Efficient Bayesian Response Surface Maximization

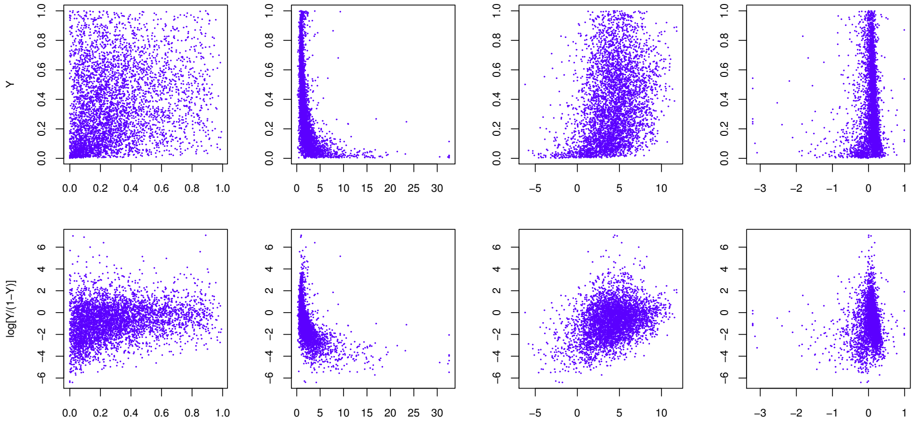
Feng Li

School of Statistics and Mathematics
Central University of Finance and Economics
Email: feng.li@cufe.edu.cn

Outline

The firm leverage data

leverage (Y):	total debt/(total debt+book value of equity), 4405 observations;
tang:	tangible assets/book value of total assets;
market2book:	(book value of total assets - book value of equity + market value of equity) / book value of total assets;
logSales:	logarithm of sales;
profit:	(earnings before interest, taxes, depreciation, and amortization) / book value of total assets.



Our interests

- Find an optimal combination of **tang**, **market2book**, **logSales**, and **profis** so that **leverage** reaches the maximum.
- We may write it down in mathematics

$$\arg\max_{x \in R^d} \; f(x)$$

  where x is defined as sample and $y = f(x) + \epsilon$ with ϵ as the noisy observation of the objective function at x.
- But note that
    - The function $f(x)$ is **unknown**, **not noise-free** and **hard to evaluate**.
    - We do not know its derivatives. Common optimization methods usually fail here.
    - The covariates space can be very sparse.

Bayesian Optimization I

1. Assume $D_{1:t} = \boldsymbol{x}_{1:t}, \boldsymbol{y}_{1:t}$ are observations, a prior distribution $P(f)$ over function $f(\cdot)$ is combined with the likelihood function $P(D_{1:t}|f)$ to produce the posterior distribution

$$P(f|D_{1:t}) \propto P(D_{1:t}|f)P(f).$$

2. Bayesian optimization it to find $\boldsymbol{x}_{t+1}$

$$\boldsymbol{x}_{t+1} = \arg\max_{\boldsymbol{x} \in R^d} \int a(\boldsymbol{x}) P(f_{t+1}|D_{1:t}) df_{t+1} \tag{1}$$

where $a(x)$ is called **acquisition function** that guides the search for the optimum that high acquisition corresponds to potentially high values of the objective function.

Bayesian Optimization II

- Common acquisition functions include **probability of improvement**, **expected improvement** and **entropy** (Kushner, 1964; Mockus et al., 1978; Jones, 2001; Cox and John, 1997; Brochu et al., 2010; Villemonteix et al., 2009)
- Recent work in Bayesian Optimization like Jones et al. (1998) Jones (2001), Bergstra and Bengio (2012) can trace back to Cox and John (1997).
- Bayesian Optimization is a popular approach in engineering but not well known in statistics.

Bayesian Optimization III

- We are interested in finding the maximum for the predictive surface

$$p(\tilde{y}_b|\tilde{y}_{-b}, x) = \int \prod_{i \in \mathcal{T}_b} p(y_i|\theta, x_i)p(\theta|\tilde{y}_{-b})d\theta,$$

- However in econometric time series this is much more complicated due to the decomposition

$$\begin{aligned} & p(Y_{(T+1):(T+p)}|Y_{1:T}, X) \\ & = \prod_{i=1}^{p} \int p(Y_{T+i}|\theta, Y_{1:(T+i-1)}, X_{T+i})p(\theta|Y_{1:(T+i-1)}, X_{1:(T+i-1)})d\theta. \end{aligned}$$

Bayesian modeling of $P(f|D_{1:t})$ I

- The function $f(x)$ is usually approximated by a Gaussian Process (Mockus, 1994; Sasena, 2002)

$$f(\mathbf{x}) = \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

  which is restrictive by $\mathcal{GP}$ itself because choosing the covariance function for the $\mathcal{GP}$ is crucial.
- We consider the **multivariate surface model** (Li and Villani, 2013) to model $f(x)$
  - The surface consists of three different components, *linear*, *surface* and *additive* as

$$\mathbf{Y} = \mathbf{X}_o \mathbf{B}_o + \mathbf{X}_s(\xi_s)\mathbf{B}_s + \mathbf{X}_a(\xi_a)\mathbf{B}_a + \mathbf{E}.$$

  - We treat the knots ξ_i as unknown parameters and let them move freely.

Bayesian modeling of $P(f|D_{1:t})$ II

- A model with a minimal number of free knots outperforms model with lots of fixed knots.
- For notational convenience, we sometimes write model in compact form

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E},$$

where $\mathbf{X} = [\mathbf{X}_o, \mathbf{X}_s, \mathbf{X}_a]$ and $\mathbf{B} = [\mathbf{B}_o{}', \mathbf{B}_s{}', \mathbf{B}_a{}']'$ and $\mathbf{E} \sim \mathbf{N}_p(\mathbf{0},\ \mathbf{\Sigma})$

The Efficient MCMC algorithm I

- The coefficients $(\mathbf{B})$ are directly sampled from normal distribution.
- We update covariance $(\boldsymbol{\Sigma})$, all knots $(\boldsymbol{\xi})$ and shrinkages $(\boldsymbol{\lambda})$ jointly by using Metropolis-Hastings within Gibbs.
- The proposal density for $\boldsymbol{\Sigma}$ is the inverse Wishart density on previous slide.
- The proposal density for $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$ is a multivariate t-density with $\nu > 2$ df,

$$\theta_p|\theta_c \sim \mathbf{MVT}\left[\hat{\theta},\ -\left(\frac{\partial^2 \ln p(\theta|\mathbf{Y})}{\partial\theta\partial\theta'}\right)^{-1}\Bigg|_{\theta=\hat{\theta}},\ \nu\right],$$

where $\hat{\theta}$ is obtained by R steps $(R \leqslant 3)$ Newton's iterations during the proposal with analytical gradients for matrices.

The Efficient MCMC algorithm II

- The Metropolis-Hastings acceptance probability is

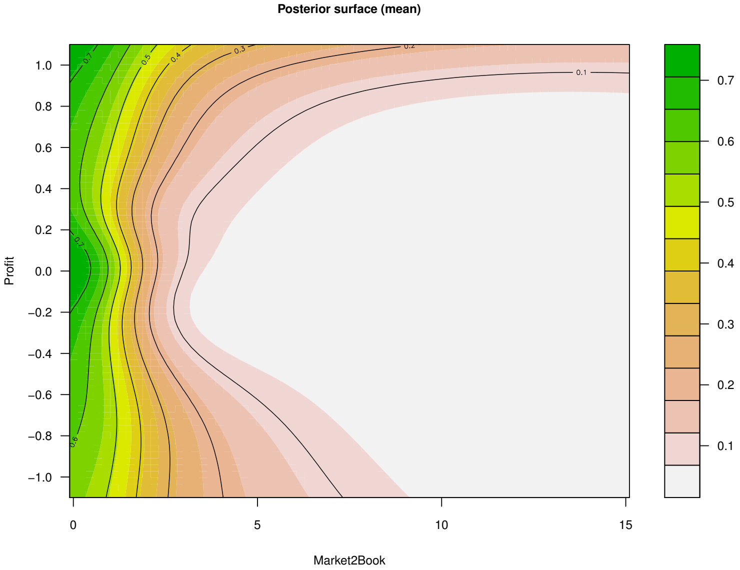
$$a\left(\theta_c \to \theta_p\right) = \min\left[1,\ \frac{p(\mathbf{Y}|\theta_p)p(\theta_p)g(\theta_c|\theta_p)}{p(\mathbf{Y}|\theta_c)p(\theta_c)g(\theta_p|\theta_c)}\right].$$

- The analytical gradients are very complicated and we have implemented it in an efficient way.
- Bayesian variable selection can be naturally applied in MCMC procedure.
- The MCMC implementations are straightforward.
- We allow the parameters to be updated via:
  - parallel mode for small datasets,
  - batched mode for big datasets.
- MCMC method allows us to evaluate the integral in Eq. (1) easily.

Optimizing the acquisition function

- With the posterior, a deterministic, derivative-free optimizer can then be used in optimizing the acquisition function (Jones et al., 1993; Mockus, 1994; Lizotte, 2008).
- By taking the advantage of MCMC, we may integrate the two steps together. *Working in progress*...

The firm leverage data, a revisit



Extensions

- Our approach can be applied experimental design.
- High dimensional response surfaces will be considered.

References I

Bergstra, J. and Bengio, Y. (2012), "Random search for hyper-parameter optimization," *The Journal of Machine Learning Research*, 13, 281–305.

Brochu, E., de Freitas, N., and Hoffman, M. (2010), "Hedging Strategies for Bayesian Optimization," Tech. rep.

Cox, D. D. and John, S. (1997), "SDO: A statistical method for global optimization," in *Multidisciplinary design optimization: state of the art*, SIAM, Philadelphia, pp. 315–329.

Jones, D. R. (2001), "A taxonomy of global optimization methods based on response surfaces," *Journal of global optimization*, 21, 345–383.

Jones, D. R., Perttunen, C. D., and Stuckman, B. E. (1993), "Lipschitzian optimization without the Lipschitz constant," *Journal of Optimization Theory and Applications*, 79, 157–181.

References II

Jones, D. R., Schonlau, M., and Welch, W. J. (1998), "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, 13, 455–492.

Kushner, H. J. (1964), "A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise," *Journal of Fluids Engineering*, 86, 97–106.

Li, F. and Villani, M. (2013), "Efficient Bayesian Multivariate Surface Regression," *Scandinavian Journal of Statistics*, 40, 706–723.

Lizotte, D. J. (2008), *Practical bayesian optimization*, University of Alberta.

Mockus, J. (1994), "Application of Bayesian approach to numerical methods of global and stochastic optimization," *Journal of Global Optimization*, 4, 347–365.

References III

Mockus, J., Tiesis, V., and Zilinskas, A. (1978), "The application of Bayesian methods for seeking the extremum," *Towards Global Optimization*, 2, 2.

Sasena, M. J. (2002), "Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations," Ph.D. thesis, University of Michigan.

Villemonteix, J., Vazquez, E., and Walter, E. (2009), "An informational approach to the global optimization of expensive-to-evaluate functions," *Journal of Global Optimization*, 44, 509–534.

Thank you!