Spline Methods

Feng Li

feng.li@cufe.edu.cn

School of Statistics and Mathematics
Central University of Finance and Economics

Revision: November 7, 2013

Today we are going to talk about...

1. Piecewise Polynomials
2. Avoiding knots selection with smoothing splines
3. Multi-dimensional splines
4. Discussions

Piecewise Polynomials

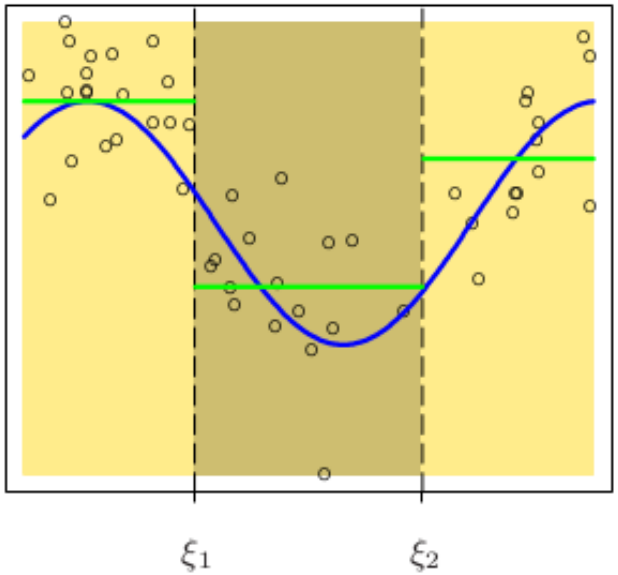
- Consider the regression model with only y and X
- Let

$$h_i(X) = (X - \xi_i)_+ = \begin{cases} X - \xi_i, & X > \xi_i \\ 0, & elsewhere \end{cases}$$

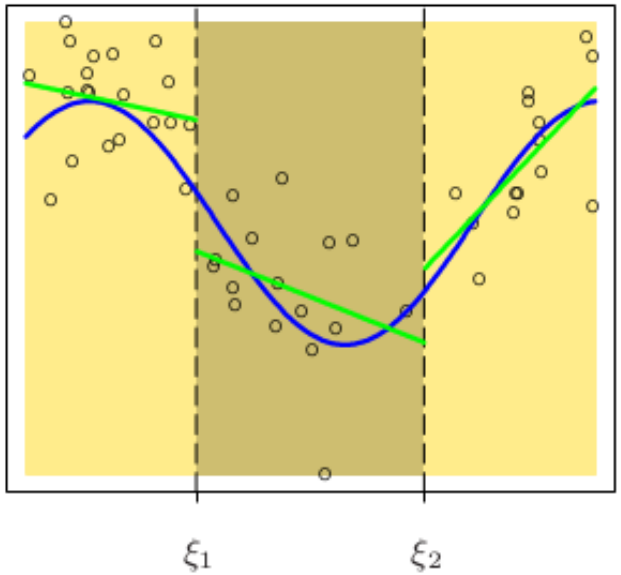
  for $i = 1, ..., p$
- Then set up a regression model

$$y = \beta_0 + \beta_1 X + \alpha_1 h_1(X) + ... + +\alpha_p h_p(X) + \epsilon$$

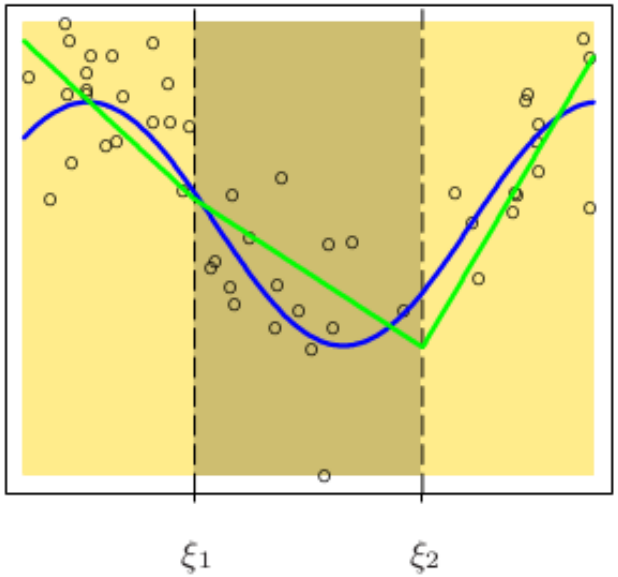
Piecewise Constant



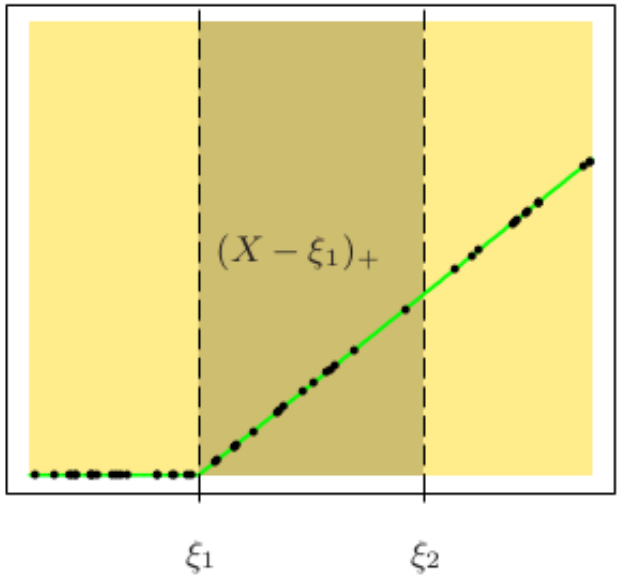
Piecewise Linear



Continuous Piecewise Linear



Piecewise-linear Basis Function



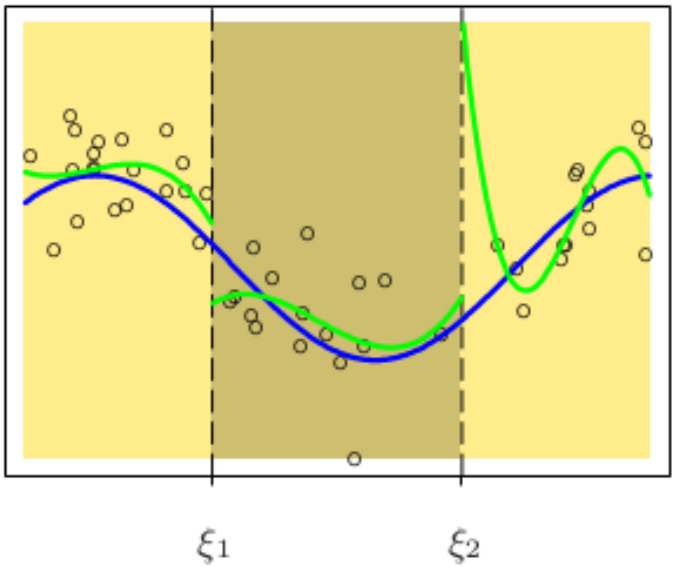
Higher order piecewise polynomials

- Let

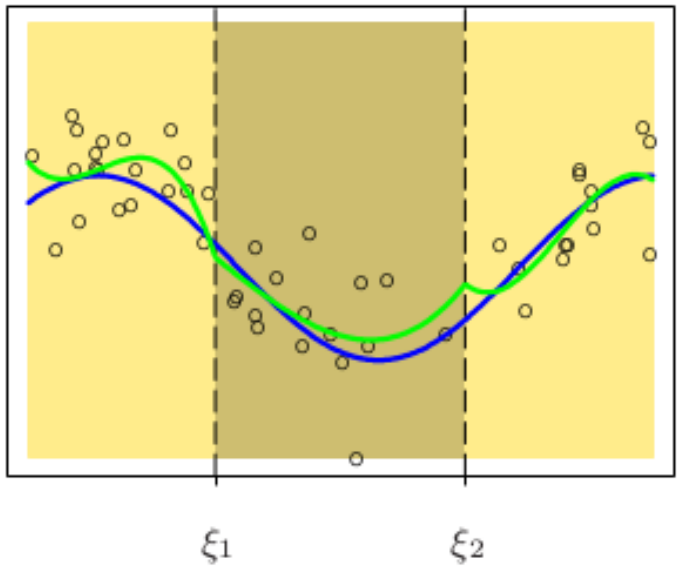
$$h_j = (X) = X^{j-1}, j = 1, ..., M$$
$$h_{M+l}(X) = (X - \xi_l)_+^{M-1}, l = 1, ..., K$$

- Then use all $h()$ with Y to setup a regression model.
- Terminologies: **basis functions**, **knots**, **knots locations**.
- **Regression splines**: when the knots are fixed

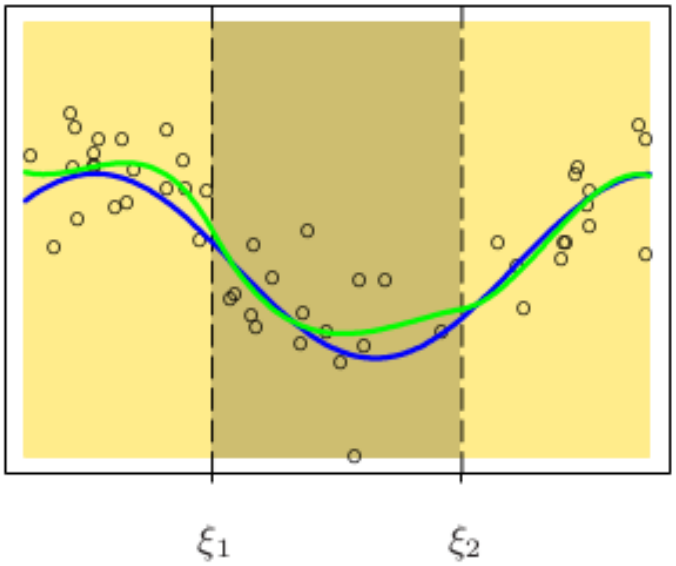
Discontinuous



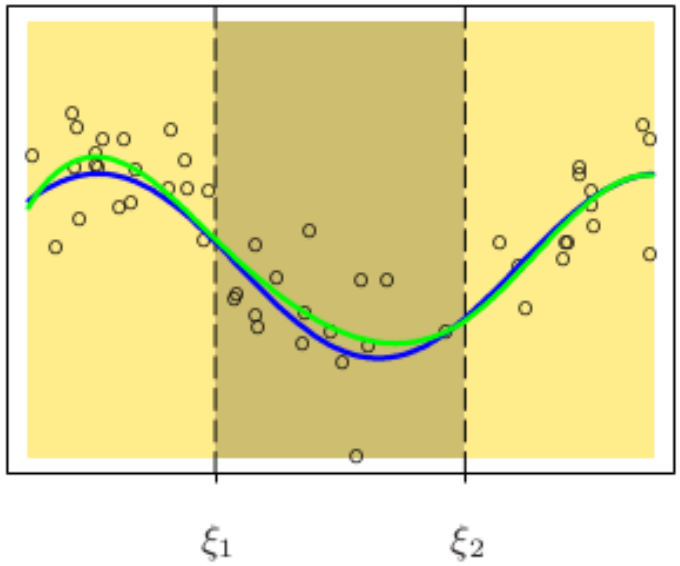
Continuous



Continuous First Derivative



Continuous Second Derivative



Natural cubic splines

- Define

$$d_k(X) = \frac{(X - \xi_k)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi - k}$$

- And the spline is defined as

$$h_k(X) = d_k(X) - d_{K-1}(X)$$

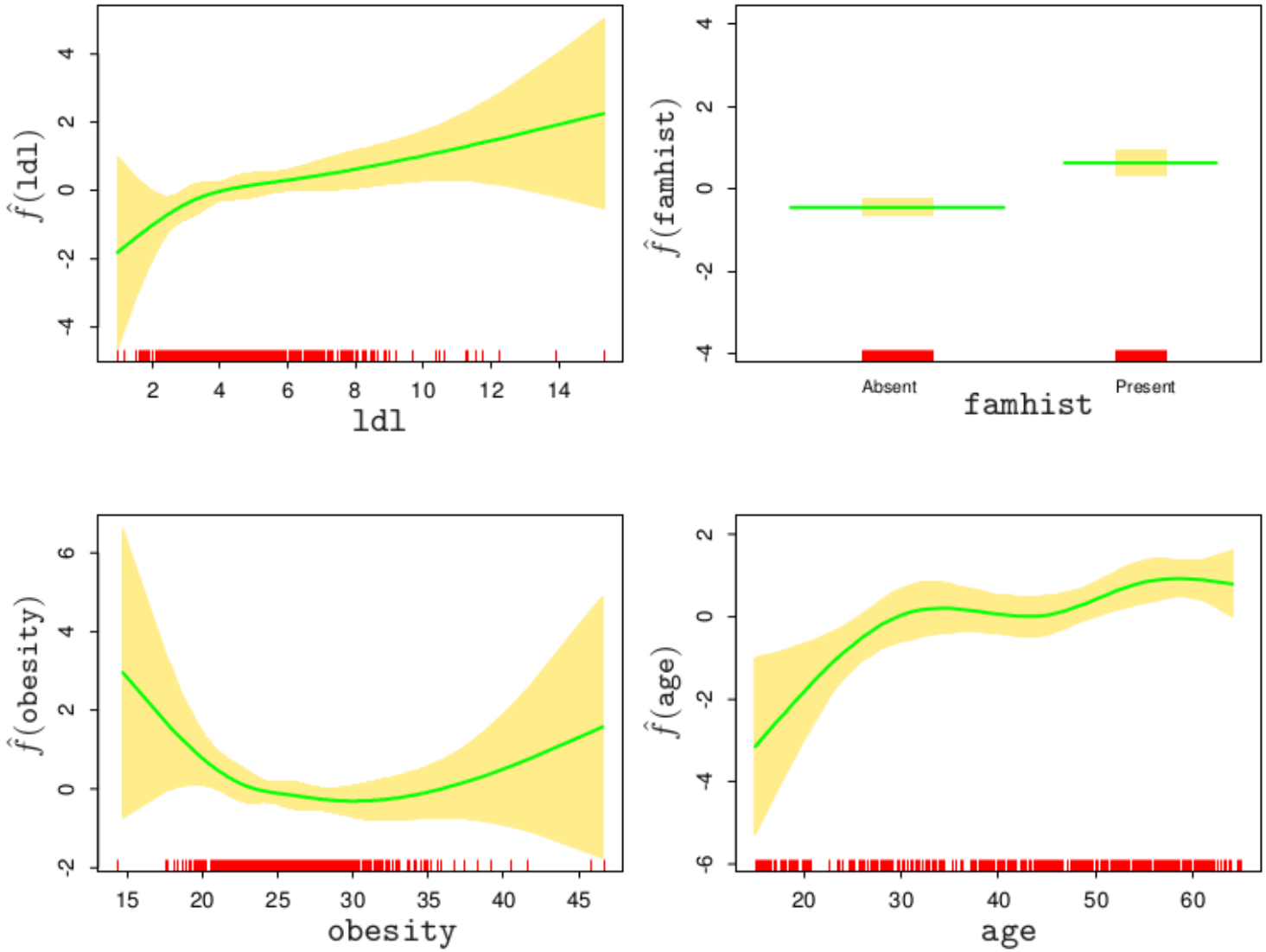


FIGURE 5.4. *Fitted natural-spline functions for each of the terms in the final model selected by the stepwise procedure. Included are pointwise standard-error bands. The* rug plot *at the base of each figure indicates the location of each of the sample values for that variable (jittered to break ties).*

B-splines

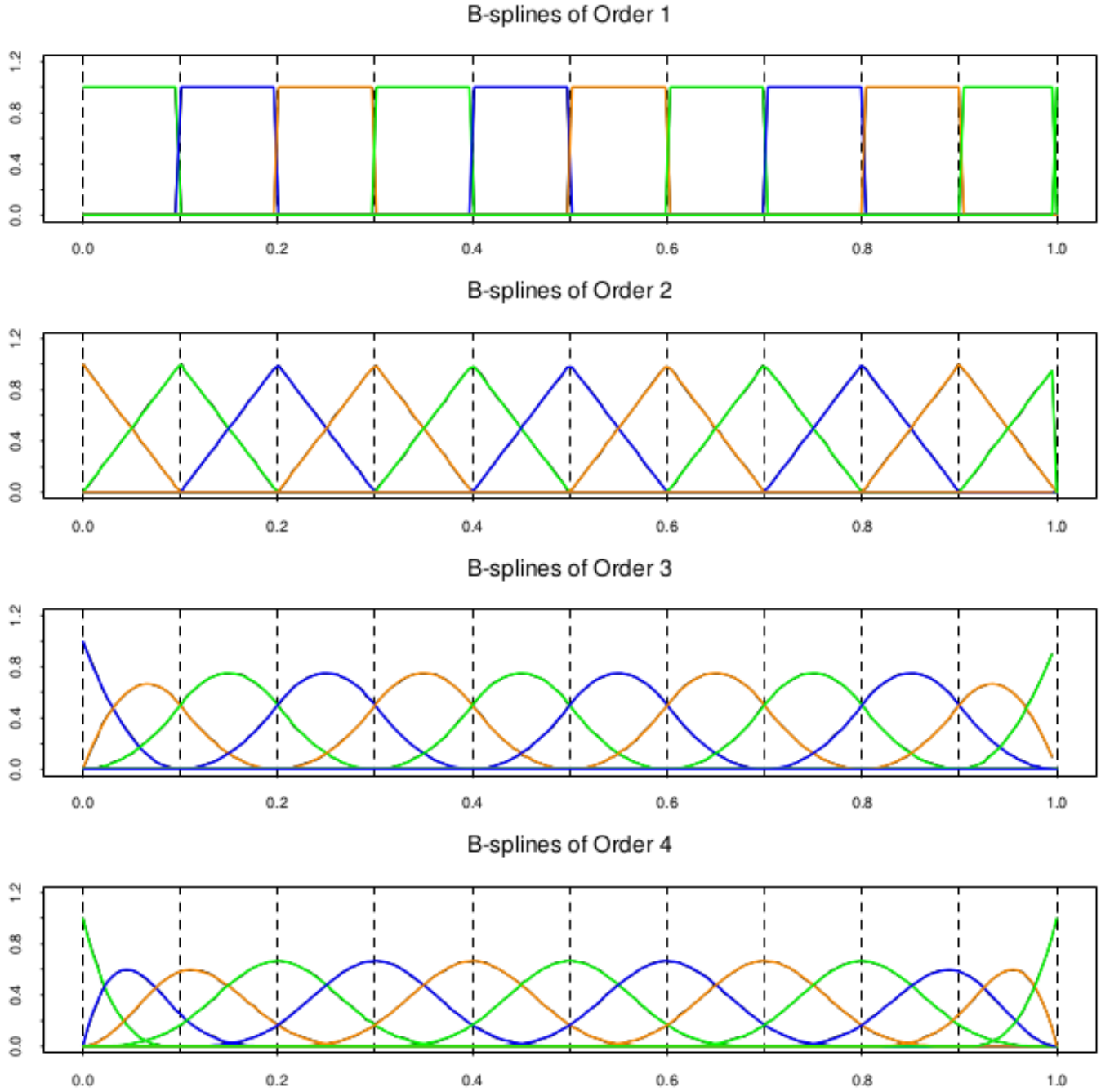
- Assume we have the two boundary knots $\xi_0 < \xi_1$ and $\xi_K < \xi_{K+1}$.
- We define a knot sequence τ such that

$$\tau_1 \leqslant \tau_2 \leqslant ... \leqslant \tau_M \leqslant \xi_o$$
$$\tau_{j+M} = \xi_j, j = 1, ..., K$$
$$\xi_{K+1} \leqslant \tau_{K+M+1} \leqslant \tau_{K+M+2} \leqslant ...\tau_{K+2M}$$

- Let $B_{i,m}(x)$ the ith B-spline function of order $m < M$ for the knot-sequence τ.

$$B_{i,1}(x) = \begin{cases} 1, & \tau_i \leqslant x \leqslant t_{i+1} \\ 0, otherwise \end{cases}$$
$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+i}} B_{i+1,m-1}(x)$$

- Properties of B-spline
  - A B-spline is a continuous function at the knots.
  - Any spline function of degree k on a given set of knots can be expressed as a linear combination of B-splines.


B-splines of Order 1
B-splines of Order 2
B-splines of Order 3
B-splines of Order 4
0.0
0.4
0.8
1.2
0.0
0.2
0.4
0.6
0.8
1.0

Avoiding knots selection with smoothing splines

- The **smoothing spline** is to minimize

$$\mathrm{RSS}(f,\lambda) = \sum_{i=1}^{N}\{y_i - f(x_i)\}^2 + \lambda\int\{f''(t)\}^2 dt$$

  where λ is the **smoothing parameter**

  - $\lambda = 0$ the usual spline fitting with not penalty.
  - $\lambda \to \infty$ the curve is moving from rough to very smooth till a regression line without knots (very heavy penalty)

Smoothing example with natural cubic splines

- For natural cubic splines

$$f(x) = \sum_{j=1}^{K} \theta_j h_j(x)$$

- The RSS is now as

$$RSS(\theta, \lambda) = \sum_{i=1}^{N} \{y_i - \sum_{j=1}^{K} h_j(x_i)\}^2 + \lambda \int \left\{ \frac{\partial^2 \sum_{j=1}^{K} h_j(t_i)}{\partial t^2} \right\}^2 dt$$

- And the solution is

$$\hat{\theta} = \left(N'N + \lambda \int \{f''(t)\}^2 dt \right)^{-1} N'y$$

  where N is the design matrix with all the data and basis functions.
- And the degree of freedom (no. of free parameters) is obtained through the trace of the hat matrix.

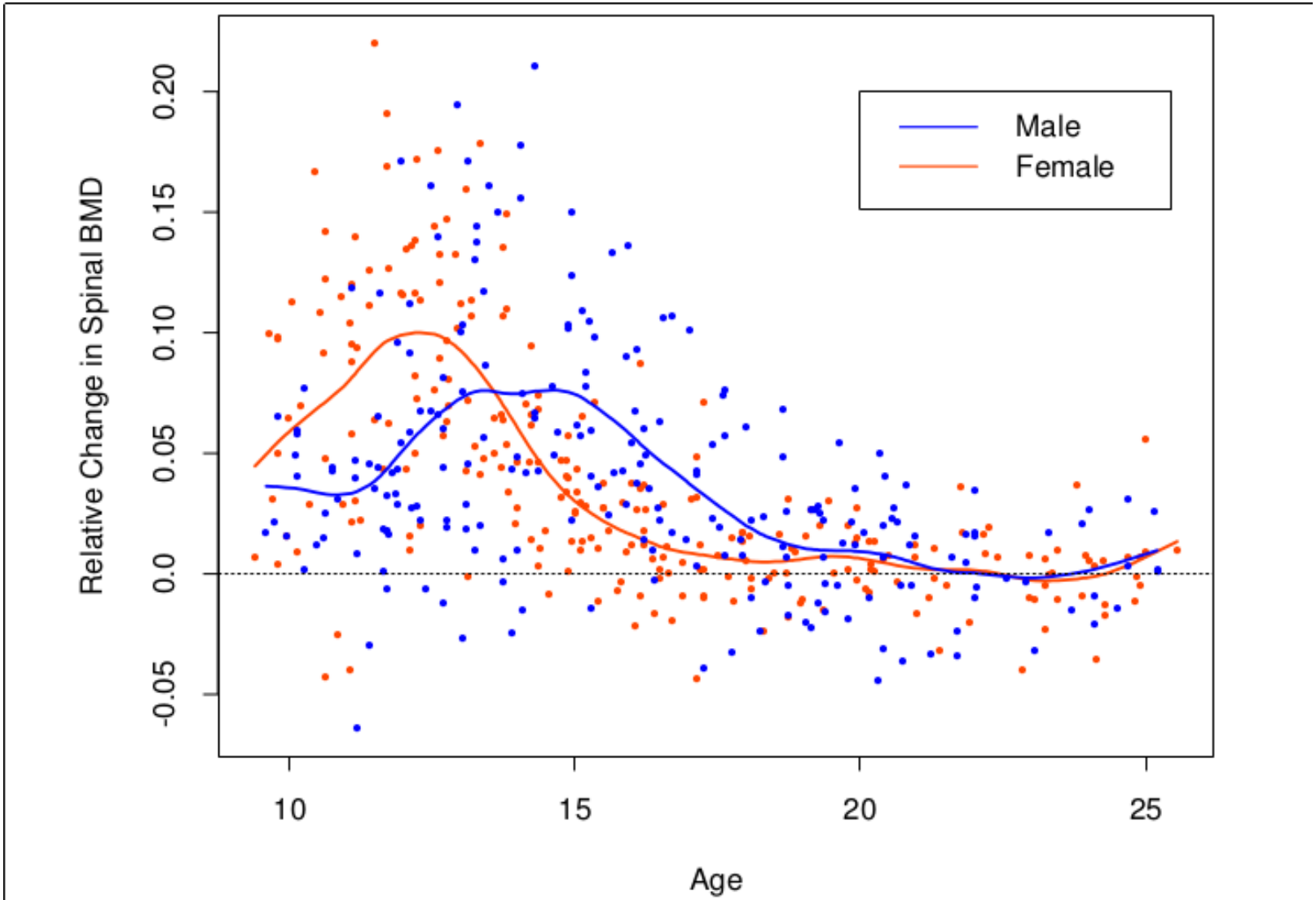


FIGURE 5.6. *The response is the relative change in bone mineral density measured at the spine in adolescents, as a function of age. A separate smoothing spline was fit to the males and females, with* $\lambda \approx 0.00022$. *This choice corresponds to about* 12 *degrees of freedom.*

Spline methods in logistic regression

- Recall the logistic regression

$$\log \frac{\Pr(G = k|X = x)}{\Pr(G = K|X = x)} = f(x)$$

  where $k = 1, 2, ..., K - 1$.
- Splines can also be used in $f(x)$.
- Need maximum likelihood method to obtain $\hat{\beta}$.
- Newton-Raphson algorithm is exactly of the same.

Multi-dimensional splines

- All the cases we considered are univariate splines.
- Multivariate splines are not so rich.
- People usually use thinplate splines

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

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

Discussions

- How do you choose from different splines?
- How do we avoid overfitting in spline method?
- How to apply shrinkage methods like LASSO in splines?
- How to choose λ with smoothing splines
- Do we obtain unbiased estimators in spline methods?
- What is the bias-variance trade off?

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

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

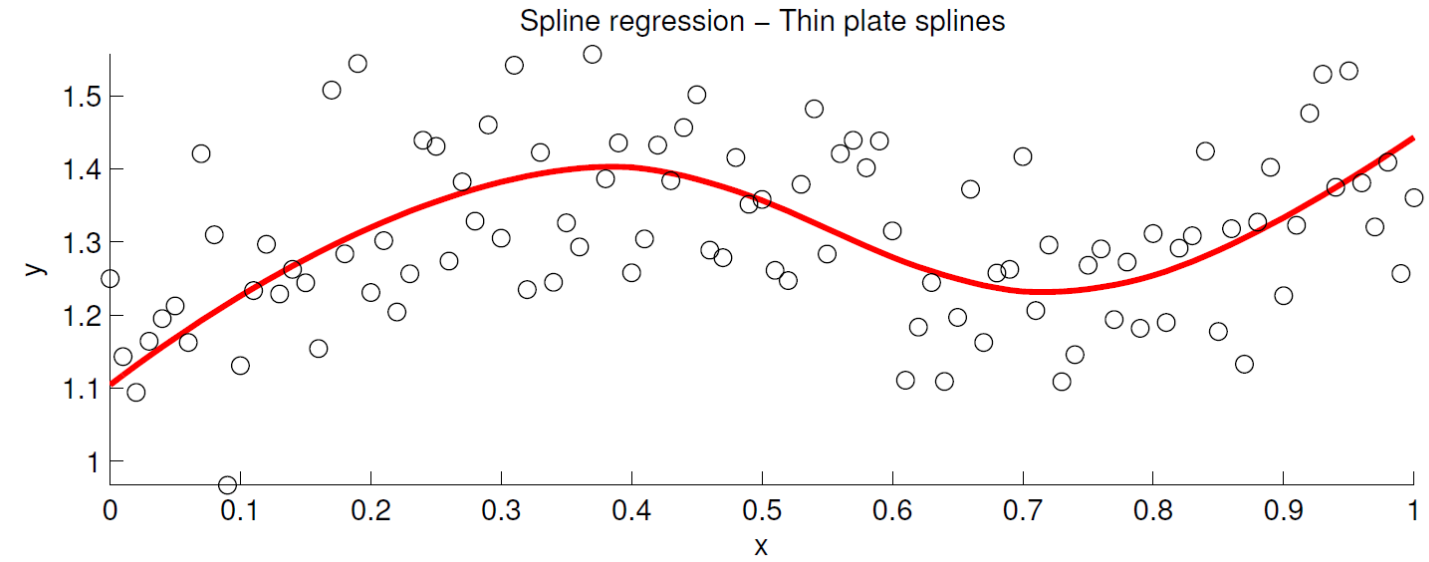
$$y = \alpha_0 + \alpha_1 x + \beta_1 (x - \xi_1)_+ + ... + \beta_q (x - \xi_q)_+ + \varepsilon$$

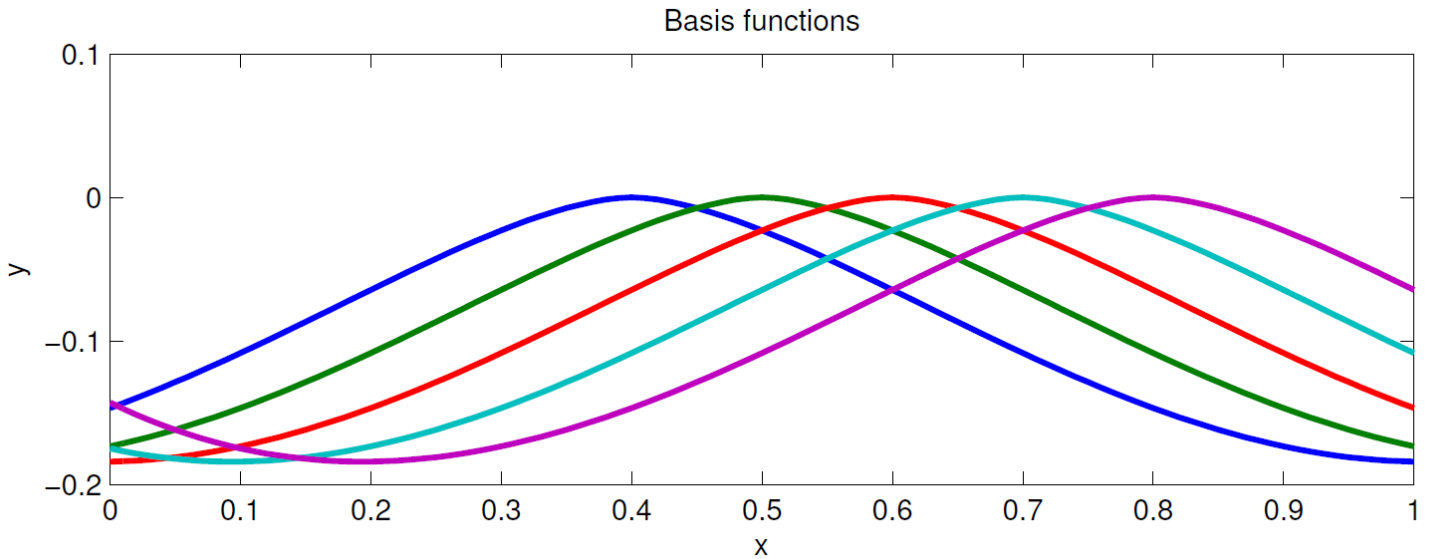
where

  - $(x - \xi_i)_+$ 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

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

  - 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

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

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

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

  - 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
↛ The model

- 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 $\mathbf{E} \sim \mathbf{N}_p(\mathbf{0},\ \mathbf{\Sigma})$

The multivariate surface model

↛ The prior

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

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

  - $\mathbf{\Lambda}_i = diag(\boldsymbol{\lambda}_i)$ 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

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

- Hence $p(\mathbf{B}|\boldsymbol{\Sigma}, \xi, \boldsymbol{\lambda}, \mathbf{Y}, \mathbf{X})$ follows the multivariate normal distribution according to the conjugacy;
- When $p = 1$, $p(\boldsymbol{\Sigma}|\xi, \boldsymbol{\lambda}, \mathbf{Y}, \mathbf{X})$ 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 $p(\boldsymbol{\Sigma}|\xi, \boldsymbol{\lambda}, \mathbf{Y}, \mathbf{X})$, the above result is a very accurate approximation. Then the marginal posterior of $\boldsymbol{\Sigma}$, ξ and $\boldsymbol{\lambda}$ is

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

The MCMC algorithm
↛ Metropolis-Hastings within Gibbs

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

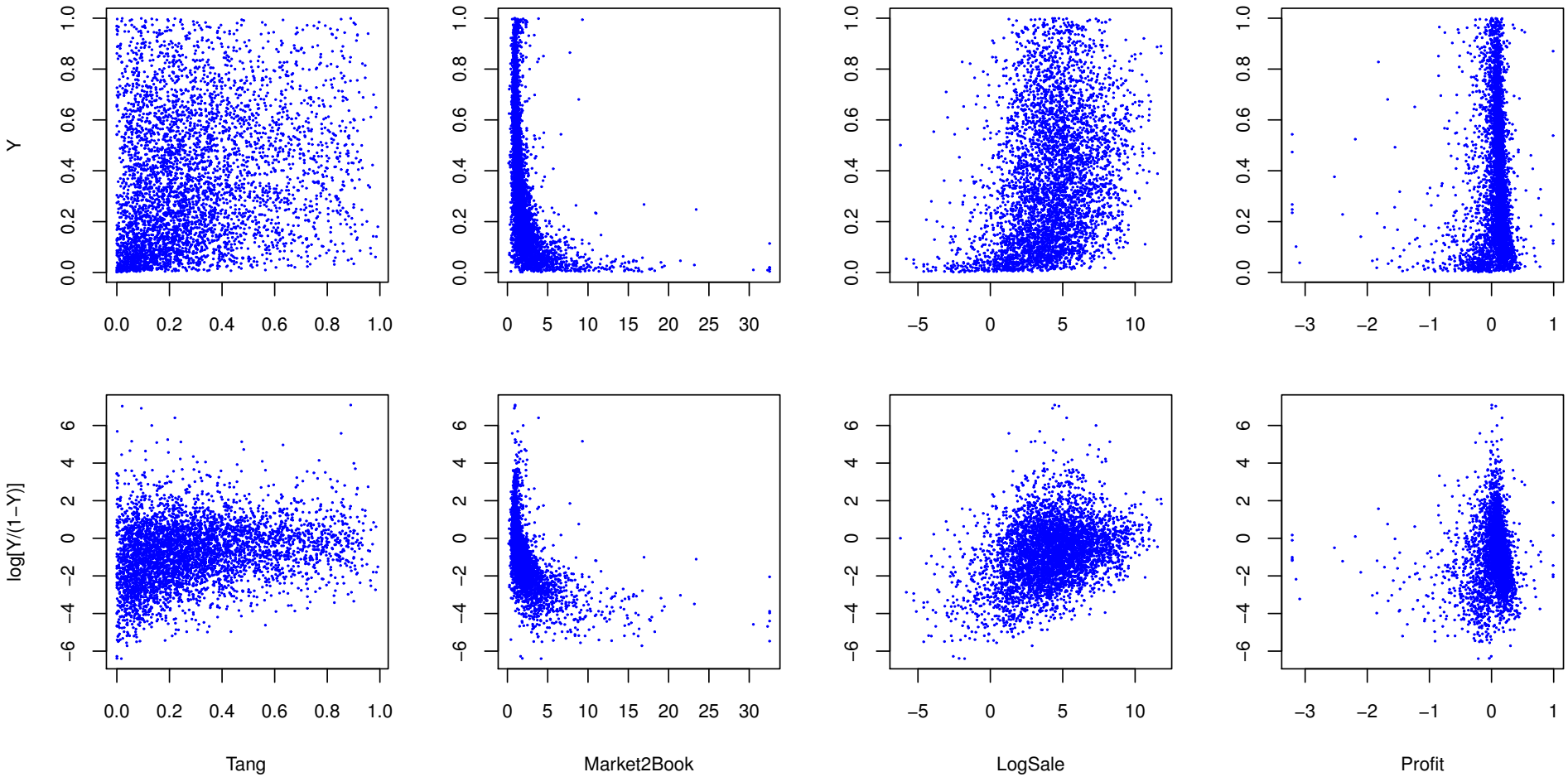
$$\theta_p|\theta_c \sim \mathbf{MVT}\left[\hat{\theta},\ -\left(\frac{\partial^2 \ln p(\theta|\mathbf{Y})}{\partial\theta\partial\theta'}\right)^{-1}\Bigg|_{\theta=\hat{\theta}},\ \nu\right],$$

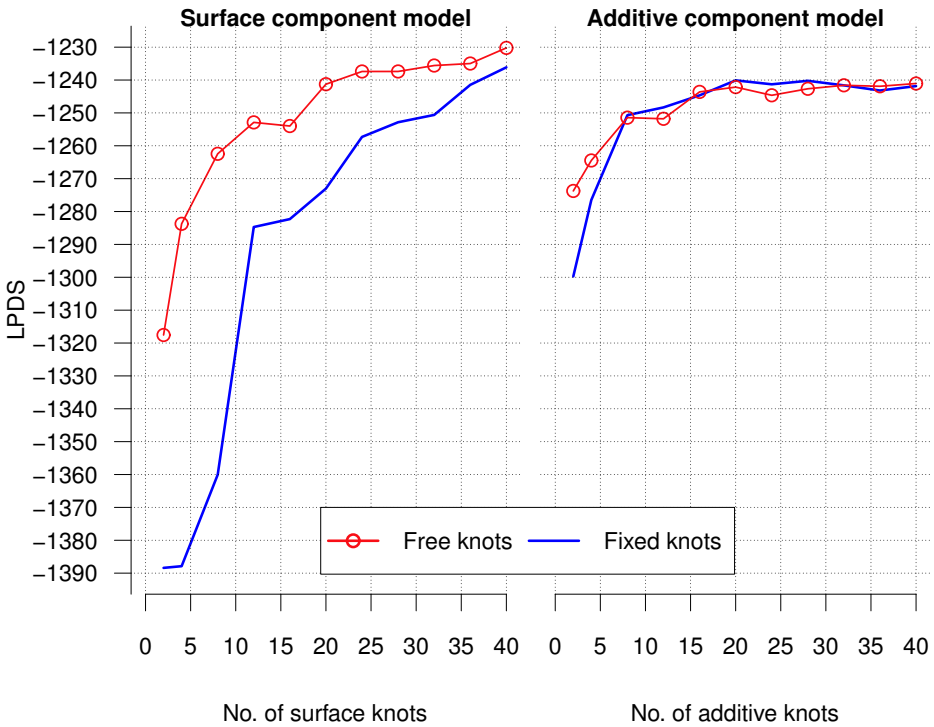
  where $\hat{\theta}$ is obtained by R steps ($R \leqslant 3$) Newton's iterations during the proposal with analytical gradients for matrices.
- The analytical gradients are very complicated and we have implemented it in an efficient way (**the key!**).

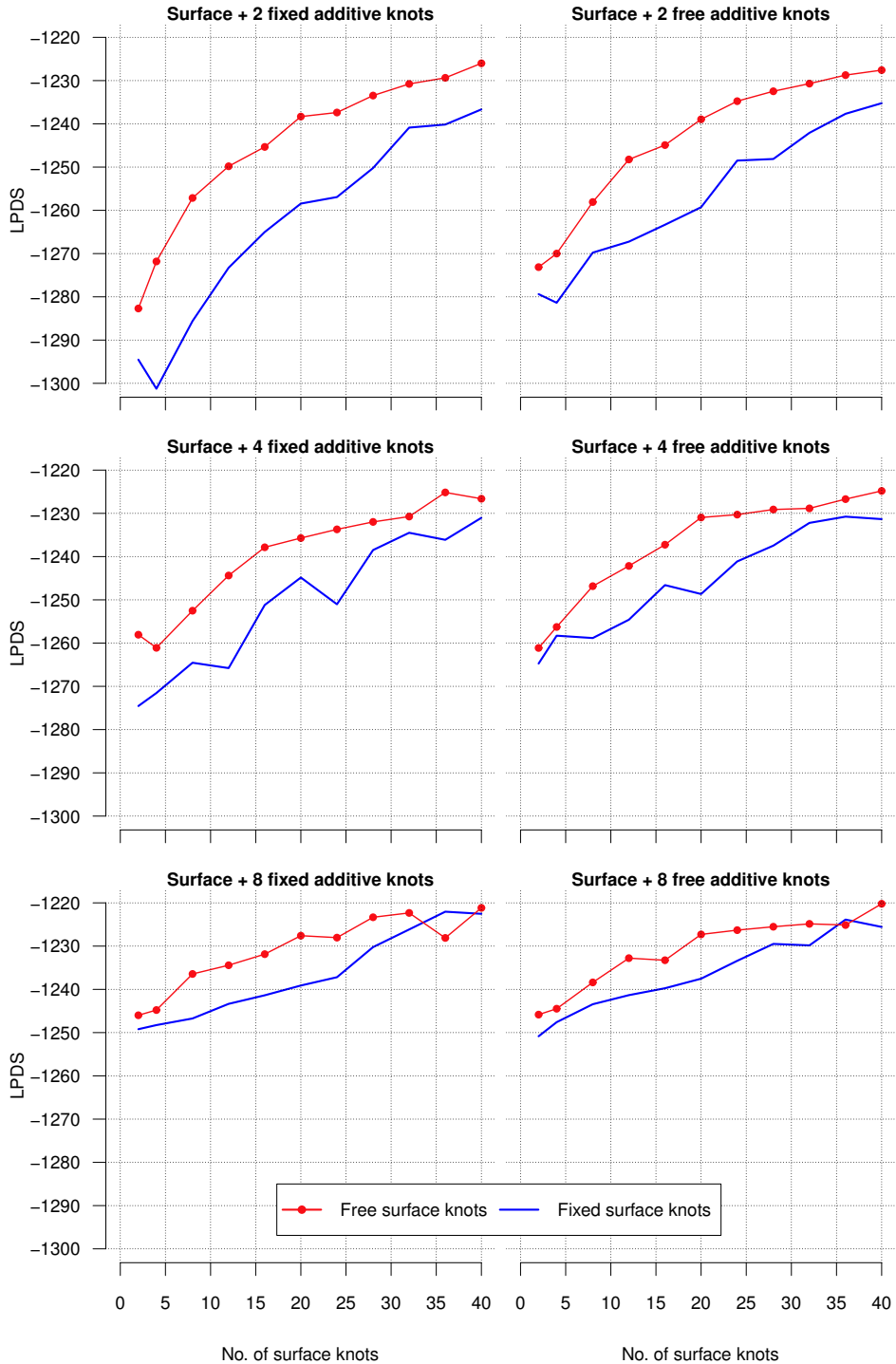
Application to firm leverage data

↛ 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.







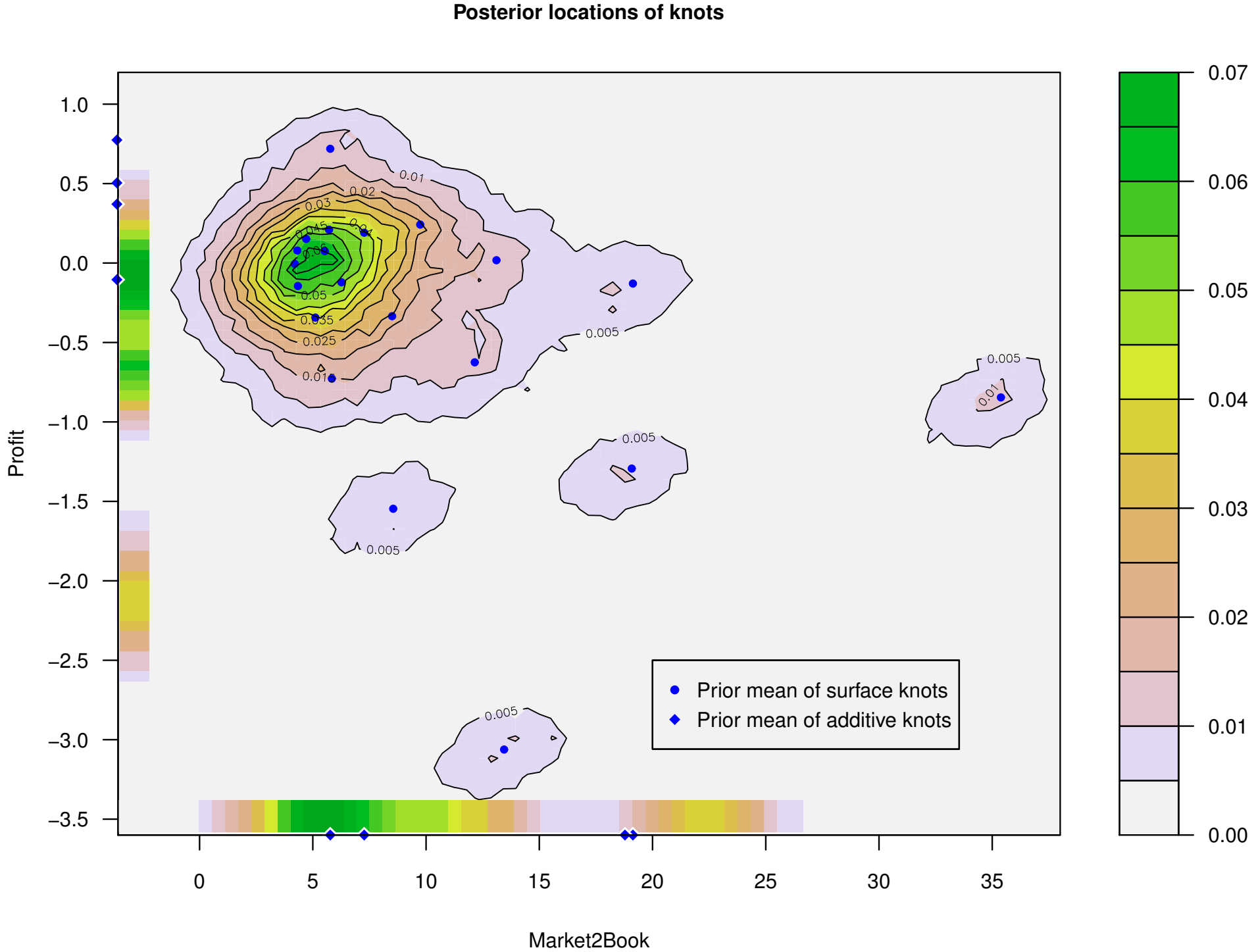
↑ Models with only surface or additive components

→ Model with both additive and surface components.

LPDS Log predictive density score which is defined as

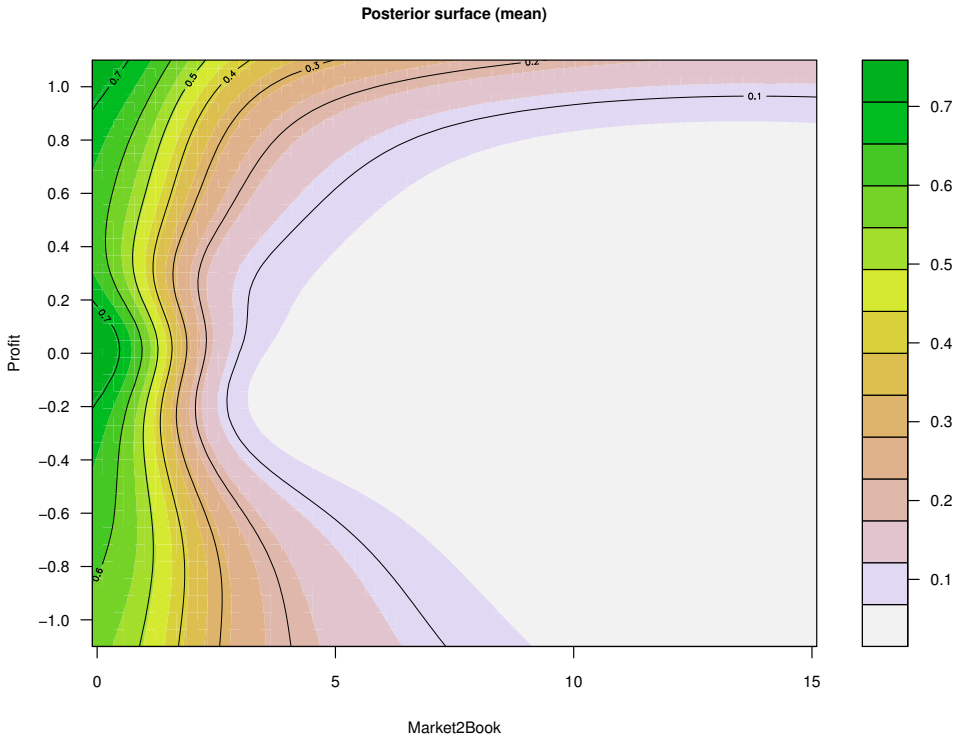
$$\begin{aligned} \mathrm{LPDS} &= \frac{1}{D}\sum_{d=1}^{D} \ln p(\tilde{Y}_d \mid \tilde{Y}_{-d}, X) \\ &= \int \prod_{i \in \tau_d} p(y_i \mid \theta, x_i)\, p(\theta \mid \tilde{Y}_{-d})\, d\theta, \end{aligned}$$

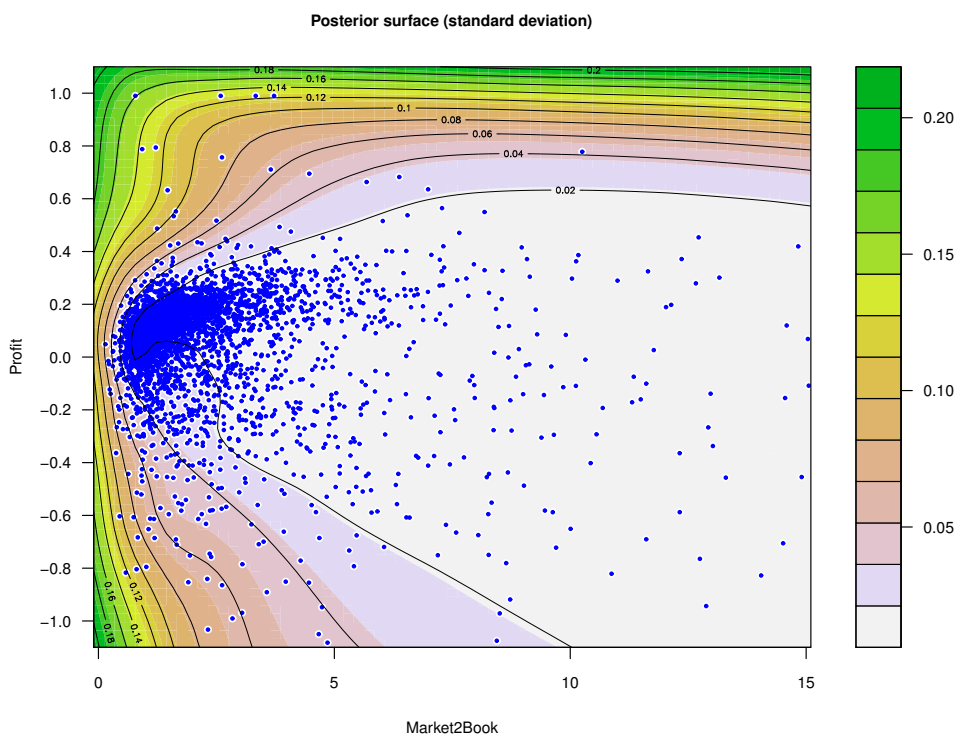
and $D = 5$ in the cross-validation.


Posterior locations of knots
Profit
Market2Book
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!