Index matching portfolios

Robert Skowron[*] and Nicholas Syring[†]

August 3, 2019

Abstract

We review approaches to solving the sparse portfolio index tracking problem in investment management. We explore in additional depth the work of George and McCulloch (1997) and elaborate on the applications of a spike and slab variable selection process for tracking portfolios. Potential further applications of this method are discussed. Finally, we discuss the potential applications of generalized fiducial inference to the problem and areas for continued research.

Keywords and phrases: Sparse portfolio, index tracking, regularization, variable selection, Bayesian inference, generalized fiducial inference, Markov chain Monte Carlo

1 Introduction

Stock indices like the S&P 500 and Russell 2000 are large, diverse groups of assets of particular relevance to investors. However, for active managers, there are many reasons why it would not it beneficial to invest even a moderate portion of the assets in these indices. There are several reasons for this including the costs associated with holding many assets, the difficulty of managing a portfolio of many assets, and the inability to deploy excess capital to higher alpha generating bets. Therefore investors may seek to closely replicate or track the performance of a whole index by carefully selecting a small, manageable subset of its assets and deploying excess capital to more tactical bets.

The most straightforward way to solve the index tracking problem is to pose it as an optimization problem in which one minimizes a measure of the tracking error over acceptable portfolios. Difficulties with this approach arise if one includes practical considerations in the optimization problem such as bounds on the number of assets and amounts invested in each, which take the form of constraints; see, e.g. Fastrich et al. (2014) and Benidis et al. (2018). These constrained optimizations often can be solved using mixed integer programming but may converge slowly for high-dimensional data.

Other approaches to index tracking involve replacing the tracking error minimization by an approximation or heuristic algorithm as in the above references or by changing

[*]McKelvey School of Engineering, Washington University in St. Louis, r.skowron@wustl.edu@wustl.edu

[†]Department of Mathematics and Statistics, Washington University in St. Louis, nasyring@wustl.edu.

the optimization criteria. For example, a Markowitz mean-variance portfolio model that minimizes portfolio variance while achieving a minimum return can approximate index tracking when the target return is the index return; see Fastrich et al. (2015) and Puelz et al. (2018).

In contrast to optimization methods, Bayesian methods for index tracking output a range of potential portfolios rather than a single solution. This can be beneficial as estimating a covariance matrix for a large number of correlated assets requires a significant time scale to avoid singularity or ill-condition. George and McCulloch (1997) regress index returns on asset returns and use a spike and slab type prior to encourage sparsity in the number of assets selected to remain in the model. The marginal posterior of included assets can quantify the uncertainty in asset selection, better enabling the portfolio manager to make a final subjective judgement. Variations on the approach in George and McCulloch (1997) involve different prior specifications of sparsity, such as the SCAD penalty of Fan and Li (2001) and the spike and slab lasso of Ročková and George (2018). A new Bayesian-like generalized fiducial approach to sparse regression was introduced in Williams and Hannig (2019). Rather than directly encouraging sparse portfolios, the generalized fiducial approach limits acceptable portfolios to those with only minimally correlated assets.

The paper is organized as follows. Section 2 reviews the optimization problem and various proposed methods for solving. Section 3 reviews Bayesian approaches to the same problem. Section 3 also includes an in-depth look at the spike and slab approach of George and McCulloch (1997) with a more detailed discussion of applications. Additionally, we discuss a possible way to solve the constrained optimization problem using a Dirichlet prior with reversible jump Markov chain Monte Carlo. Section 4 investigates the generalized fiducial inference (GFI) approach and discusses potential applications of such a method. Finally, the paper is concluded in section 6 with suggestions for additional avenues of future investigation.

2 The Optimization Problem

Adopting the notation of Benidis et al. (2018) let $\mathbf{r}^b = (r_1^b, ..., r_T^b)^\top \in \mathbb{R}^T$ and $X = [\mathbf{r}_1, ..., \mathbf{r}_T]^\top \in \mathbb{R}^{T\times N}$ denote the returns of the index and the N assets of the index over T time periods. Let $\mathbf{b} \in \mathbb{R}_+^N$ denote the normalized index weights of each asset, i.e. $b^\top 1_{T\times 1} = 1$ and $X\mathbf{b} = \mathbf{r}^b$.

A portfolio is defined as a weight vector $\mathbf{w} = (w_1, ..., w_N)$ giving the proportion invested in each asset. For example, when the investor is limited to long positions the portfolio satisfies $w_i \geq 0$ for every $i = 1, ..., N$ and $\mathbf{w}^t op 1 = 1$. Then, the tracking error of the portfolio can be measured in many ways, the primary of which is the L_2-error or empirical tracking error

$$\mathrm{ETE}(\mathbf{w}) = \frac{1}{T}||X\mathbf{w} - \mathbf{r}^b||_2^2.$$

As described in the introduction, we desire a portfolio with only a small number of assets relative to the size of the index. However, sparse minimization of the empirical tacking error of a long portfolio is not a trivial problem. Benidis et al. (2018) defines the sparse optimization problem

$$\begin{aligned} &\text{minimize}_{\mathbf{w}} \frac{1}{T}||X\mathbf{w} - \mathbf{r}^b||_2^2 + \lambda||\mathbf{w}||_0 \\ &\text{subject to} \mathbf{w}^\top 1_{N\times 1} = 1, \\ &\qquad \mathbf{w} \geq 0_{N\times 1}. \end{aligned} \tag{1}$$

The primary challenge is the presence of the nonconvex penalty $\lambda||\mathbf{w}||_0$. Benidis et al. (2018) introduce a convex approximation to this penalty and perform the minimization using their LAIT and related procedures. The resulting solution produces a sparse, long-only portfolio with good tracking performance.

Other potential methods of solving this problem include mixed integer programming, genetic algorithms. However, it is not possible to guarantee optimality using these methods and they often come with a considerable performance cost. Other heuristic methods can also be applied. For example, one could solve the unconstrained optimization problem and identify the most significant contributors. Then, secondary optimizations could be run with the limited set of assets and constraints can be specified as a constant per choice of assets.

Regardless of the methodology, these methods all result in singular solutions. Given that we are using high dimensional, correlated data, the the empirical covariance matrix can often be ill-conditioned which can result in highly sensitive results. In the next section, we discuss Bayesian methods which provide distributional results which may help alleviate some of those issues.

3 Bayesian methods

George and McCulloch (1997) presents methods for variable selection in regression from a Bayesian viewpoint. A primary application is construction of index-tracking portfolios. They begin with a regression model of the index returns $\mathbf{r}^b = X\mathbf{w} + \epsilon$ where $\epsilon \sim \mathsf{N}_T(0, \sigma^2 I_{T\times T})$. Such a regression model is somewhat artificial because the index returns actually are a *deterministic* linear combination of the constituent asset returns. However, since the Gaussian kernel contains the (negative) empirical tracking error, maximizing the likelihood of (a sparse version of) the model is equivalent to minimizing the empirical tracking error. To achieve sparsity in the predictors, and hence a small portfolio, George and McCulloch (1997) consider independent Bernoulli priors of the form

$$\pi(\gamma) = \prod_{i=1}^{N} \alpha_i^{\gamma_i}(1-\alpha_i)^{(1-\gamma_i)} \tag{2}$$

for $\alpha_i \in (0, 1)$ and where $\gamma \in \{0, 1\}^N$ denotes which assets are included in the portfolio. Then, conditional on such γ, the β_i are distributed as

$$\pi(\beta_i|\gamma) = (1-\gamma_i)N(0, v_{0_{\gamma_i}}) + \gamma_i N(0, v_{1_{\gamma_i}}) \tag{3}$$

where v_0 and v_1 can be constant or varying based say on the number of non-zero gammas. For the problem at hand, this construction poses two challenges. First, it does not guarantee that β_i will not be less than zero (i.e. a long-only constraint) and secondly, that $\sum \beta_i = 1$ (i.e. no leverage constraint).

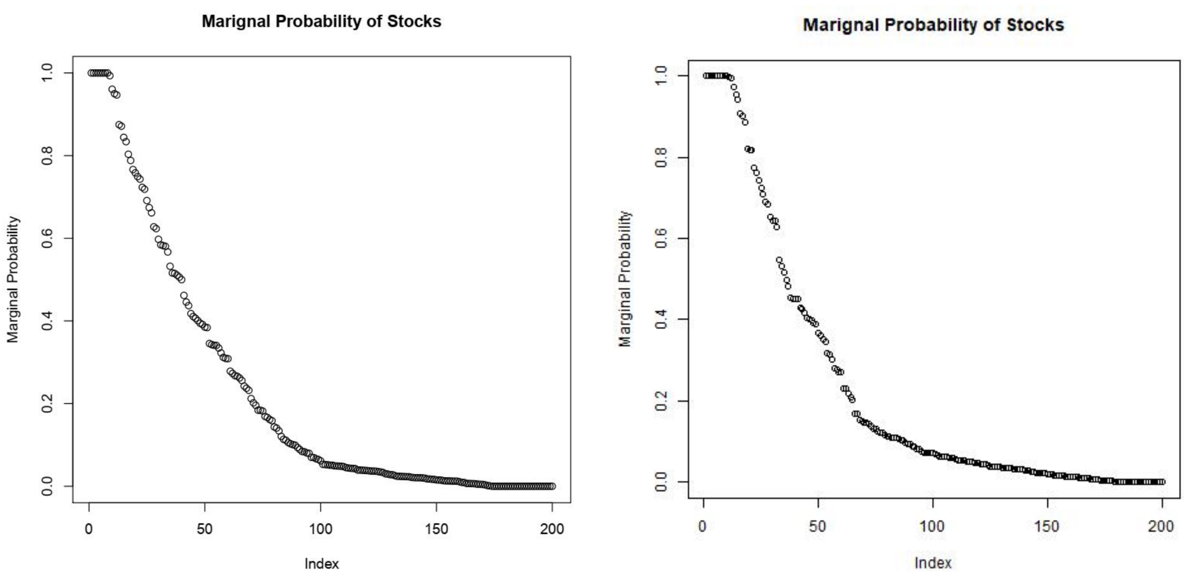
Solving this problem is accomplished by running MCMC with a Gibbs sampler.

Algorithm 1 MCMC for spike and slab

$\gamma_i = 0 \forall i$
while $n < N$ **do**
  Choose i and set $\gamma_i = 1 - \gamma_i$
  Compute the change in the log posterior $\pi(\gamma|Y) = r$
  Randomly sample from a $Binomial(1, \min(e^r, 1))$
  if Sample = 1 **then**
    Accept the proposed γ
  else
    Reject the proposed γ
  end if
end while

3.1 Revisiting George and McCulloch (1997)

To begin our analysis, we start by refreshing the example performed in George and McCulloch (1997) and providing some additional insights. Using the Wharton Research Data Services (WRDS) we obtain weekly returns for stocks in the S&P 500 from January 2012 to December 2018. Following the methodology laid out in Section 6 of George and McCulloch (1997) we compute the marginal probabilities of inclusion for the stocks. Utilizing 200 randomly chosen stocks, as in their example, we can observe a similar distribution of marginal inclusion probabilities where a small number are nearly always included and rapidly fall off with fewer than 50 stocks being selected more than a handful of iterations.

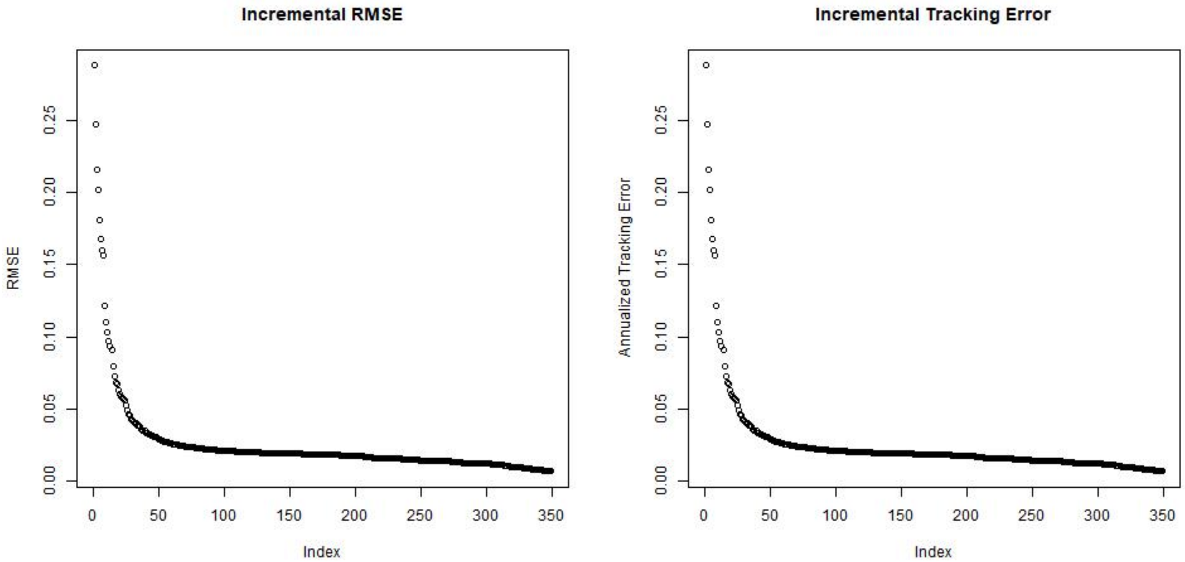


(a) Replication of 1985-1991 analysis

(b) Algorithm applied to 2012-2018

Figure 1: Marginal probability of inclusion by applying George and McCulloch (1997) to the weekly returns for S&P 500 from 2012-2018

G&M continue by running nested regressions and identifying the incremental explanatory power as measured by R^2 of each additional candidate stock. Here, we break from G&M and explore two different measures, in-sample and post tracking error.



(a) Incremental root mean squared error

(b) Incremental tracking error

Figure 2: In sample results applying George and McCulloch (1997) to the weekly returns for S&P 500 from 2012-2018. Values are computed by sequentially including assets based on marginal probability

G&M also makes no comment about the selection of the stocks. For example, looking at those top several stocks that appear to be chosen almost always, do those remain consistent over multiple iterations? Or, particularly due to the correlation structure inherent in the stock market, do we see patterns emerge where highly correlated stocks might be chosen as equivalents? As it turns out, the top stocks selected are frequently the same. Over numerous iterations, we identify stocks such as AAPL, AMZN, BA, XOM, & MSFT being selected frequently. Annecdotally, this makes sense as these correspond to the largest market cap stocks or significant drivers of returns in recent years.

However, it also begs the question of whether or not a naive approach to stock selection would do just as well. To investigate this, we use the same period of data (2012-2018) and compare both the in and out of sample performance of the G&M algorithm to five different naive algorithms:

1. Selecting the top 10 stocks by market cap
2. Selecting the top stock in each GICS Sector by market cap (≈ 10)
3. Selecting the top stock in each GICS Group by market cap (≈ 25)
4. Selecting the top stock in each GICS Industry by market cap (≈ 70)
5. Selecting the top stock in each GICS Sub-Industry by market cap (≈ 150)

where GICS refers to the Global Industry Classification Standard. The G&M algorithm is run on 250 days of data (approximately 1 year) and any stock that has a marginal inclusion probability of greater than 80% is selected. For the same date, we construct the above naive portfolios using equal weighted contributions from the selected stocks. The portfolios are then held for 20 days (approximately 1 month) after which we rebalance the portfolios using the same logic and repeat until the end of 2018.

Add plots here showing tracking error and performance

Optimization
- Implement Benidis et al. (2018) optimization approaches for setups we find interesting, setups meaning different constraints present on either number of assets or amount held or long-only portfolios, etc.

Bayesian
- Extend the basic approach in George and McCulloch (1997) to include portfolio constraints as in Benidis et al. (2018). The point of this would be to compare to the optimization approaches in Benidis et al. (2018) and see how valuable the Bayesian uncertainty quantification could be. How much does the optimal solution differ from the posterior? How wide is the posterior and how far can you get from the optimum while maintaining acceptable performance.

Gen. fid.
- Implement Williams and Hannig (2019) approach on index data.
- Tweak Williams and Hannig (2019) definition 2.1 of $\varepsilon-$admissibility to account for additional constraints, like those in Benidis et al. (2018).
- Compare performance to George and McCulloch (1997), optimization approaches.

3.2 With constraints

The basic hierarchical model of George and McCulloch (1997) places a normal prior on the asset weights w, which does not take into account any type of holding constraints, e.g. long only portfolios. Here we present an alternative model for long-only portfolios:

$$\mathbf{r}^b \sim \mathsf{N}_T(X\mathbf{w}, \sigma^2 I_{T\times T}) \tag{4}$$

$$\gamma_i \overset{ind.}{\sim} \mathsf{Ber}(\alpha_i) \tag{5}$$

$$\sigma^2|\gamma \sim \mathrm{IG}(\nu/2,\ \nu\lambda_\gamma/2) \tag{6}$$

$$\mathbf{w}|\gamma,\sigma^2 \sim Dir(||\gamma||_0; \beta). \tag{7}$$

Here the asset weights are given a Dirichlet prior conditional on the included assets. This prior enforces the constraints $\mathbf{w}^\top 1_{N\times 1} = 1$ and $\mathbf{w} \geq 0_{N\times 1}$. Implementing this model is difficult in practice as it requires us to utilize Reversible Jump MCMC in order to change candidate stock inclusions. Roughly, the algorithm is as follows:

The posterior density π_n of $(\gamma, \mathbf{w})$ given the data is proportional to the likelihood times the prior density, and can be sampled using, for instance, Metropolis-Hastings within a Gibbs sampler.

1. Begin with the i^{th} sample $(\gamma^i, \mathbf{w}^i)$.

2. Propose a new sample of γ, call it $\gamma^\star$, by drawing it at random from the proposal distribution. For now, let's say the proposal is Bernoulli with some high probability of 1 if $\gamma_j^i = 1$ and some low probability of 1 if $\gamma_j^i = 0$; such a proposal will tend to slowly change the included assets.

3. Compute the acceptance ratio (on the log scale):

$$a = \log \pi_n(\gamma^\star, \mathbf{w}^i) - \log \pi_n(\gamma^i, \mathbf{w}^i) + \log \pi(\gamma^i) - \log \pi(\gamma^\star)$$

4. Generate $U \sim \mathsf{Unif}(0, 1)$ and accept $\gamma^{i+1} = \gamma^\star$ if $\exp(a) > U$.

However, while theoretically an alternative option to optimization problems, the implementation requires considerable time to converge. As an example, over several tens of thousands of iterations, if the number of candidate stocks was larger than 25 the algorithm rarely moved significantly away from equal weighting. This limitation puts this methodology out of consideration for many portfolio managers. For more information about the reversible jump MCMC, see Hastie and Green (2012).

4 Generalized Fiducial Inference

Fiducial Inference is an area of statistical inference original proposed by R.A. Fisher to attempt to find distributional representations of variables without requiring priors. While originally dismissed, recently research in Generalized Fiducial Inference (GFI) has identified some potentially valuable applications. In particular, we want to point out the recent paper of Williams and Hannig (2018). In their paper, they explore the use of GFI for variable selection in cases where the data is highly collinear. Particularly in the financial world where assets can be highly correlated, this appears to be an interesting topic to explore.

Williams and Hanning (2018) describe an approach where the underlying premise is that the non-zero parameters are non-redundant in that they contain the minimal amount of information to explain or predict the observed data. These subsets are referred to as *ϵ-admissible*. Utilizing the code provided by Mr. Williams, with some adjustments, we attempted to apply the algorithm to our 2012-2018 returns data. Our observations were that the approach did correctly narrow down to approximately 6-10 stocks to best track. However, the selections were widely varying as contrasted with the George and McCulloch algorithm.

There are additional areas for future research here. First, if the top stocks selected are indeed optimal, can we identify the related ϵ-admissible subsets to use in analyzing highly correlated alternative choices. That is, do the ϵ-admissible subsets beget some information about the inherent structure of the market? Secondly, given the fact that it does appear to accurately reduce the dimension of the problem, can this be applied as an initial filter to the standard optimization problem, similar to how we have shown as an extension of the George and McCulloch algorithm?

5 Conclusion

In this paper we review some of the materials and common methods used in asset management to identify sparse sets of securities for index tracking. We refreshed the algorithm put forth in George and McCulloch (1997) with more recent data

Additionally, there are several future avenues of research. First, does the density or distribution of the selection probabilities from the George and McCulloch algorithm provide any information about the structure of the market? Secondly, the premise of the variable selection methodology set forth in Williams and Hannig (2018) seems particularly useful for this scenario. Perhaps not solely in the stock selection realm but in the determination of the ϵ-admissible subsets and whether or not those provide any information about relationships.

Use of the stocks selected to reduce the parameter set

All code utilized in the analysis for this paper can be found on Github: https://github.com/bob-skowron/SU19-Independent-Study

References

Benidis, K., Feng, Y., and Palomar, D. P. (2018). Sparse Portfolios for High-Dimensional Financial Index Tracking. *IEEE Transactions on Sigmal Processing* 66(1), 155-170.

Fan, J., and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of American Statistical Association* 96, 1348–1360.

Fastrich, B., Paterlini, S., and Winker, P. (2014). Cardinality versus q-norm constraints for index tracking. *Quantitative Finance* 14(11), 2019-2032, DOI: 10.1080/14697688.2012.691986

Fastrich, B., Paterlini, S. and Winker, P. (2015). Constructing optimal sparse portfolios using regularization methods. *Comput Manag Sci* 12(3) 417–434. https://doi.org/10.1007/s10287-014-0227-5

George, E. I., and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* 7, 339-373.

Hastie, D. I. and Green, P. J. (2012), Model choice using reversible jump Markov chain Monte Carlo. Statistica Neerlandica, 66: 309-338. doi:10.1111/j.1467-9574.2012.00516.x

Puelz, D., Hahn, R. and Carvalho, Carlos M. (2018). Sparse Mean-Variance Portfolios: A Penalized Utility Approach. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2729504`.

Ročková, Veronika and George, Edward. (2018). The Spike-and-Slab LASSO. Journal of the American Statistical Association. 113. 10.1080/01621459.2016.1260469.

Williams, J. P., and Hannig, J. (2019). Non-Penalized Variable Selection In High-Dimensional Linear Model Settings Via Generalized Fiducial Inference. *Annals of Statistics.* 47(3), 1723-1753.