Index matching portfolios

Robert Skowron* and Nicholas Syring†

June 30, 2019

Abstract

Keywords and phrases: 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 5000 are large, diverse groups of assets especially attractive to investors. However, most investors cannot afford to invest in even a moderate portion of the assets these indices include due to several reasons such as the high costs associated with holding many different assets and the difficulty of managing a portfolio including so many different assets. Therefore investors seek to closely replicate or track the performance of a whole index by carefully selecting a small, manageable subset of its assets.

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 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. George and McCulloch (1997) regress index returns on asset returns and use a spike and slab type prior to

*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.

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 financial analyst 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). 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.

2 What should/can we do?

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.

Optimization

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 tho portfolio satisfies $w_i \geq 0$ for every $i = 1, ..., N$ and $\mathbf{w}^t op1 = 1$. Then, the tracking error of the portfolio can be measured in many ways, one 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

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

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 portfolio with good tracking performance.

3 Basic Bayesian methods

George and McCulloch (1997) presents methods for variable selection in regression from a Bayesian viewpoint. Their 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.

3.1 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:

$$\mathbf{r}^b \sim \mathsf{N}_T(X\mathbf{w}, \sigma^2 I_{T\times T}) \tag{3}$$

$$\gamma_i \overset{ind.}{\sim} \mathsf{Ber}(\alpha_i) \tag{4}$$

$$\sigma^2|\gamma \sim \mathrm{IG}(\nu/2,\, \nu\lambda_\gamma/2) \tag{5}$$

$$\mathbf{w}|\gamma, \sigma^2 \sim Dir(||\gamma||_0; \beta). \tag{6}$$

I'm unconvinced it is important to include σ^2 rather than just set it equal to 1.

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 $\mathbf{w} \geq 0_{N\times 1}$.

It is not obvious how to sample the posterior, which is not of a simple conjugate form. Most likely, we would use a reversible jump MCMC algorithm.

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.

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`.

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.