How hard is it to pick the right model? 
MCS and backtest overfitting

Diego Aparicioa,∗ and Marcos López de Pradob,1

aDepartment of Economics, Massachusetts Institute of Technology, Cambridge, MA, USA 
bTrue Positive Technologies, New York, NY, USA

Abstract. Recent advances in machine learning, artificial intelligence, and the availability of billions of high frequency data signals have made model selection a challenging and pressing need. However, most of the model selection methods available in modern finance are subject to backtest overfitting. This is the probability that one will select a financial strategy that outperforms during backtest, but underperforms in practice. We evaluate the performance of the novel model confidence set (MCS) introduced in Hansen et al. (2011a) in a simple machine learning trading strategy problem. We find that MCS is not robust to multiple testing and that it requires a very high signal-to-noise ratio to be utilizable. More generally, we raise awareness on the limitations of model selection in finance.

Keywords: Forecasting, model confidence set, machine learning, model selection, multiple testing

JEL Codes: G17, C52, C53

1. Introduction

With recent advances in machine learning, parallel computing, and large historical millisecond-based financial datasets, it is not rare for industry engineers to backtest hundreds or thousands different investment strategies in order to search for the most profitable model.2 Likewise, the availability of unprecedented quantities of individual-level data also means that A/B experimentation and data-driven designs are becoming the gold standard in online platforms, retail, technology companies, medicine, and even policy (Lazer et al. (2009), Kohavi et al. (2007), Bakshy et al. (2014), Varian (2014), Bastani & Bayati (2015), Athey (2017), Lada et al. (2018)). But as we test an increasing number of strategies and predictive features, or repeat the same experiment many times, it becomes more likely that some of the estimated effects will be extraordinarily effective. How, then, should we evaluate and select the right models? And in which situations does it matter? The multiple testing problem is now more pervasive and affects both practitioners and academics alike. And therefore it is important that analysts use tougher standards to test their models in a robust and unbiased way.

This paper evaluates the performance of the ’model confidence set’ (MCS) introduced in Hansen et al. (2011a). The MCS procedure, described in Section 2, starts with a collection of models, and sequentially prunes the worst performing models one by one, according to some user-defined loss function,
until the first non-rejection takes place. These surviving models, found to be statistically similar, define the estimated model confidence set \( \hat{M}^\ast \). MCS presents many appealing features relative to other techniques. It allows the user to reduce the baseline set of models to a smaller set (model confidence set); the confidence set need not be just one model; the user can define the meaning of best (i.e. the loss function); it avoids \( p \)-values concerns from multiple pairwise comparisons; and it also avoids the somewhat arbitrary decision to choose a benchmark model against which all models are evaluated. Hansen et al. (2011a) thus present a substantial contribution to the long-standing discussion on model selection, which, in our view, is nowadays becoming more and more pervasive. While we favor these advances, however, we find that the properties assumed in MCS are not well-suited to aid the model selection problem in modern finance.

In the introduction to their article, for instance, the authors suggest that MCS can be used to select ’treatment effects’ or ’trading rules with the best Sharpe ratio’ (p. 454). Our simulations suggest that the coverage properties in MCS are not adequate to winnow out trading strategies in practice. Analysts may use MCS for initial screening or forecasting combination, but not as sufficient evidence to select investment strategies. Similarly, academics should not rely solely on MCS as sufficient evidence to defend a given macroeconomic or forecasting model. We hope that our discussions here raise awareness of the limitations of similar model selection methods, but also of the need for further research in this area.

This paper relates to an extensive literature on model selection and forecast evaluation in economics (Corradi & Distaso (2011), Elliott & Timmermann (2016), Clark & McCracken (2013)). More generally, our work relates to a deeper discussion of the implications and challenges of data-driven model selection (Leeb & Potscher (2005)). Concerns about false discoveries due to \( p \)-hacking, or data snooping, are not limited to finance, but arguably affect all observational or experimental studies (Ioannidis (2005), John et al. (2012), Simonsohn et al. (2014)). A growing variety of methods address the multiple testing problem (White (2000), Benjamini & Hochberg (1995), Benjamini & Yekutieli (2001), Storey (2002), Romano & Wolf (2005), Romano et al. (2010); see also Bailey et al. (2014) and Harvey et al. (2016) for recent methodologies in finance).

The rest of the paper is structured as follows. Section 2 describes the MCS and its limitations. Section 3 presents simulation results from the perspective of selecting financial strategies. Section 4 concludes.

2. MCS

We begin this Section with a brief overview of the model confidence set (MCS) from Hansen et al. (2011a), and then introduce the main limitations of applying it to a forecasting problem. We encourage the reader to see Hansen et al. (2011a), Hansen et al. (2011b), and Hansen et al. (2014) for additional details on the methodology. We stress that our discussions here should not be understood as a naive critique of the MCS. MCS presents a substantial contribution to the model selection problem; we find, however, that the requirements in MCS are not adequate to many of the modern model-selection problems faced in practice.

MCS starts with a collection of models \( \mathcal{M}^0 \) predefined by the user. Without loss of generality, consider the hypothetical case of a manager whose problem is to decide the best investment strategy. In deciding which strategy to invest in, the manager might simulate and analyze the backtesting performance of several hundred, possibly thousands, of different strategies. In fact, recent advances in multi-processing computing and big data allow us to very easily simulate thousands of investment strategies at the same time, all of them using an arsenal of millisecond-long transactions, while fine-tuning the best feature set combination. MCS provides a framework that facilitates the model-selection problem. In particular, MCS yields a model confidence set, \( \hat{M}^\ast_{1-\alpha} \), that contains a (possibly smaller) set of the best models with a given level of confidence. That is, \( \lim_{n \to \infty} P(\mathcal{M}^\ast \subset \hat{M}^\ast_{1-\alpha}) \geq 1 - \alpha \). \( \hat{M}^\ast \) can potentially be equal to just one strategy, but could also contain all the initial models if these are found to be statistically similar.

Following the notation in Hansen et al. (2011a), MCS is based on an equivalence test, \( \delta_{\mathcal{M}} \), and an elimination rule, \( e_{\mathcal{M}} \). The algorithm can be described as follows.

- Initially set \( \mathcal{M} = \mathcal{M}^0 \), where \( \mathcal{M}^0 \) contains a finite number of models indexed by \( i = 1, \cdots, m_0 \). These objects will be evaluated according to a user-defined loss function \( L_{i,t} = L(Y_t, \hat{Y}_{i,t}) \). For instance, \( L_{i,t} = (\pi_t - \hat{\pi}_{i,t})^2 \) could be defined as the squared error from the actual inflation \( \pi_t \) and the forecast \( \hat{\pi}_{i,t} \) from model \( i \).
Test the hypothesis $H_{0,M}: \mu_{ij} = 0$ at level $\alpha$, for all $i, j \in M$, where $\mu_{i,j} \equiv E(d_{ij,t})$, and $d_{ij,t} \equiv L_{i,t} - L_{j,t}$ denotes the relative performance.

- If $H_{0,M}$ is accepted, then $\hat{M}^*_{\alpha} = M$. Otherwise use the elimination rule, $e_M$, to drop the worst model from $M$ and repeat the procedure. Different test statistics are proposed in Hansen et al. (2011a). In the case of the $T_{Range}$ statistic, for instance, the worst model is such that $e_M \equiv \arg\max_{i \in M} \frac{d_i}{\sqrt{\text{var}(d_i)}}$.

When the procedure ends, MCS yields $\hat{M}^*_{\alpha}$, or the set of 'surviving' models, in the sense that there is no object whose relative performance is found to be significantly inferior to the other elements.

Although MCS is easy to compute (there are several statistical software packages available) and has many attractive features, we find that its use is limited in practice. The methodology requires the true superior models to have an unrealistically high signal-to-noise ratio. The low power of the test is in part due to not defining a benchmark. In Section 3, for example, we show that a superior model would need to have an annualized Sharpe ratio greater than 7 to be picked up as the single model in $\hat{M}^*$. Practitioners are not likely to face such profitable strategies, and if they do, those strategies may be highly overfitted. Moreover, MCS does not fully penalize the test with the number of trials in the experiment.

Model selection criteria that do not severely penalize for multiple testing tend to select models that have experienced a high backtesting performance when, in reality, they are of the same quality as many others with a poorer performance. The problem is exacerbated with large $N$ trials, similarly to testing individual coefficients in a regression. If there are dozens of coefficients, on average there will be a few that appear strongly significant. If we run hundreds of trading strategies, some of them will yield extraordinarily large Sharpe ratios and MCS will select them.3

3. Simulation exercise

This Section presents simulation results in a financial engineering problem. However, the results are relevant to a wider range of model selection applications. Data scientists are regularly testing usage time or conversion rates under different features via A/B experimentation in retail, online platforms, and mobile apps (Kohavi et al. (2007), Aparicio & Prelec (2017); prediction methodologies to improve decision-making are also becoming popular in policy Athey (2017)).

We take the stand of a hedge fund manager who has to choose between different investment strategies. A manager will typically simulate $M$ different strategies, each of them generated using different features, data signals, and machine learning methods, and potentially choose those with the highest backtesting performance. We simulate $M$ series of financial returns as follows.4,5

1. Let $M$ be the number of models (strategies) to be simulated. Assume each model $m$ generates $T$ returns according to a random walk with drift. We also assume that daily returns experience a Poisson jump-diffusion process, similar to Merton (1976). When this event takes place, returns jump upwards or downwards an amount equal to ten times the (daily) volatility. In discrete form,

$$\hat{r}_{m,t} \sim N(\mu', \sigma'), \, m \in M, \, t \in T$$

$$r_{m,t} = \hat{r}_{m,t} + b_{m,t}(\lambda)(10*\sigma')$$

where $b = \{-1, +1\}$ with equal probability, and takes place according to a Poisson process with occurrence rate $\lambda = 3\%$. We vary $T$ from 1 to 3 years of daily returns; and $M$ from 10 to 100. Returns in equation (1) are generated to have mean annual return $\mu = 10\%$ and annual return volatility, $\sigma$, from 3% to 30%; e.g. $\mu' = \frac{\mu}{T}$ and $\sigma' = \frac{\sigma}{\sqrt{T}}$ when $T = 1$ year (250 trading days).

2. We introduce one true superior strategy, which is defined as having a (‘multiplier’) times higher expected returns. That is, using the notation from equation (1), $E(\hat{r}_{1,t}) = a * \mu'$. We let $a$ fixed within each simulation, but vary $a$ from 1 to 20 across different specifications. When $a = 1$ all models are equally good.

3. In order to evaluate the performance of a strategy, we define the loss function as the excess returns over the expected returns:

4The data generating process (DGP) in the following simulations is a simplified yet standard assumption in the literature (e.g., Harvey & Liu (2014)).

5The Python codes to reproduce the results are available on the authors’ webpage.
$L_{m,t} = - (r_{m,t} - \bar{r})$

Where $\bar{r} = \frac{1}{M} \left( \sum_{t=1}^{T} r_{m,t} \right)$ is the estimated average daily return across all models. Therefore models with high returns have lower errors.

4. Finally, for each Monte Carlo simulation and parameter combination, we apply the MCS procedure and analyze the in-sample as well as the out-of-sample performance of the selected and excluded models. As defined earlier, $\mathcal{M}_{1-\alpha}^*$ corresponds to the set of surviving models that are equally good in a statistical sense at level $\alpha$. The following results correspond to $\alpha = 10\%$, statistical level $\alpha = 10\%$, and 400 repetitions per simulation (for a total of about 2 million MCS simulations).

3.1. Results

To first illustrate the lack of power or signal-to-noise ratio problem in the MCS procedure, we narrow the simulations to the case where $M = 50, 100$ and $T = 250$ (about a year of daily trading data). Figure 1 shows the number of selected models in $\mathcal{M}^*$ as a function of the in-sample Sharpe ratio. The Sharpe ratio is calculated as $SR = \frac{\hat{\mu}}{\hat{\sigma}}$, where $\hat{\mu}$ and $\hat{\sigma}$ denote the estimated mean return and standard deviation, respectively, during the in-sample period. Throughout the paper we follow Lo (2002) to annualize Sharpe ratios.\(^6\) We first show results for the MCS specification using the $T_{Range,M}$ test statistic, a moving-block bootstrap of length $\ell = 5$, and $B = 500$ bootstrap samples. Results are similar under alternative specifications of the $T_{Range,M}$ statistic. However, we find somewhat inconsistent results using the $T_{max,M}$ test statistic. See Section 3.3.

We find that the superior model needs to have a Sharpe ratio greater than 7 to be picked up as the sole best model in $\mathcal{M}^*$.\(^7\) In some sense, this is to be expected because MCS does not require a benchmark model (contrary to, e.g., White (2000) and Romano & Wolf (2005)) and thus there is greater uncertainty that exacerbates the need for a high signal-to-noise ratio. In contexts of uncertainty over many models, it is plausible that MCS can provide an interesting strategy to create pooled forecasts based on (possibly many) MCS selected strategies. Even simple forecast combination schemes are hard to outperform in the forecasting literature (Faust et al. (2013), Aparicio & Bertolotto (2016), and references therein).

Figure 1 also suggests that the MCS’s threshold Sharpe ratios uniformly penalize for the number of trials. This concern can be related to a growing literature on the false discovery rate (FDR) or family-wise error rate (FWER). The most common example is that of using individual $t$-tests in multiple testing. Suppose that we backtest $N$ independent investment strategies and find that the most profitable one has a Sharpe ratio that is highly significant at the $1\%$ level. Even for small $N$, such as $N = 25$, the implied probability of observing such $t$-statistic is high: $p = Pr(max SR_i \geq \hat{r}) = 1 - (1 - \hat{p})^N = 22\%$. Several methods have been proposed to account for the FDR or FWER: Bonferroni’s adjusted $p$-values, Holm’s step-down $p$-values (Holm (1979)), White’s reality check (White (2000)), FDR-based tests (Benjamini & Hochberg (1995), Storey (2002)). See also Romano & Wolf (2005) and Romano et al. (2008). Analysts should consider applying tougher adjusted $p$-values into the MCS test to further strengthen their estimated model confidence sets.

This concern is relevant here because the multiple testing problem is particularly worrisome in

\(^6\)The DGP assumes $M$ independent strategies, although we note that in practice some will tend to be correlated. Correlated returns would reduce the variance of $d_{ij} \equiv (L_{i,t} - L_{j,t})$, and therefore reduce the sample size required in MCS to identify the superior model.

\(^7\)Such Sharpe ratios are rarely seen in practice. As a reference, the S&P 500 Sharpe ratio is estimated at 0.38 during 1996–2014; even the best-performing hedge funds typically have average Sharpe ratios below 2 (Titman & Tiu (2010), Getmansky et al. (2015)).
finance (Barras et al. (2010), Bailey et al. (2014), Lo (2002)). Hedge funds managers can be tempted to backtest hundreds of trading strategies, and then present to their clients those with the highest performance. By selecting investment \( i \), where \( i = \arg\max_m \{SR_m\} \), one might end up picking the one with the highest backtesting overfitting probability (which is therefore likely to underperform out-of-sample).\(^8\)

Figure 2 generalizes the results from Fig. 1 using all parameter specifications. In particular, the 3D-surface shows the percentage of models included in \( \hat{M}^* \) as a function of the number of models \( M \) and the Sharpe ratio of the superior model. In all cases, we use \( T = 250 \) in-sample returns observations, and limit the Sharpe ratio to 10 for better visualization. We find that, for a given in-sample Sharpe ratio, the percentage of models in \( \hat{M}^* \) is very similar across \( M \).

3.2. Out-of-sample performance

Finally, we illustrate what we observe out-of-sample when we use the MCS algorithm to select financial strategies. We restrict the data to the case where \( T = 250 \) in-sample observations, \( T = 125 \) out-of-sample observations, \( M = 100 \) initial models, one superior strategy with \( a = 10 \), and \( \mu = 10\% \) and \( \sigma = 9\% \) (results are similar under alternative specifications). We first compare the in-sample and out-of-sample performance of the MCS selected models, \( \hat{M}^* \). And we then evaluate the out-of-sample performance of both MCS selected and excluded models, that is \( \hat{M}^* \) and \( \hat{M}^C*, \) respectively.

Figure 3 shows that the out-of-sample performance of the selected models in \( \hat{M}^* \) is significantly worse than their corresponding in-sample performance. This behavior is suggestive of backtest overfitting. In fact, Fig. 4 shows that the out-of-sample mean returns of the selected strategies, \( \hat{M}^* \), is no better than that of the eliminated strategies, \( \hat{M}^C* \). Figures exclude the true superior model for better visual comparison. Not fully penalizing for the multiplicity of trials leads us to more easily select strategies that, out of too many equally good models, were just lucky instances during backtesting.

3.3. Alternative specifications

We now discuss robustness results from two alternative specifications. First, we extend the simulation to select financial strategies based on a collection of Sharpe ratios. In particular, we follow the steps from Section 3 and simulate three years of daily returns. For each strategy, we compute twelve annualized Sharpe ratios based on their quarterly performance (similar results are obtained using monthly or bi-monthly SRs). We then compute the number of MCS selected models as a function of the superior model’s
Fig. 3. The out-of-sample performance of the selected models in $M^*$ is significantly worse than their corresponding in-sample performance. This figure shows the histogram of in-sample and out-of-sample Sharpe ratios of the MCS selected strategies (probability density function).

Fig. 4. The out-of-sample mean returns of the selected strategies, $\hat{M}^*$, are no better than that of the eliminated strategies, $\hat{MC}^*$. This figure shows the histogram of mean out-of-sample returns of the MCS selected and excluded strategies (probability density function).

Sharpe ratio, its return multiplier $a$ (relative to a baseline 10% annual return), as well as the out-of-sample performance of the selected (and excluded) trading strategies. In this case, the loss function is computed for each strategy-period SR as opposed to daily returns. The results, shown in Figs. 5–7, are similar to those in the previous Section. We find that MCS will select strategies with backtest overfitting and that the selected and excluded strategies perform equally good out-of-sample. Figure 9 (Appendix) also shows that MCS excludes a large fraction of models even when all strategies are equally good ($a = 1$).

Finally, we discuss the case when MCS is used with the $T_{max,M}$ statistic instead of $T_{Range,M}$. In the case of the test statistic $T_{max,M}$ the elimination rule is $e_{max,M} = \arg\max_{i \in M} t_i$. Where $t_i = \frac{\bar{d}_i}{\sqrt{\text{var}(\bar{d}_i)}}$, $\bar{d}_i = m^{-1} \sum_{j \in M} \bar{d}_{ij}$, and $\bar{d}_{ij} = n^{-1} \sum_{t=1}^{n} d_{ij,t}$ measures the relative sample loss between $i$ and $j$ models. We find that MCS is very sensitive to the choice between $T_{max,M}$ and $T_{Range,M}$. In particular, the former yields conservative model confidence sets, and in fact MCS will not pick up the right model for any reasonable Sharpe ratio. For in-sample Sharpe ratios below 10, MCS will select almost all models regardless the number of starting models $m_0$. The results are shown in Figs. 8 and 10 (Appendix).\(^{10}\)

\(^{10}\)Consistent with these results, it has come to our attention that the $T_{max,M}$ statistic, and therefore the elimination rule $e_{max,M}$, is not recommended in practice. See Corrigendum (Hansen et al. (2014)).
Fig. 6. Figures 6 and 7 show that the MCS selected strategies are subject to backtest overfitting, i.e. the strategies experience a larger in-sample Sharpe ratio although they are equally good out-of-sample. This figure shows the histogram of mean in-sample and out-of-sample Sharpe ratios of the MCS selected strategies (probability density function). Results are qualitatively similar to those in Section 3.

Fig. 7. This figure shows the histogram of the mean out-of-sample Sharpe ratios of the MCS selected and excluded strategies (probability density function).

4. Conclusions

Traditional testing and evaluation methods need to be reconsidered in light of recent advances in big data and technology. Portfolio managers, for instance, can now generate thousands of different trading strategies at little computational cost, and then present those with the highest backtesting performance to their investors. Similarly, data scientists in industry can design experiments to test each of the new features, and even repeat these experiments many times. In finance, this means that machine learning strategies will be subject to backtest overfitting: we will tend to select strategies that, out of so many, just happened to experience high backtesting performance. We therefore need new tools that can severely penalize for the multiplicity of trials but remain powerful enough to be utilized in practice.

We test the performance of the model confidence set introduced in Hansen et al. (2011a) using a variety of financial strategies simulated from the perspective of a portfolio manager. We find that MCS is not adequate to solve an analyst’s model selection problem, and more generally we hope that our work raises awareness of the challenges of model selection in modern finance.

Acknowledgments

We are grateful to Mike Lock, Isaiah Andrews, Valentina Corradi, Peter Hansen, Anna Mikusheva, Michael Lewis, and Kevin Sheppard for helpful comments. All views expressed in this paper are those of the authors, and do not necessarily reflect those of True Positive Technologies. All errors are our own.
Appendix Figures

Fig. 9. Generalizes the results on the collection of quarterly Sharpe ratios using all parameter specifications. This figure shows the number of models in $\hat{M}^*$ as a function of the number of models $M$ and the return multiplier, $a$, of the superior model.

Fig. 10. We simulate trading strategies following Section 3 and use the $T_{\text{max}, M}$ statistic instead of $T_{\text{Range}, M}$. See Section 3.3 for additional details. Figure 10 shows the probability that $\hat{M}^*$ is exactly equal to the true superior model $M^*$ as a function of its estimated in-sample Sharpe ratio. A cubic spline is added as visual guide.