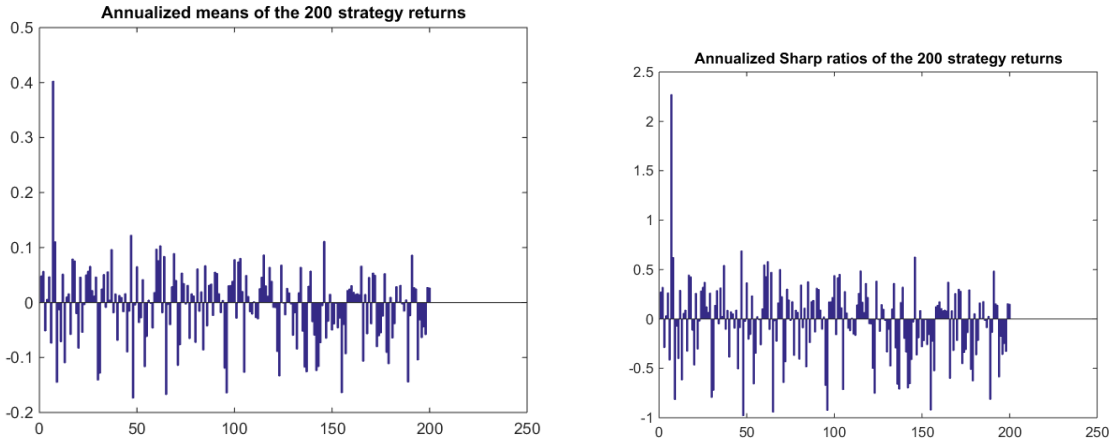


Figure 18. The modified annualized mean returns and Sharp ratios

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

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. The 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 true discovery is surprisingly low 88.7%.

Nevertheless, it should be noted that in the MC simulations based on o 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.

	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	-	2.354	2.087	11.3%	0.014	96.4%	0.011
Bayes mod. 2	0.887	1.752	1.439	17.8%	0.067	91.8%	0.067

Table 2. Summary of the tests' results

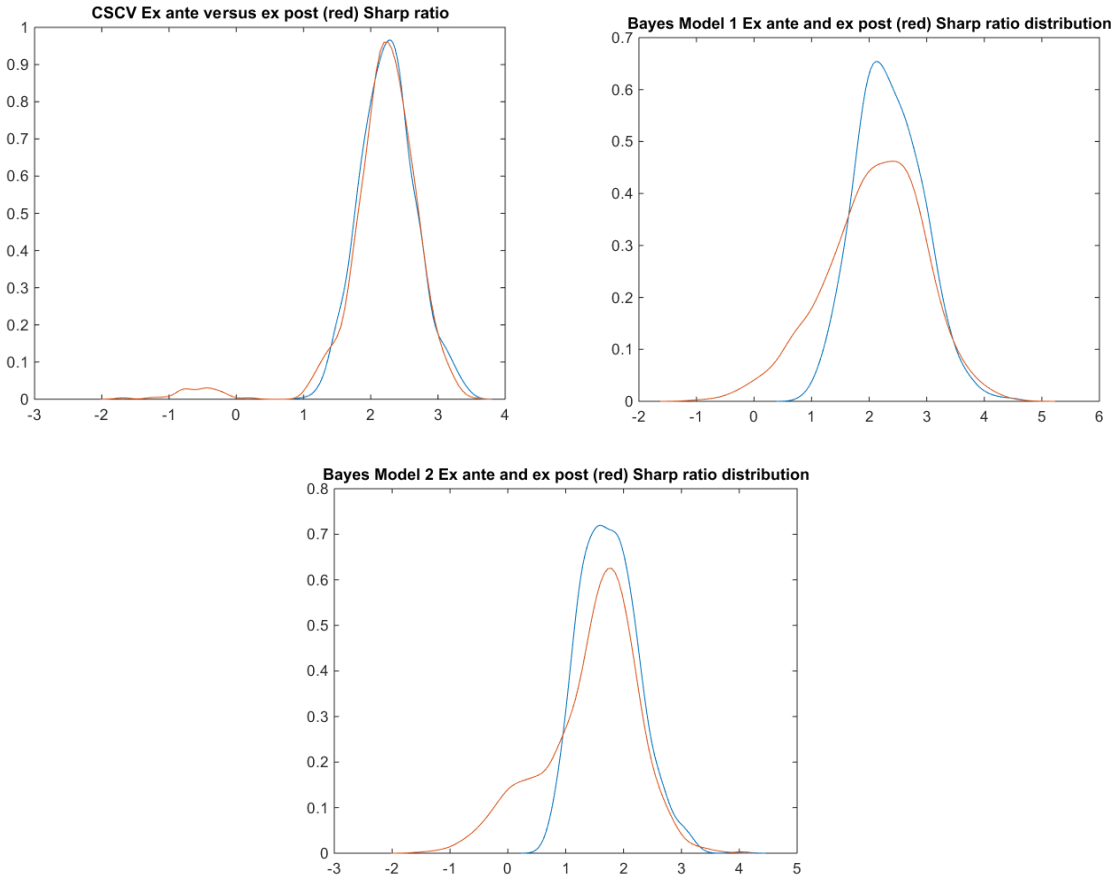


Figure 19. The simulations of the ex-ante and ex post SR densities

Finally, we will 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 that the methods to reject existence of a positive strategy.

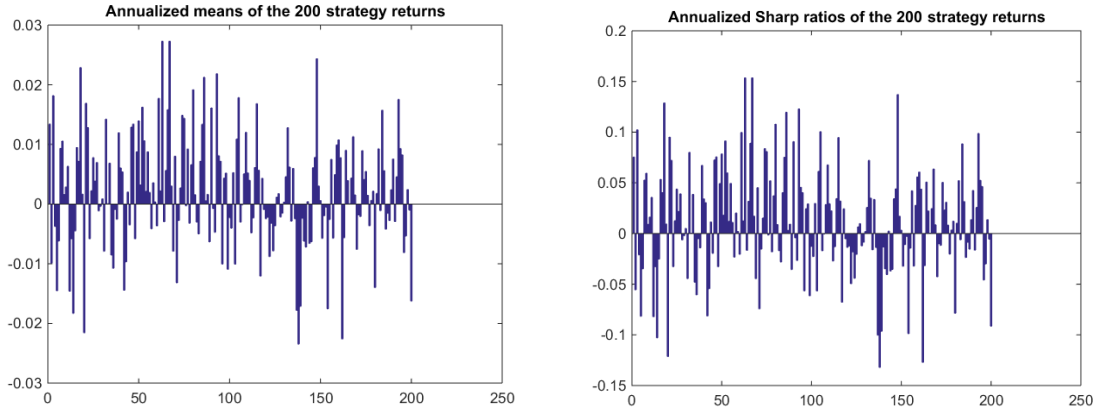
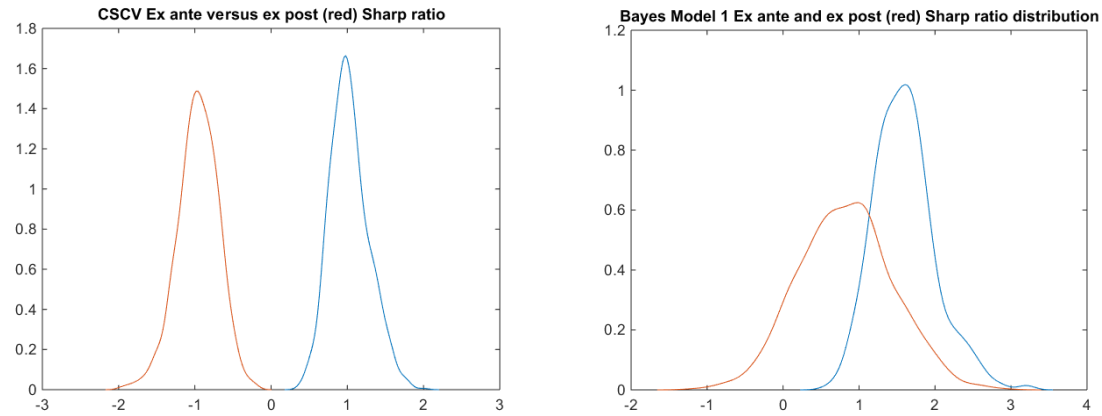


Figure 20. The modified annualized mean returns and Sharp ratios

	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	-	1.596	0.825	48.3%	0.084	80.5%	0.121
Bayes mod. 2	0.314	1.139	0.057	95%	0.452	53.5%	0.465

Table 3. Summary of the tests' results



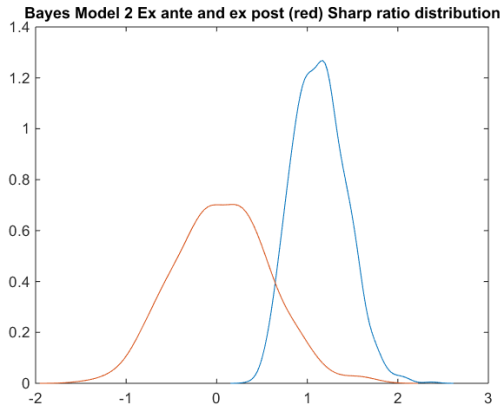


Figure 21. The simulations of the ex-ante and ex post SR densities

All the methods, with the exception of the Bayes Model 2, clearly refute existence of a positive strategy (Table 3). The surprisingly optimistic results of the 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 IS/OOS the scatter plots in Figure 22.

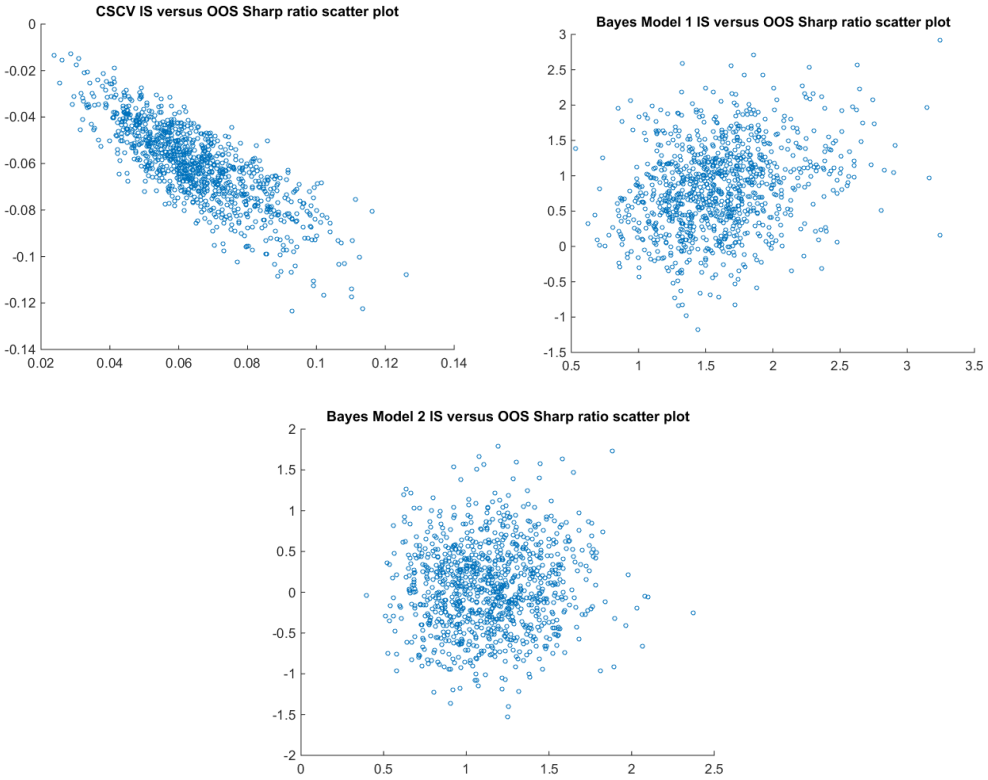


Figure 22. Scatter plots of IS versus OOS Sharp ratios generated by the three models

5. Conclusions

The classical methods to adjust single test p-value for the effect of multiple testing when selecting a trading strategy out of many possibilities like Bonferroni, Holms, or BHY work relatively well but provide very conservative estimations due to their approximate nature. Certain improvement can be achieved 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, Liu (2015) are rather heuristic and in our view not 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 computational the most demanding since all strategies must be replicated for each sequence of bootstrapped asset prices. Moreover, it cannot be used of the strategies are not known or depend on other economic series. The CSCV method (Bailey et al., 2016) is relatively computational 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. Finally, we have 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 in our view the most consistent results and a useful tool to analyze properly the issue of backtest overfitting. Besides the probability of loss and backtest overfitting (PBO) it estimates the posterior probabilities whether each individual model is a true discovery and at the same time the probability making a true discovery (and the complementary FDR). It should be emphasized that the goal of the investigated methods is to analyze the effect of the backtest overfitting keeping relatively parsimonious assumptions on the underlying data generating model.

Literature

Bailey, D. H., Borwein, J. M., Lopez de Prado, M., & Zhu, Q. J. (2016). The probability of backtest overfitting.

Bailey, D. H., Borwein, J. M., de Prado, M. L., & Zhu, Q. J. (2014). Pseudomathematics and financial charlatanism: The effects of backtest over fitting on out-of-sample performance. Notices of the AMS, 61(5), 458-471.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. The Review of Financial Studies, 29(1), 5-68.

Harvey, C. R., & Liu, Y. (2014). Evaluating trading strategies. The Journal of Portfolio Management, 40(5), 108-118.

Harvey, C. R., & Liu, Y. (2015). Backtesting. The Journal of Portfolio Management, 42(1), 13-28.

Harvey, Campbell R. and Liu, Yan, Lucky Factors (2017). Available at SSRN: https://ssrn.com/abstract=2528780 or http://dx.doi.org/10.2139/ssrn.2528780 , June 1

Harvey, Campbell R. and Liu, Yan, Multiple Testing in Economics (2013). Available at SSRN: https://ssrn.com/abstract=2358214

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. Journal of the American Statistical association, 89(428), 1303-1313.

De Prado, M. L. (2015). The future of empirical finance. The Journal of Portfolio Management, 41(4), 140-144.

Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. Journal of statistical planning and inference, 136(7), 2144-2162.

Sullivan, R., Timmermann, A., & White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap. The journal of Finance, 54(5), 1647-1691.

White, H. (2000). A reality check for data snooping. Econometrica, 68(5), 1097-1126