How Minimum Performance Thresholds Bias Backtests

Bayesian Estimation for Sharpe Ratios Under Selection Bias

Joseph Mulligan
Imperial College London
Qube Research & Technologies

LOW Mathematical Finance Workshop Jan 2025

IMPERIAL QRT

The Manager's Dilemma

- Manager's need to assess the strategies proposed to them for allocation.
- The Sharpe ratio is by far the most popular metric for measuring risk-adjusted returns [Ame+08].
- But how do we estimate the Sharpe ratio for a strategy?

The Manager's Dilemma

- Manager's need to assess the strategies proposed to them for allocation.
- The Sharpe ratio is by far the most popular metric for measuring risk-adjusted returns [Ame+08].
- But how do we estimate the Sharpe ratio for a strategy?

Naive Approach: Blindly trust in-sample Sharpe ratios.

The Manager's Dilemma

- Manager's need to assess the strategies proposed to them for allocation.
- The Sharpe ratio is by far the most popular metric for measuring risk-adjusted returns [Ame+08].
- But how do we estimate the Sharpe ratio for a strategy?

Naive Approach: Blindly trust in-sample Sharpe ratios.
Slightly Better Approach: Apply a flat haircut to all Sharpe ratios.

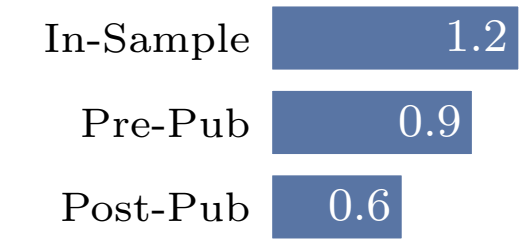
The Manager's Dilemma

- Manager's need to assess the strategies proposed to them for allocation.
- The Sharpe ratio is by far the most popular metric for measuring risk-adjusted returns [Ame+08].
- But how do we estimate the Sharpe ratio for a strategy?

Naive Approach: Blindly trust in-sample Sharpe ratios.
Slightly Better Approach: Apply a flat haircut to all Sharpe ratios.
Much Better Approach: Use a model to estimate out-of-sample Sharpe ratios.

Why Backtests Disappoint

Why can we not simply take backtests at face value?



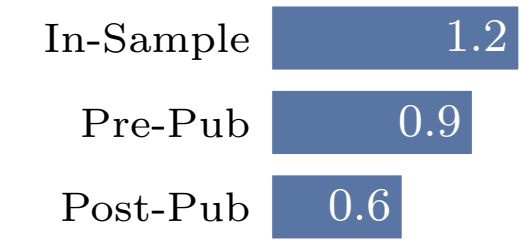
Dataset: 206 predictors from [CZ22].

CAPM β: In-Sample 1929-1968, Pre-Pub 1968-1973, Post-Pub 1973-2024.

Why Backtests Disappoint

Why can we not simply take backtests at face value?

- Transaction costs,



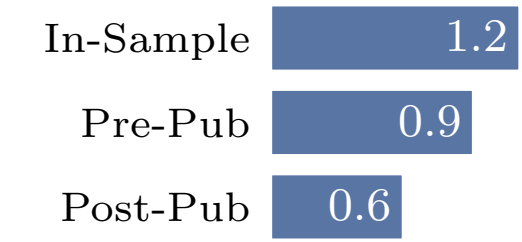
Dataset: 206 predictors from [CZ22].

CAPM β: In-Sample 1929-1968, Pre-Pub 1968-1973, Post-Pub 1973-2024.

Why Backtests Disappoint

Why can we not simply take backtests at face value?

- Transaction costs,
- Alpha decay,



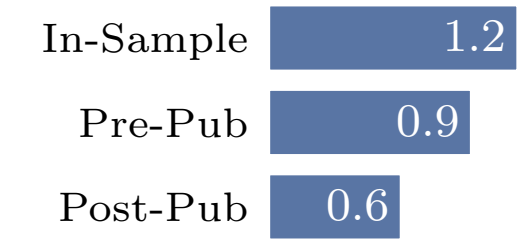
Dataset: 206 predictors from [CZ22].

CAPM β: In-Sample 1929-1968, Pre-Pub 1968-1973, Post-Pub 1973-2024.

Why Backtests Disappoint

Why can we not simply take backtests at face value?

- Transaction costs,
- Alpha decay,
- **Statistical bias in estimation.**



Dataset: 206 predictors from [CZ22].

CAPM β: In-Sample 1929-1968, Pre-Pub 1968-1973, Post-Pub 1973-2024.

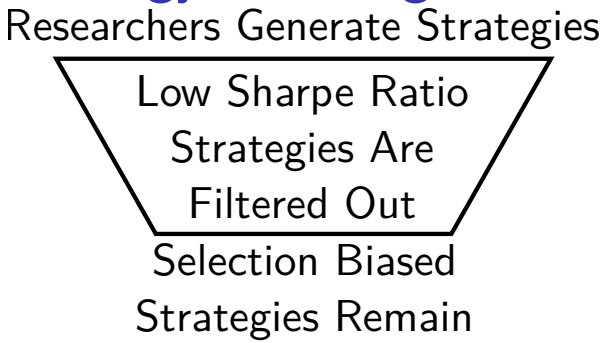
Selection Bias

- How can we end up with bias in our backtests?

Selection Bias

- How can we end up with bias in our backtests?

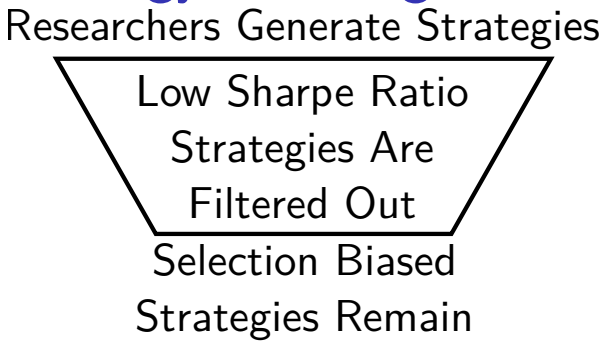
Strategy Filtering



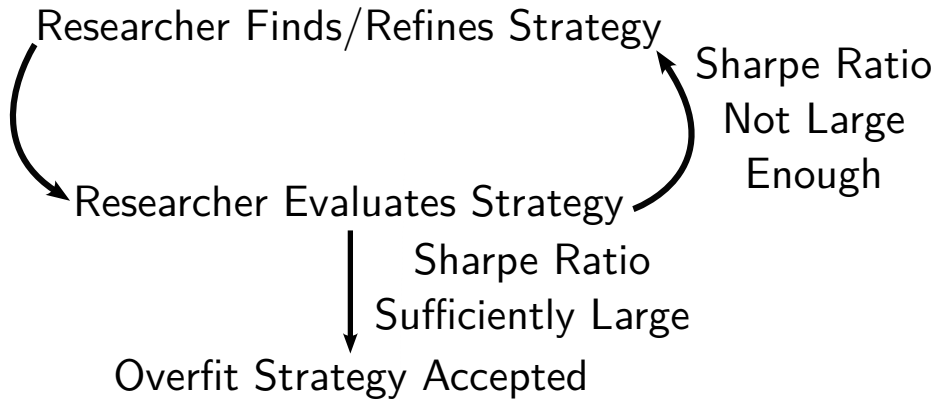
Selection Bias

- How can we end up with bias in our backtests?

Strategy Filtering

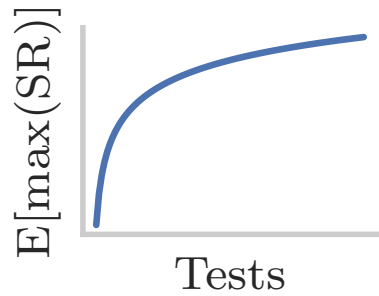


Strategy Iteration



The Issue With Multiple Testing

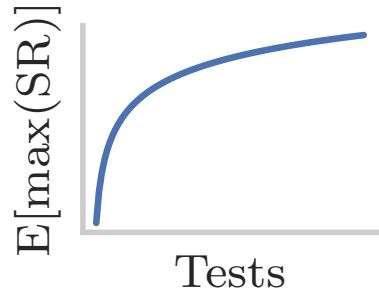
- Historically, this has been studied from a multiple testing perspective [BL14; HLZ16].



Expected Maximum Sharpe after N tests [BL14].

The Issue With Multiple Testing

- Historically, this has been studied from a multiple testing perspective [BL14; HLZ16].

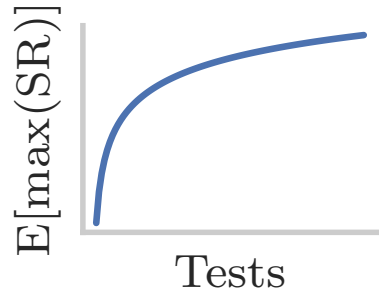


Expected Maximum Sharpe after N tests [BL14].

- However these approaches share a key problem:

The Issue With Multiple Testing

- Historically, this has been studied from a multiple testing perspective [BL14; HLZ16].

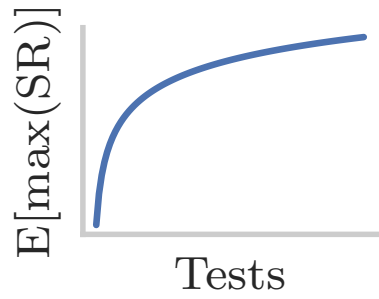


Expected Maximum Sharpe after N tests [BL14].

- However these approaches share a key problem: **The number of tests conducted almost always goes untracked, or unreported.**

The Issue With Multiple Testing

- Historically, this has been studied from a multiple testing perspective [BL14; HLZ16].



Expected Maximum Sharpe after N tests [BL14].

- However these approaches share a key problem: **The number of tests conducted almost always goes untracked, or unreported.**
- Instead, we model *why* researchers conduct multiple tests: **To search for higher Sharpe ratios.**

Question: How do we estimate a Sharpe ratio, accounting for selection bias?

Proposition: Use Bayesian methods to correct our Sharpe ratio estimates under selection bias, and leverage a dataset of observed in-sample and out-of-sample Sharpe ratios to fit our priors.

Findings: This particularly improves estimates for short backtests, and gives non-linear adjustments which diminish for high Sharpe ratios.

Chen and Zimmermann [CZ20]:

- The authors model **publication bias** in econometrics literature, looking at in-sample return only.
- Focuses on estimation **without** out-of-sample data.

Chen and Zimmermann [CZ20]:

- The authors model **publication bias** in econometrics literature, looking at in-sample return only.
- Focuses on estimation **without** out-of-sample data.

Our Contribution:

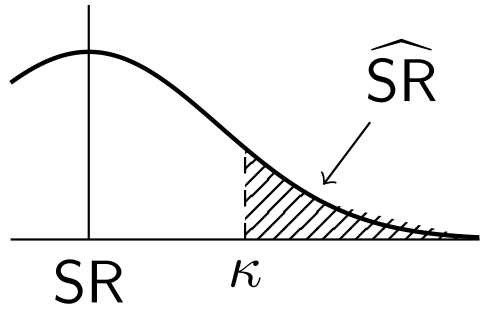
- Seek the **best Bayesian estimate** for **out-of-sample Sharpe ratios**.
- Utilise both **in-sample and out-of-sample** data to fit our model.

The Three Sharpe Ratios

- The **In-Sample** Sharpe ratio $\widehat{\mathrm{SR}} = \frac{\widehat{\mu}}{\widehat{\sigma}}$ is the Sharpe ratio you observe in your backtest,
- The **True** Sharpe ratio $\mathrm{SR} = \frac{\mu}{\sigma}$ is the unobserved population Sharpe ratio for your strategy,
- The **Out-of-Sample** Sharpe ratio $\widetilde{\mathrm{SR}} = \frac{\widetilde{\mu}}{\widetilde{\sigma}}$ is the Sharpe ratio you'll observe in live trading.

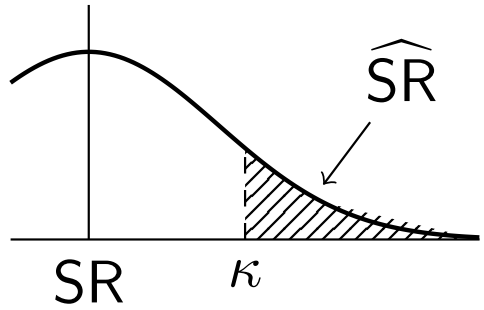
How Selection Creates Bias

Imagine we have selected a strategy based on a sample Sharpe ratio $\widehat{\text{SR}}$ which clears a minimum threshold κ.



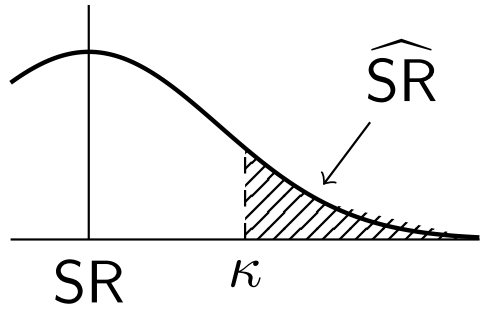
How Selection Creates Bias

Imagine we have selected a strategy based on a sample Sharpe ratio $\widehat{\text{SR}}$ which clears a minimum threshold κ.



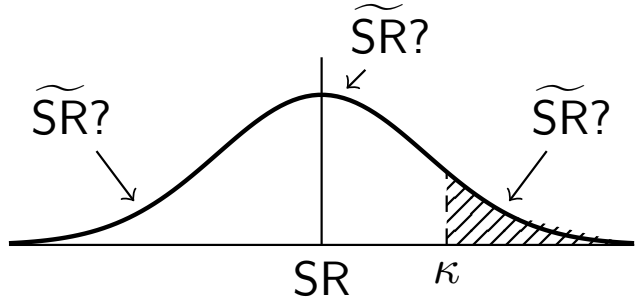
The sample Sharpe ratio $\widehat{\text{SR}}$ has been **biased upwards** from the true value SR due to selection from a **truncated distribution**.

Imagine we have selected a strategy based on a sample Sharpe ratio $\widehat{\text{SR}}$ which clears a minimum threshold κ.



The sample Sharpe ratio $\widehat{\text{SR}}$ has been **biased upwards** from the true value SR due to selection from a **truncated distribution**.

The out-of-sample Sharpe ratio will be a new draw from the sampling distribution, and will be less than our threshold $F_{\widetilde{\text{SR}}}(\kappa)\%$ of the time.



Sample Sharpe Ratio Distribution

- What is the sampling distribution of the Sharpe ratio?
- The sample Sharpe ratio is a scaled t-statistic, $t = \sqrt{T}\widehat{\mathrm{SR}}$.

Sample Sharpe Ratio Distribution

- What is the sampling distribution of the Sharpe ratio?
- The sample Sharpe ratio is a scaled t-statistic, $t = \sqrt{T}\widehat{\mathrm{SR}}$.
- Assuming the payoffs of the strategy are Normal, the sample Sharpe ratio has distribution $\sqrt{T}\widehat{\mathrm{SR}} \sim t_{T-1}\left(\sqrt{T}\mathrm{SR}\right)$.
- Although Normality is a strong (wrong? [Con01]) assumption, it's useful and doesn't hugely affect the results [Pav21].

- Taking into account a hard selection threshold, the truncated distribution of the sample Sharpe ratio is given by

$$f_{\widehat{\mathsf{SR}}|\widehat{\mathsf{SR}}>\kappa}\left(\widehat{\mathsf{SR}} \mid \mathsf{SR}, T, \kappa\right) = \frac{f_{\widehat{\mathsf{SR}}}\left(\widehat{\mathsf{SR}} \mid \mathsf{SR}, T\right) \mathbb{1}_{\widehat{\mathsf{SR}}>\kappa}}{1 - F_{\widehat{\mathsf{SR}}}\left(\kappa \mid \mathsf{SR}, T\right)},$$

where $f_{\widehat{\mathsf{SR}}}$ and $F_{\widehat{\mathsf{SR}}}$ denote the original density and CDF of the sample Sharpe ratio.

- Taking into account a hard selection threshold, the truncated distribution of the sample Sharpe ratio is given by

$$f_{\widehat{\mathsf{SR}}|\widehat{\mathsf{SR}}>\kappa}\left(\widehat{\mathsf{SR}} \mid \mathsf{SR}, T, \kappa\right) = \frac{f_{\widehat{\mathsf{SR}}}\left(\widehat{\mathsf{SR}} \mid \mathsf{SR}, T\right) \mathbb{1}_{\widehat{\mathsf{SR}}>\kappa}}{1 - F_{\widehat{\mathsf{SR}}}\left(\kappa \mid \mathsf{SR}, T\right)},$$

  where $f_{\widehat{\mathsf{SR}}}$ and $F_{\widehat{\mathsf{SR}}}$ denote the original density and CDF of the sample Sharpe ratio.
- We can then numerically compute the expected in-sample Sharpe ratio $\widehat{\mathsf{SR}}$ given a true Sharpe ratio SR, threshold κ and backtest length T.
- But if we take $\mathsf{SR} = 0$, then we have a closed form result.

Proposition

Let $\mathsf{SR} = 0$*, then the expected sample Sharpe is given by,*

$$\mathbb{E}\left[\widehat{\mathsf{SR}} \,\middle|\, \widehat{\mathsf{SR}} > \kappa, \mathsf{SR} = 0\right] = \frac{U}{\sqrt{T}}$$

where

$$U = \gamma \frac{T-1}{T-2}\left(1 + \frac{T}{T-1}\kappa^2\right)^{-\frac{T-2}{2}}$$

$$\gamma = \frac{\Gamma\left(\frac{T}{2}\right)}{\alpha_0 \Gamma\left(\frac{T-1}{2}\right)\sqrt{(T-1)\pi}}, \qquad \alpha_0 = 1 - F_t\left(\sqrt{T}\kappa; T-1\right).$$

Expected Bias Over True Sharpe Ratio

- Setting any threshold causes the expected in-sample Sharpe ratio to be larger than zero.

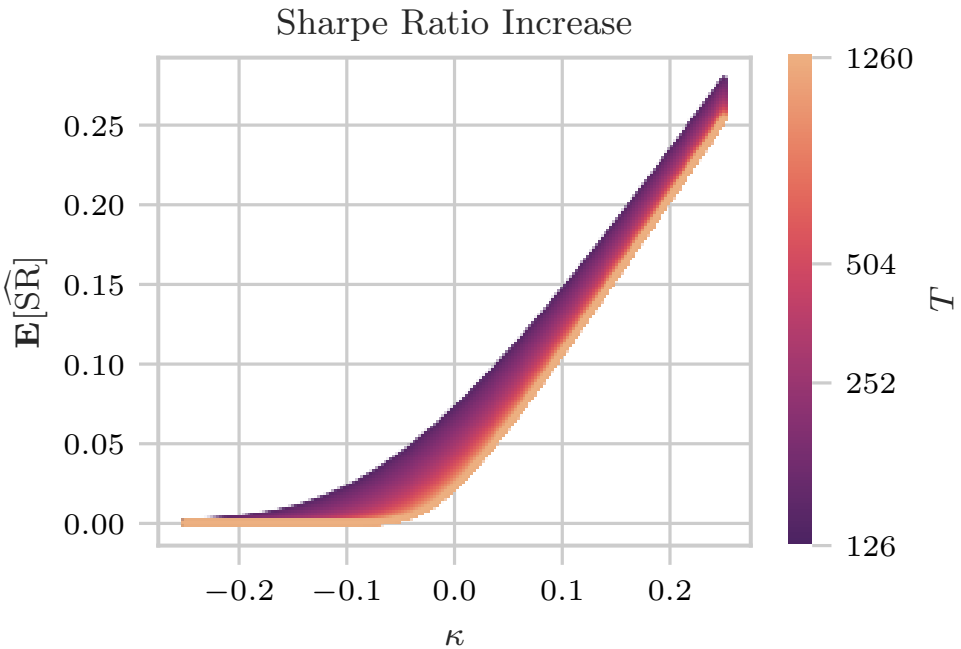
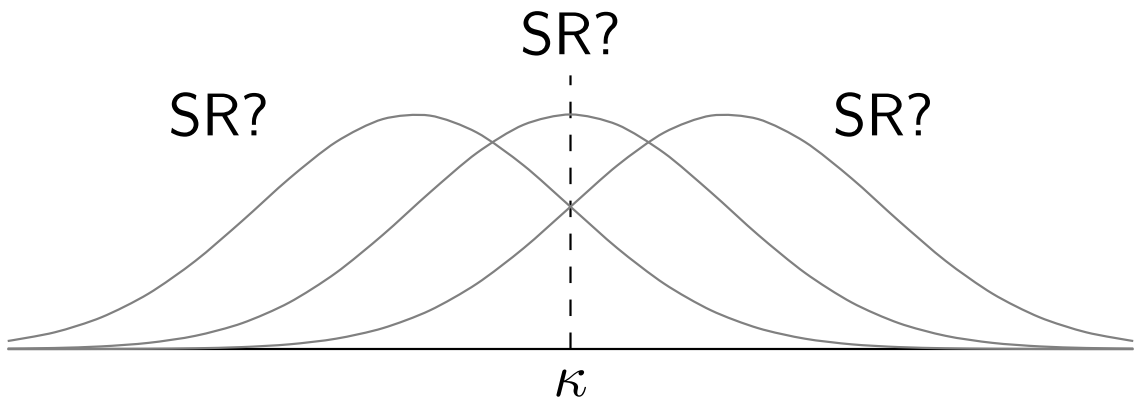


Figure: Sharpe ratio inflation by κ and T. Values are in terms of daily Sharpe ratios, and T is in days.

In reality, the true SR is unknown, and we don't know if our threshold is above or below the truth.



So how could we estimate the true Sharpe, given an observed sample Sharpe ratio and a threshold?

We need:

- A prior for the Sharpe ratio $p\left(\mathsf{SR} \mid \Theta\right)$,
- An appropriate likelihood which is aware of the threshold bias $p\left(\widehat{\mathsf{SR}} \mid \mathsf{SR}, \widehat{T}, \Theta, \mathsf{acc}\right)$.

We can then compute the posterior for the true Sharpe ratio,

$$p(\mathsf{SR} \mid \widehat{\mathsf{SR}}, \widehat{T}, \Theta, \mathsf{acc}) \propto p\left(\mathsf{SR} \mid \Theta\right) p\left(\widehat{\mathsf{SR}} \mid \mathsf{SR}, \widehat{T}, \Theta, \mathsf{acc}\right).$$

Sharpe Prior

If the payoffs are Normal, then a Normal-Inverse-Gamma prior is the natural choice for the variance and the Sharpe ratio, [Pav21]

$$\sigma^2 \sim \Gamma^{-1}\left(\frac{m_0}{2}, \sigma_0^2 \frac{m_0}{2}\right),$$

$$\text{SR} \mid \sigma^2 \sim \mathcal{N}\left(\frac{\mu_0}{\sigma}, \frac{1}{n_0}\right),$$

Sharpe Prior

If the payoffs are Normal, then a Normal-Inverse-Gamma prior is the natural choice for the variance and the Sharpe ratio, [Pav21]

$$\sigma^2 \sim \Gamma^{-1}\left(\frac{m_0}{2}, \sigma_0^2 \frac{m_0}{2}\right),$$
$$\mathrm{SR} \mid \sigma^2 \sim \mathcal{N}\left(\frac{\mu_0}{\sigma}, \frac{1}{n_0}\right),$$

which yields a marginal prior for the Sharpe ratio,

$$\sqrt{n_0}\mathrm{SR} \sim \lambda'\left(\sqrt{n_0}\mathrm{SR}_0, m_0\right),$$

where λ' denotes Lecoutre's lambda prime distribution.[1]

1. The lambda prime distribution is related to the t distribution by their CDFs, $F_{\lambda'}(x; t, \nu) = 1 - F_t(t; x, \nu)$.

As in [CZ20] we extend to a soft rather than hard threshold,

$$p_{\text{acc}}\left(x \mid \kappa, \ell\right) = \frac{1}{1 + \exp(-\ell(x - \kappa))},$$

As in [CZ20] we extend to a soft rather than hard threshold,

$$p_{\text{acc}}\left(x \mid \kappa, \ell\right) = \frac{1}{1 + \exp(-\ell(x - \kappa))},$$

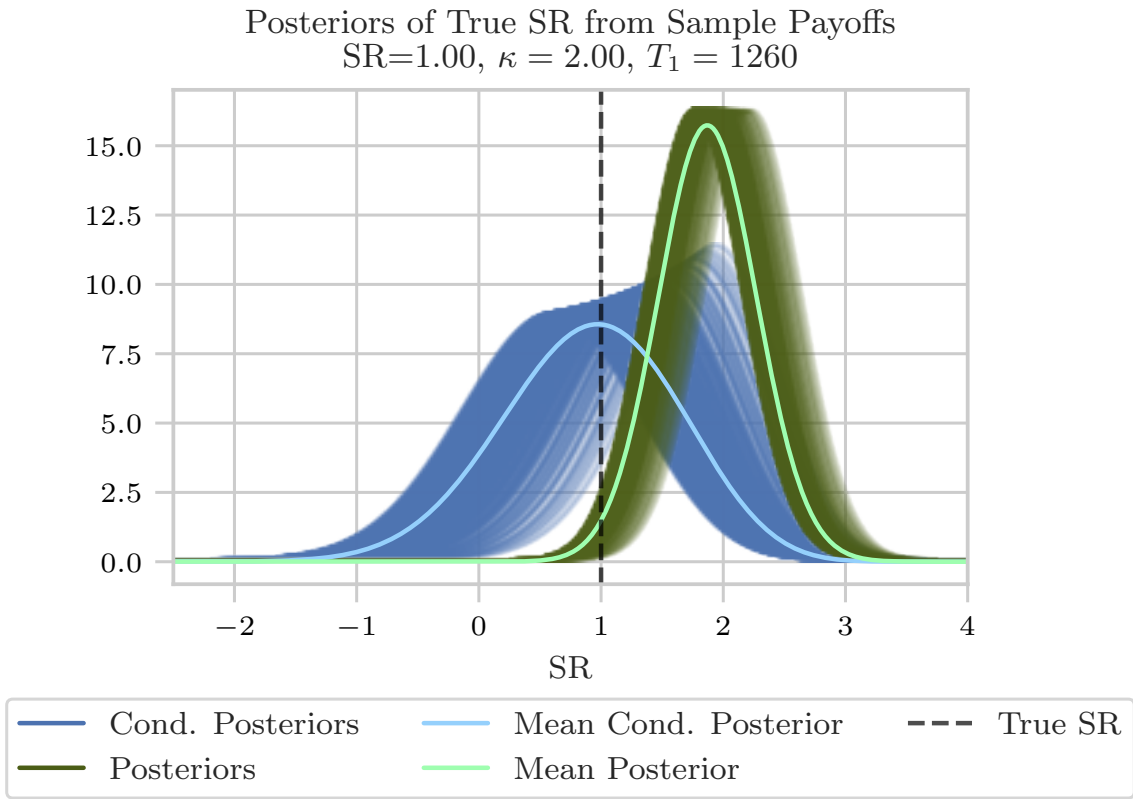
and we already know the sample distribution of the Sharpe ratio, $p_{\widehat{\text{SR}}}(\widehat{\text{SR}} \mid \text{SR}, \widehat{T})$, so our likelihood is given by

$$p\left(\widehat{\text{SR}} \mid \text{SR}, \widehat{T}, \Theta, \text{acc}\right) = \frac{p_{\widehat{\text{SR}}}(\widehat{\text{SR}} \mid \text{SR}, \widehat{T}) p_{\text{acc}}(\widehat{\text{SR}} \mid \kappa, \ell)}{\mathbb{E}_{\widehat{\text{SR}}}\left[p_{\text{acc}}(\widehat{\text{SR}} \mid \kappa, \ell)\right]}.$$

We can now use our bias-aware posterior to recover the true Sharpe ratio!

Voila!

Given a sample Sharpe ratio, threshold, and backtest length we can more accurately recover the true Sharpe ratio than a naive approach:



But What is Θ?

However, a problem remains: **How do we pick Θ?**

But What is Θ?

However, a problem remains: **How do we pick Θ?**

We propose using empirical Bayes to estimate the hyperparameters. This has a few key advantages:

1. We "leverage" a dataset of observed in-sample *and* out-of-sample Sharpe ratios to fit our parameters.

But What is Θ?

However, a problem remains: **How do we pick Θ?**

We propose using empirical Bayes to estimate the hyperparameters. This has a few key advantages:

1. We "leverage" a dataset of observed in-sample *and* out-of-sample Sharpe ratios to fit our parameters.
2. This enables us to correct for the selection bias, and also improve performance estimates in a limited information environment.

But What is Θ?

However, a problem remains: **How do we pick Θ?**

We propose using empirical Bayes to estimate the hyperparameters. This has a few key advantages:

1. We "leverage" a dataset of observed in-sample *and* out-of-sample Sharpe ratios to fit our parameters.
2. This enables us to correct for the selection bias, and also improve performance estimates in a limited information environment.
3. This is particularly useful for short backtests, where we have limited information to accurately estimate the Sharpe ratio of the strategy.

- We can use maximum likelihood estimation to estimate the parameters in $\Theta = \{n_0, m_0, \mathsf{SR}_0, \kappa, \ell\}$ from a dataset of strategies.
- If we have a dataset of strategies, we likely have both in-sample and out-of-sample performance for each strategy.
- We would like to use all the data available to us, so we need the joint density of the in-sample and out-of-sample Sharpe ratio for a singular strategy:

$$p\left(\widehat{\mathsf{SR}}_i, \widetilde{\mathsf{SR}}_i \mid \widehat{T}_i, \widetilde{T}_i, \Theta, \mathsf{acc}\right)$$

Joint Density of In-Sample and Out-of-Sample Sharpe Ratios

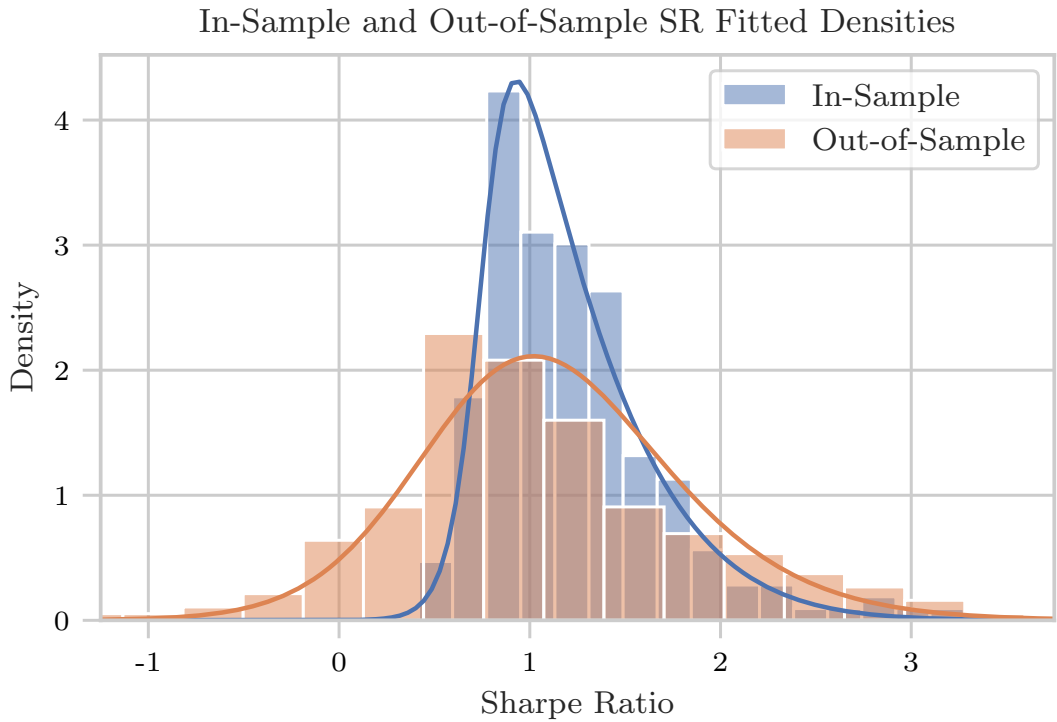
- To get the joint density of $\widehat{\mathrm{SR}}, \widetilde{\mathrm{SR}}$ we use the joint density of $\widehat{\mu}, \widehat{\sigma}^2, \widetilde{\mu}, \widetilde{\sigma}^2$ and apply the necessary transform.
- The joint density of the sample means and variances is given by,

$$p\left(\widehat{\mu}, \widehat{\sigma}^2, \widetilde{\mu}, \widetilde{\sigma}^2 \mid \widehat{T}, \widetilde{T}, \Theta, \mathrm{acc}\right) = \overbrace{q(\widetilde{\mu}, \widetilde{\sigma}^2 \mid \widetilde{T}, \Theta_1)}^{\text{OOS stats}} \overbrace{\frac{q(\widehat{\mu}, \widehat{\sigma}^2 \mid \widehat{T}, \Theta_0) p_{\mathrm{acc}}(\widehat{\mu}/\widehat{\sigma} \mid \kappa, \ell)}{\int_0^\infty \int_{-\infty}^\infty q(m, s^2 \mid \widehat{T}, \Theta_0) p_{\mathrm{acc}}(m/s \mid \kappa, \ell)\, \mathrm{d}m\, \mathrm{d}s^2}}^{\text{IS stats}},$$

  where $q(x, y)$ is the joint likelihood of the sample mean and variance given a Normal-Inverse-Gamma prior, which we have marginalised out, and Θ_j refers to either the prior, or posterior updated, parameters of the prior.
- The likelihoods of the out-of-sample and in-sample statistics are not independent!

Voila Encore!

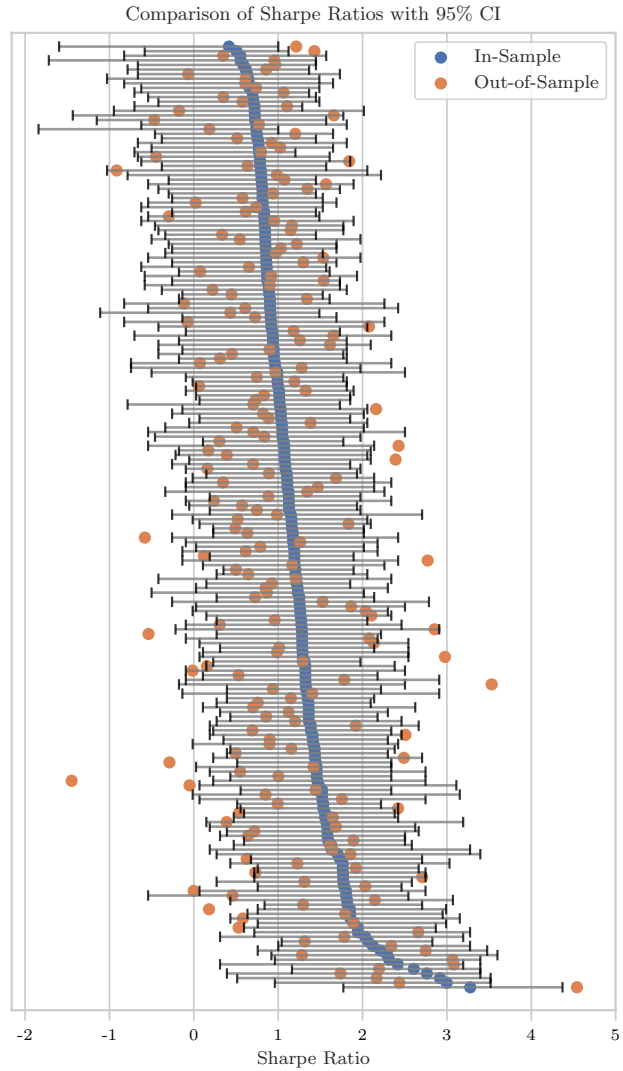
- Using this likelihood to fit our parameters to a real dataset provides a very good fit.



Dataset: 206 predictors from [CZ22].

Voila Encore!

And using the fitted model we can compute confidence intervals for the out-of-sample performance for each strategy.



Dataset: 206 predictors from [CZ22].

1. We can fit a Bayesian model to a dataset of in-sample and out-of-sample Sharpe ratios.

1. We can fit a Bayesian model to a dataset of in-sample and out-of-sample Sharpe ratios.
2. The model takes into account the selection criteria used to choose strategies for trading.

Conclusion

1. We can fit a Bayesian model to a dataset of in-sample and out-of-sample Sharpe ratios.
2. The model takes into account the selection criteria used to choose strategies for trading.
3. And we can use this model to improve estimates of out-of-sample performance of new candidate strategies!

Bibliography I

[Ame+08] Noël Amenc et al. *EDHEC European Investment Practices Survey*. Tech. rep. EDHEC, Jan. 2008, p. 156.

[BL14] David H. Bailey and Marcos López De Prado. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality". In: *The Journal of Portfolio Management* 40.5 (Sept. 2014), pp. 94–107. ISSN: 0095-4918, 2168-8656. DOI: 10.3905/jpm.2014.40.5.094.

[Con01] R. Cont. "Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues". In: *Quantitative Finance* 1.2 (Feb. 2001), pp. 223–236. ISSN: 1469-7688. DOI: 10.1080/713665670.

[CZ20] Andrew Y Chen and Tom Zimmermann. "Publication Bias and the Cross-Section of Stock Returns". In: *The Review of Asset Pricing Studies* 10.2 (June 2020). Ed. by Jeffrey Pontiff, pp. 249–289. ISSN: 2045-9920, 2045-9939. DOI: 10.1093/rapstu/raz011.

[CZ22] Andrew Y. Chen and Tom Zimmermann. "Open Source Cross-Sectional Asset Pricing". In: *Critical Finance Review* 11.2 (May 2022), pp. 207–264. ISSN: 2164-5744, 2164-5760. DOI: 10.1561/104.00000112.

[HLZ16] Campbell R. Harvey, Yan Liu, and Heqing Zhu. "... and the Cross-Section of Expected Returns". In: *The Review of Financial Studies* 29.1 (Jan. 2016), pp. 5–68. ISSN: 0893-9454. DOI: 10.1093/rfs/hhv059.

[Pav21] Steven E. Pav. *The Sharpe Ratio: Statistics and Applications*. New York: Chapman and Hall/CRC, Sept. 2021. ISBN: 978-1-00-318105-7. DOI: 10.1201/9781003181057.