Efficient Bayesian Multivariate Surface Regression

Feng Li
feng.li@cufe.edu.cn

School of Statistics and Mathematics
Central University of Finance and Economics

Outline of the talk

1. Introduction to flexible regression models
2. The multivariate surface model
3. Application to firm leverage data
4. Extensions and future work

Flexible regression models
↪ Introduction

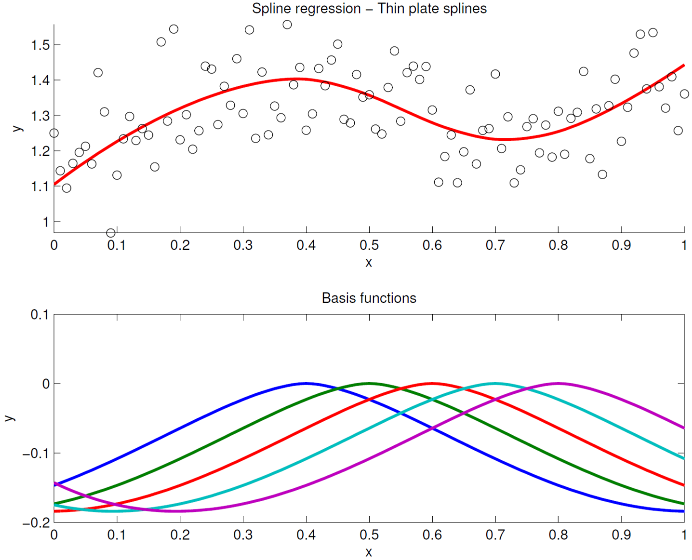
- Flexible models of the regression function $E(y|x)$ has been an active research field for decades.
- Attention has shifted from kernel regression methods to spline-based models.
- Splines are regression models with flexible mean functions.
- Example: a simple spline regression with only one explanatory variable with truncated linear basis function can be like this

$$y = \alpha_0 + \alpha_1 x + \beta_1 (x - \xi_1)_+ + \ldots + \beta_q (x - \xi_q)_+ + \varepsilon$$

  where
  - $(x - \xi_i)_+$ are called the basis functions,
  - ξ_i are called knots (the location of the basis function).

Flexible regression models

↪ Spline example (single covariate with thinplate bases)



Flexible regression models

↪ Spline regression with multiple covariates

- Additive spline model
  - Each knot $\xi_{j.}$ (scaler) is connected with only one covariate

$$y = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_q x_q + \left[\sum_{j_1=1}^{m_1} \beta_{j_1} f\left(x_1, \xi_{j_1}\right) + \ldots + \sum_{j_q=1}^{m_q} \beta_{j_q} f\left(x_q, \xi_{j_q}\right) \right] + \varepsilon$$

  - Good and simple if you know there is no interactions in the data a priori.
- Surface spline model
  - Each knot $\boldsymbol{\xi}_j$ (vector) is connected with more than one covariate

$$y = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_q x_q + \left[\sum_{j=1}^{m} \beta_j g\left(x_1, \ldots x_q, \boldsymbol{\xi}_j\right) \right] + \varepsilon$$

  - A popular choice of $g\left(x_1, \ldots x_q, \boldsymbol{\xi}_j\right)$ can be e.g. the multi-dimensional thinplate spline

$$g\left(x_1, \ldots x_q, \boldsymbol{\xi}_j\right) = \left\|\mathbf{x} - \boldsymbol{\xi}_j\right\|^2 \ln \left\|\mathbf{x} - \boldsymbol{\xi}_j\right\|$$

  - Can handle the interactions but the model complexity increase dramatically with the interactive knots.

The challenges

- How many knots are needed?
  - Too few knots lead to a bad approximation; too many knots yield overfitting.
- Where to place those knots?
  - Equal spacing for the additive model,
  - which is obviously not efficient with the surface model.
- Common approaches to the two problems:
  - place enough many knots and use variable selection to pick up useful ones.
    - not truly flexible
  - use reversible jump MCMC to move among the model spaces with different numbers of knots
    - very sensitive to the prior and not computational efficient
  - clustering the covariates to select knots
    - does not use the information from the responses
- How to choose between additive spline and surface spline?
  - NA

- The multivariate surface model 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.
  - 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 $\mathbf{E} \sim \mathbf{N}_p(\mathbf{0},\ \boldsymbol{\Sigma})$

The multivariate surface model
↛ The prior

- Conditional on the knots, the prior for $\mathbf{B}$ and $\mathbf{\Sigma}$ are set as

$$\text{vec}\mathbf{B}_i|\mathbf{\Sigma},\ \lambda_i \sim \mathbf{N}_q\left[\mu_i,\ \mathbf{\Lambda}_i^{1/2}\mathbf{\Sigma}\mathbf{\Lambda}_i^{1/2}\otimes\mathbf{P}_i^{-1}\right],\ i\in\{o,s,a\},$$
$$\mathbf{\Sigma}\sim\mathbf{IW}\left[n_0\mathbf{S}_0,\ n_0\right],$$

  - $\mathbf{\Lambda}_i = diag(\boldsymbol{\lambda}_i)$ are called the shrinkage parameters, which is used for overcome overfitting through the prior.
  - If $\mathbf{P}_i = \mathbf{I}$, can prevent singularity problem, like the ridge regression estimate.
  - If $\mathbf{P}_i = \mathbf{X}_i'\mathbf{X}_i$: use the covariates information, also a compressed version of least squares estimate when $\boldsymbol{\lambda}_i$ is large.
- The shrinkage parameters are estimated in MCMC
  - A small $\boldsymbol{\lambda}_i$ shrinks the variance of the conditional posterior for $\mathbf{B}_i$
  - It is another approach to selection important variables (knots) and components.
- We allow to mixed use the two types priors ($\mathbf{P}_i = \mathbf{I}$, $\mathbf{P}_i = \mathbf{X}_i'\mathbf{X}_i$) in different components in order to take the both the advantages of them.

The multivariate surface model
↛ The Bayesian posterior

- The posterior distribution is conveniently decomposed as

$$p(\mathbf{B}, \boldsymbol{\Sigma}, \xi, \lambda|\mathbf{Y}, \mathbf{X}) = p(\mathbf{B}|\boldsymbol{\Sigma}, \xi, \lambda, \mathbf{Y}, \mathbf{X})p(\boldsymbol{\Sigma}|\xi, \lambda, \mathbf{Y}, \mathbf{X})p(\xi, \lambda|\mathbf{Y}, \mathbf{X}).$$

- Hence $p(\mathbf{B}|\boldsymbol{\Sigma}, \xi, \lambda, \mathbf{Y}, \mathbf{X})$ follows the multivariate normal distribution according to the conjugacy;
- When $p = 1$, $p(\boldsymbol{\Sigma}|\xi, \lambda, \mathbf{Y}, \mathbf{X})$ follows the inverse Wishart distribution

$$\mathbf{IW}\left[n_0 + n, \left\{n_0 \mathbf{S}_0 + n\tilde{\mathbf{S}} + \sum_{i \in \{o,s,a\}} \boldsymbol{\Lambda}_i^{-1/2}(\tilde{\mathbf{B}}_i - \mathbf{M}_i)'\mathbf{P}_i(\tilde{\mathbf{B}}_i - \mathbf{M}_i)\boldsymbol{\Lambda}_i^{-1/2}\right\}\right]$$

- When $p \geqslant 2$, no closed form of $p(\boldsymbol{\Sigma}|\xi, \lambda, \mathbf{Y}, \mathbf{X})$, the above result is a very accurate approximation. Then the marginal posterior of $\boldsymbol{\Sigma}$, ξ and λ is

$$\begin{aligned} p\left(\boldsymbol{\Sigma}, \xi, \lambda|\mathbf{Y}, \mathbf{X}\right) =& c \times p(\xi, \lambda) \times |\boldsymbol{\Sigma}_{\beta}|^{-1/2}|\boldsymbol{\Sigma}|^{-(n+n_0+p+1)/2}|\boldsymbol{\Sigma}_{\tilde{\beta}}|^{-1/2} \\ & \times \exp\left\{-\frac{1}{2}\left[\mathrm{tr}\boldsymbol{\Sigma}^{-1}\left(n_0\mathbf{S}_0 + n\tilde{\mathbf{S}}\right) + \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu}\right)'\boldsymbol{\Sigma}_{\beta}^{-1}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\mu}\right)\right]\right. \end{aligned}$$

The MCMC algorithm
$\hookrightarrow$ Metropolis-Hastings within Gibbs

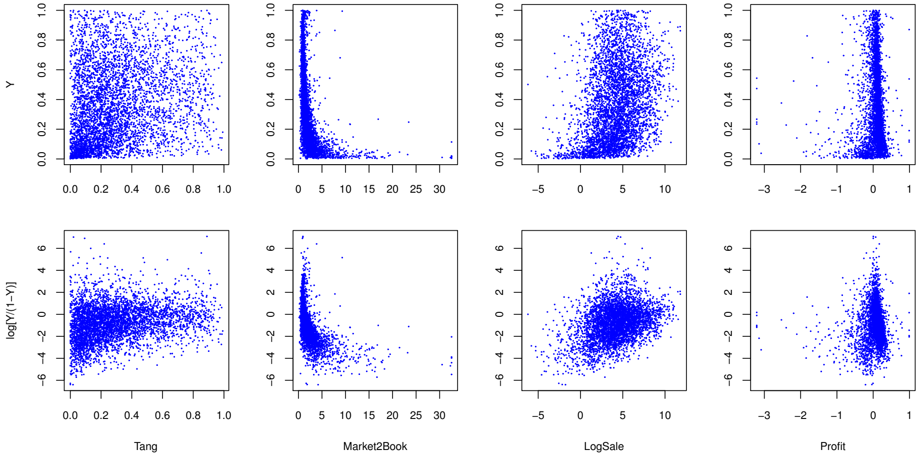
- The coefficients $(\mathbf{B})$ are directly sampled from normal distribution.
- We update covariance $(\boldsymbol{\Sigma})$, all knots $(\boldsymbol{\xi})$ and shrinkages $(\boldsymbol{\lambda})$ jointly by using Metropolis-Hastings within Gibbs.
- The proposal density for $\boldsymbol{\Sigma}$ is 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,

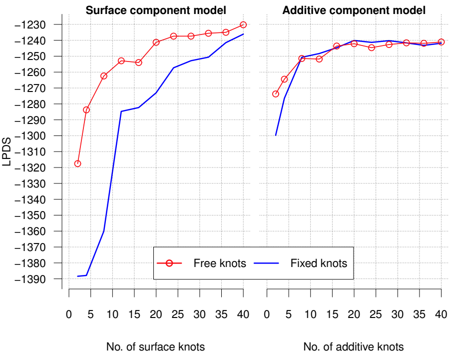
$$\theta_p|\theta_c \sim \mathbf{MVT}\left[\hat{\theta},\ -\left(\frac{\partial^2 \ln p(\theta|\mathbf{Y})}{\partial\theta\partial\theta'}\right)^{-1}\Bigg|_{\theta=\hat{\theta}},\ \nu\right],$$

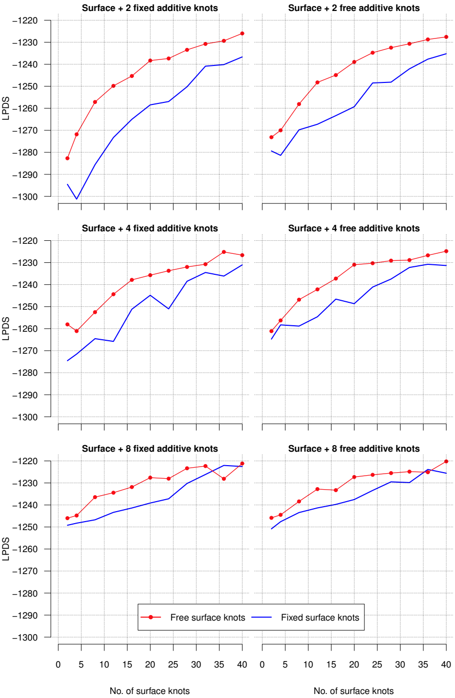
  where $\hat{\theta}$ is obtained by R steps ($R \leqslant 3$) Newton's iterations during the proposal with analytical gradients for matrices.
- The analytical gradients are very complicated and we have implemented it in an efficient way (**the key!**).

Application to firm leverage data
↪ The 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.



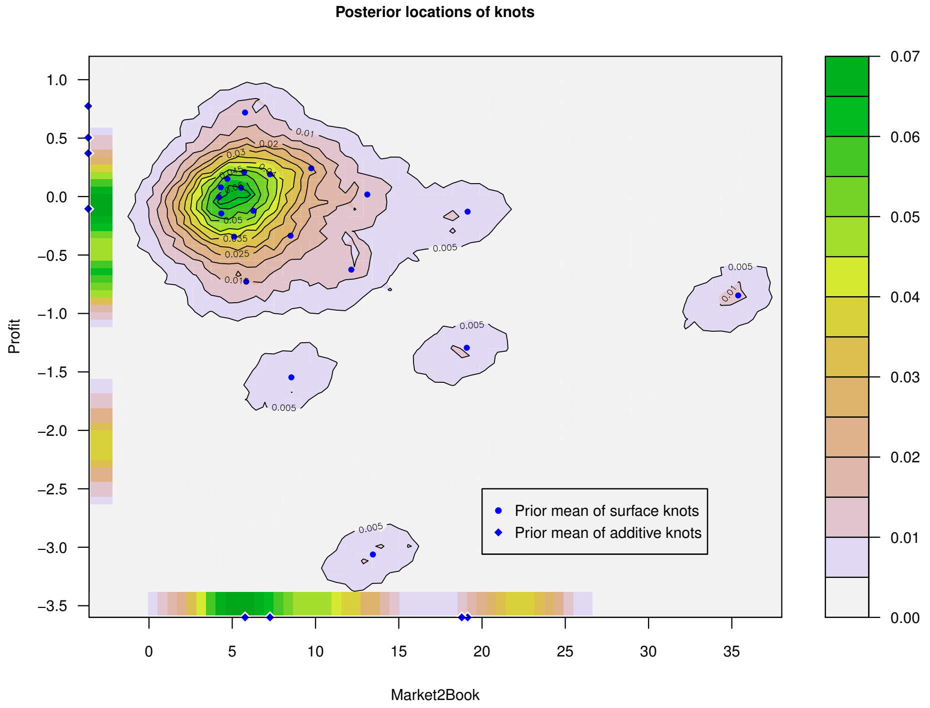




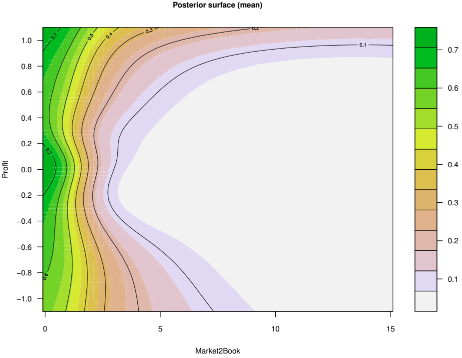
- $\uparrow$ Models with only surface or additive components
- $\rightarrow$ Model with both additive and surface components.
- **LPDS** Log predictive density score which is defined as

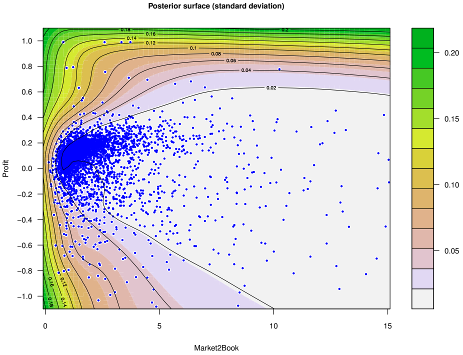
$$
\begin{aligned}
\mathrm{LPDS} &= \frac{1}{D}\sum_{d=1}^{D} \ln p(\tilde{\mathbf{Y}}_d \mid \tilde{\mathbf{Y}}_{-d}, \mathbf{X}) \\
&= \int \prod_{i \in \tau_d} p(\mathbf{y}_i \mid \boldsymbol{\theta}, \mathbf{x}_i)\, p(\boldsymbol{\theta} \mid \tilde{\mathbf{Y}}_{-d})\, d\boldsymbol{\theta},
\end{aligned}
$$

and $D = 5$ in the cross-validation.


Posterior locations of knots
Profit
Market2Book
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
−3.5
0
5
10
15
20
25
30
35
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0.005
0.01
0.02
0.03
0.025
0.035
0.05
Prior mean of surface knots
Prior mean of additive knots

Application to firm leverage data
↪ Posterior mean surface(left) and standard deviation(right)





Extensions and future work

- The model and the methods we used are very general.
- It is easy to generalize the model to GLM framework.
- Variable selection is possible for knots.
- Dirichlet precess prior can be plugged into the model when heteroscedasticity is the problem.
- And the copula...

Thank you!