# Efficient Bayesian Multivariate Surface Regression

**Fan Cheng**

**Central University of Finance and Economics**

**chengfan9477@163.com**

**April 20, 2018**

# Outline

# 1 Flexible regression models

## 1.1 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, e.g. radial basis function,
  - $\xi_i$ are called knots (the location of the basis function).

# 1 Flexible regression models

## 1.1 Introduction

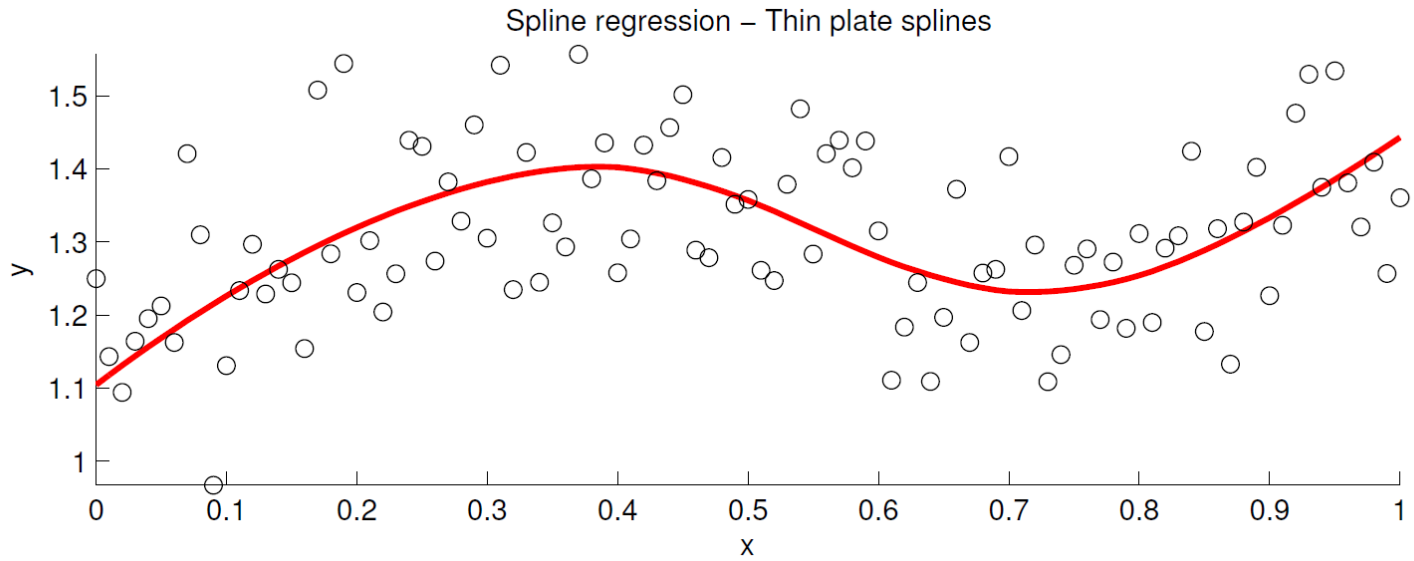
- Example: a simple spline regression with only one explanatory variable with truncated linear basis function can be like this

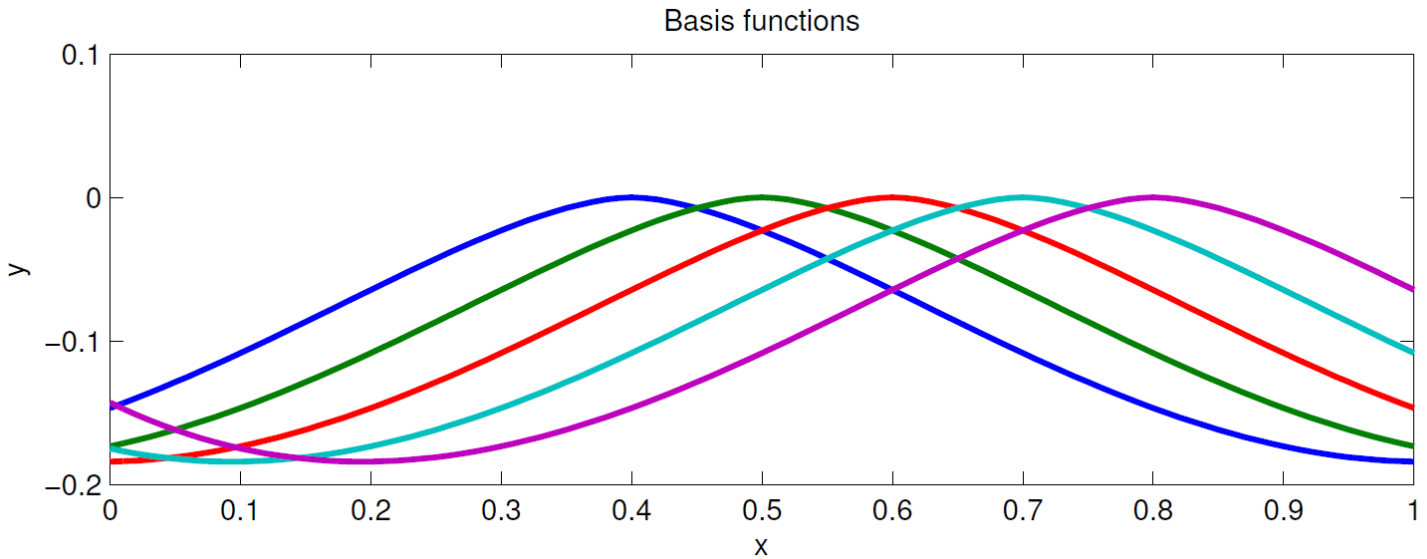
$$y = \alpha_0 + \alpha_1 x + \beta_1 (x - \xi_1)_+ + ... + \beta_q (x - \xi_q)_+ + \varepsilon$$

  where
  - $(x - \xi_i)_+$ are called the basis functions, e.g. radial basis function,
  - $\xi_i$ are called knots (the location of the basis function).
- A spline is a linear regression on a set of nonlinear basis functions of the original regressors. Each basis function is defined from a knot in regressor space and the knots determine the points of flexibility of the fitted regression function.
- Benefit of splines
  - determine the points of flexibility of the fitted regression function
  - locally polynomial model with continuity at the knots.

# 1 Flexible regression models

## ↛ 1.2 Spline example (single covariate with thinplate bases)



# 1 Flexible regression models

## ↛ 1.3 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.
- Surface spline model
  - Each knot $\xi_j$ (vector) is connected with more than one covariate

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

  - A popular choice of $g(x_1, ...x_q, \xi_j)$ can be e.g. the multi-dimensional thinplate spline

$$g(x_1, ...x_q, \xi_j) = \|x - \xi_j\|^2 \ln \|x - \xi_j\|$$

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

# 2 The challenges in spline regression

- How many knots are needed?
  - Too few knots lead to a bad approximation.
    - ⋆ Q-dimensional knots are necessarily sparse in $\mathbb{R}^q$, which is the curse of dimensionality.
  - Too many knots yield overfitting.
    - ⋆ Bayesian variable selection methods are used to prevent overfitting using Markov chain Monte Carlo (MCMC) techniques (Smith & Kohn, 1996).
- Where to place those knots?
  - Equal spacing for the additive model,
  - which is obviously not efficient with the surface model.

# 2 The challenges in spline regression

- 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 (RJMCMC) 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

# 3 The multivariate surface model
## ↛ The model

- The multivariate surface model consists of three different components, *linear*, *surface* and *additive* as

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

where

  - $\mathbf{Y}(n \times p)$ contains $n$ observations on $p$ response variables,
  - $\mathbf{X}_o(n \times q_o)$ contains the original regressors (first column is a vector of ones for the intercept). $\mathbf{B}_o$ holds the corresponding regression coefficients,
  - The $q_a$ columns of the matrix $\mathbf{X}_a(\xi_a)$ are additive splines functions of the covariates in $\mathbf{X}_o$. $\mathbf{X}_a$ depends on the knots $\xi_a$. The knots in $\xi_a$ are scalars,
  - $\mathbf{X}_s(\xi_s)$ contains the surface, or interaction, part of the model. The knots in $\xi_s$ are $q_o$-dimensional vectors,
  - the rows of $\mathbf{E}$ are error vectors assumed to be independent and identically distributed (*iid*) as $N_p(\mathbf{0}, \mathbf{\Sigma})$.

# 3 The multivariate surface model

## ↛ The model

- The multivariate surface model consists of three different components, *linear*, *surface* and *additive* as

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

- The additive part of the model captures the main part of the nonlinearities so that the number of knots in $\mathbf{X}_s$ is kept to a minimum.
- We treat the knots $\xi_i$ as unknown parameters and let them move freely.
  - A model with a minimal number of free knots outperforms model with lots of fixed knots.
- We use thin-plate splines as the basis function.

$$x_{sj}(\xi_{sj}) = \|x_o - \xi_{sj}\|^2 \log \|x_o - \xi_{sj}\|, \ j = 1, ..., q_s, \tag{1}$$

$$x_{aj}(\xi_{aj}) = \sum_{j=1}^{q_a} |x_o - \xi_{aj}|^2 \log |x_o - \xi_{aj}|, \ j = 1, ..., q_a. \tag{2}$$

- For notational convenience, we sometimes write model in compact form

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},$$

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

# 4 The MCMC algorithm[1]

- Markov Chains
- Metropolis Algorithm
- Metropolis-Hastings
- Metropolis-Hastings within Gibbs

[1] Please refer to `https://feng.li/files/bda/L5.3-Monte-Carlo-Methods-with-Details` for more details.

# Markov Chains

- The goal of today's lecture is to learn about the **Metropolis Hastings algorithm**
- The Metropolis Hastings algorithm allows us to simulate from any distribution as long as we have the kernel of the density of the distribution.
- To understand the Metropolis Hastings algorithm, we must learn a little bit about **Markov chains**

# Basic Probability Rules

- **Law of conditional probability**

$$\Pr(A = a, B = b) = \Pr(A = a|B = b)\Pr(B = b) \tag{1}$$

- More general conditional probability

$$\Pr(A = a, B = b|C = c) = \Pr(A = a|B = b, C = c) \times \Pr(B = b|C = c) \tag{2}$$

# Basic Probability Rules

- **Marginalizing** (for a discrete variable)

$$\Pr(A = a) = \sum_b \Pr(A = a, B = b) \tag{3}$$

- More general

$$\Pr(A = a|C = c) = \sum_b \Pr(A = a, B = b|C = c) \tag{4}$$

# Independence

- Two variables are **independent** if

$$\Pr(A = a, B = b) = \Pr(A = a)\Pr(B = b) \quad \forall a, b \tag{5}$$

- Dividing both sides by Pr(B=b) gives

$$\Pr(A = a|B = b) = \Pr(A = a) \quad \forall a, b \tag{6}$$

# Conditional Independence

- Two variables A and B are **Conditionally Independent** if

$$\Pr(A = a, B = b|C = c) = \Pr(A = a|C = c) \times \Pr(B = b|C = c) \quad \forall a, b, c \tag{7}$$

- Dividing both sides by $\Pr(B = b|C = c)$ gives

$$\Pr(A = a|B = b, C = c) = \Pr(A = a|C = c) \quad \forall a, b, c \tag{8}$$

# A simple game

- Player A and Player B play a game. The probability that Player A wins each game is 0.6 and the probability that Player B wins each game is 0.4.
- They play the game N times.
- Each game is **independent**.
- Let
  - $X_i = 0$ if Player A wins game $i$
  - $X_i = 1$ if Player B wins game $i$
- Also assume there is an initial Game called Game 0 ($X_0$)

# Some simple questions

- What is the probability that Player A wins Game 1 ($(X_1 = 0)$) if
  - If $X_0 = 0$ (Player A wins Game 0)
  - If $X_0 = 1$ (Player B wins Game 0)
- What is the probability that Player A wins Game 2 ($(X_2 = 0)$) if
  - If $X_0 = 0$ (Player A wins Game 0)
  - If $X_0 = 1$ (Player B wins Game 0)
- Since each game is independent all answers are 0.6.

# A different game: A Markov chain

- Now assume that both players have a better chance of winning Game $i+1$ if they already won Game $i$.

$$\Pr(X_{i+1} = 0 | X_i = 0) = 0.8 \tag{9}$$
$$\Pr(X_{i+1} = 1 | X_i = 1) = 0.7 \tag{10}$$

- Assume nothing other than game $i$ has a direct effect on Game $i+1$.
- This is called the **Markov Property**. Mathematically

$$\Pr(X_{i+1} | X_i, X_{i-1}, \ldots, X_1, X_0) = \Pr(X_{i+1} | X_i) \tag{11}$$

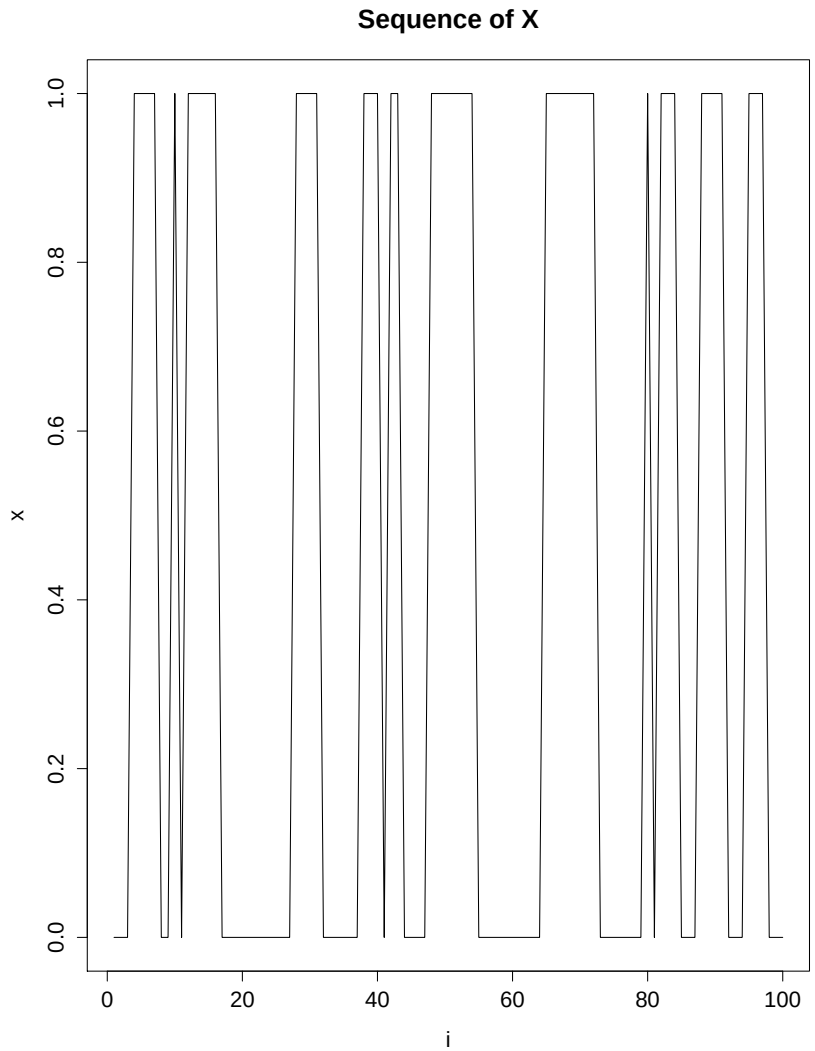
# Markov Property

- Another way to define the Markov property is to notice that $X_{i+1}$ and $X_{i-1}, \ldots, X_0$ are **independent** conditional on $X_i$
- This may be a model for the stock market, all the valuable information about tomorrow's stock price is contained in today's price.
- This is related to the **Efficient Market Hypothesis**, a popular theory in finance.
- Now back to the simple game.

# Simulating from a Markov chain

- Now let's simulate a sequence $X_1, X_2, \ldots, X_{100}$ from the Markov chain.
- Initialize at $x_0 = 0$. Then inside a loop
- Code the following using *if*.
  - if $X_i = 0$ then $X_{i+1} = \begin{cases} 0 & \text{with probability } 0.8 \\ 1 & \text{with probability } 0.2 \end{cases}$
  - if $X_i = 1$ then $X_{i+1} = \begin{cases} 0 & \text{with probability } 0.3 \\ 1 & \text{with probability } 0.7 \end{cases}$
- Try it

# Markov chain



# Simple questions again

- What is the probability that Player A wins the first game (i.e $(X_1 = 0)$) if
  - If $X_0 = 0$ (Player A wins initial game)
  - If $X_0 = 1$ (Player B wins initial game)
- The answers are 0.8 and 0.3.
- What is the probability that Player A wins the second game $(X_2 = 0)$ if
  - If $X_0 = 0$ (Player A wins initial game)
  - If $X_0 = 1$ (Player B wins initial game)

# Solution

- Let $X_0 = 0$. Then $\Pr(X_2 = 0|X_0 = 0)$

$$
\begin{aligned}
&= \sum_{x_1=0,1} \Pr(X_2 = 0, X_1 = x_1|X_0 = 0) \\
&= \sum_{x_1=0,1} \Pr(X_2 = 0|X_1 = x_1, X_0 = 0)\Pr(X_1 = x_1|X_0 = 0) \\
&= \sum_{x_1=0,1} \Pr(X_2 = 0|X_1 = x_1)\Pr(X_1 = x_1|X_0 = 0) \\
&= 0.8 \times 0.8 + 0.3 \times 0.2 \\
&= 0.7
\end{aligned}
$$

- What if $X_0 = 1$?

# Recursion

- Notice that the distribution of $X_i$ depends on $X_0$
- The sequence is no longer independent.
- How could you compute $\Pr(X_n = 0|X_0 = 0)$ when $n = 3$, when $n = 5$, when $n = 100$?
- This is hard, but the Markov Property does make things simpler
- We can use a recursion to compute the probability that Player A wins any game.

# Recursion

Note that $\Pr(X_i = 0 | X_0 = 0)$

$$
\begin{aligned}
&= \sum_{x_{i-1}} \Pr(X_i = 0, X_{i-1} = x_{i-1} | X_0 = 0) \\
&= \sum_{x_{i-1}} \Pr(X_i = 0 | X_{i-1} = x_{i-1}, X_0 = 0) \Pr(X_{i-1} = x_{i-1} | X_0 = 0) \\
&= \sum_{x_{i-1}} \Pr(X_i = 0 | X_{i-1} = x_{i-1}) \Pr(X_{i-1} = x_{i-1} | X_0 = 0)
\end{aligned}
$$

We already applied this formula when $i = 2$. We can continue for $i = 3, 4, 5, \ldots, n$

# Recursion

$$\Pr(X_i = 0|X_0 = 0) = \sum_{x_{i-1}} \Pr(X_i = 0|X_{i-1} = x_{i-1})\Pr(X_{i-1} = x_{i-1}|X_0 = 0)$$

- Start with $\Pr(X_1 = 0|X_0 = 0)$
- Get $\Pr(X_1 = 1|X_0 = 0)$
- Use these in formula with $i = 2$
- Get $\Pr(X_2 = 0|X_0 = 0)$
- Get $\Pr(X_2 = 1|X_0 = 0)$
- Use these in formula with $i = 3$
- Get $\Pr(X_3 = 0|X_0 = 0)$
- $\vdots \quad \vdots \quad \vdots \quad \vdots$

# Matrix Form

It is much easier to do this calculation in matrix form (especially when $X$ is not binary). Let $P$ be the transition matrix

| | $X_i = 0$ | $X_i = 1$ |
|---|---|---|
| $X_{i-1} = 0$ | $Pr(X_i = 0 \mid X_{i-1} = 0)$ | $Pr(X_i = 1 \mid X_{i-1} = 0)$ |
| $X_{i-1} = 1$ | $Pr(X_i = 0 \mid X_{i-1} = 1)$ | $Pr(X_i = 1 \mid X_{i-1} = 1)$ |

# Matrix Form

In our example:

| | $X_i = 0$ | $X_i = 1$ |
|---|---|---|
| $X_{i-1} = 0$ | 0.8 | 0.2 |
| $X_{i-1} = 1$ | 0.3 | 0.7 |

$$P = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix} \tag{12}$$

# Matrix Form

Let $\pi_i$ be a $1 \times 2$ row vector which denotes the probabilities of each player winning Game $i$ conditional on the initial Game

$$\pi_i = (\Pr(X_i = 0|X_0), \Pr(X_i = 1|X_0)) \tag{13}$$

In our example if $X_0 = 0$

$$\pi_1 = (0.8, 0.2) \tag{14}$$

In our example if $X_0 = 1$

$$\pi_1 = (0.3, 0.7) \tag{15}$$

# Recursion in Matrix form

- The recursion formula is

$$\pi_i = \pi_{i-1} P \tag{16}$$

  Therefore

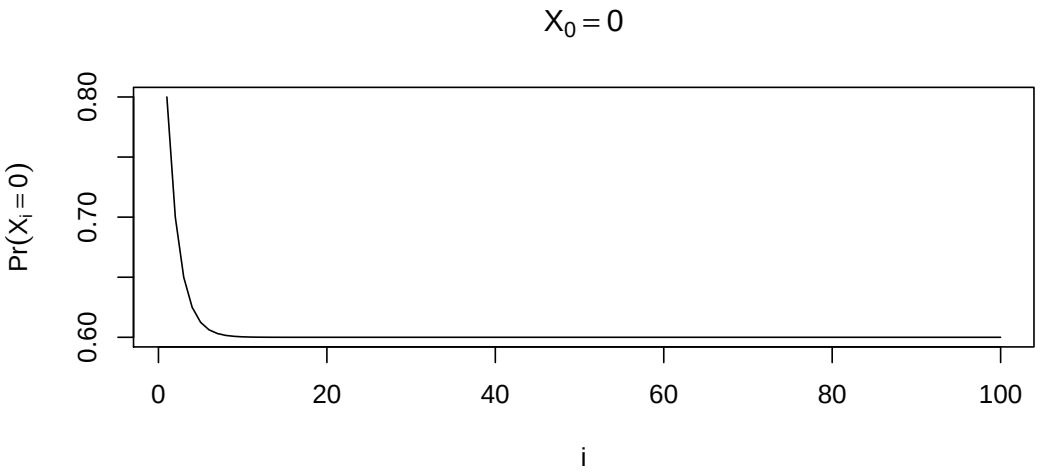
$$\pi_n = \pi_1 P \times P \times \ldots \times P \tag{17}$$

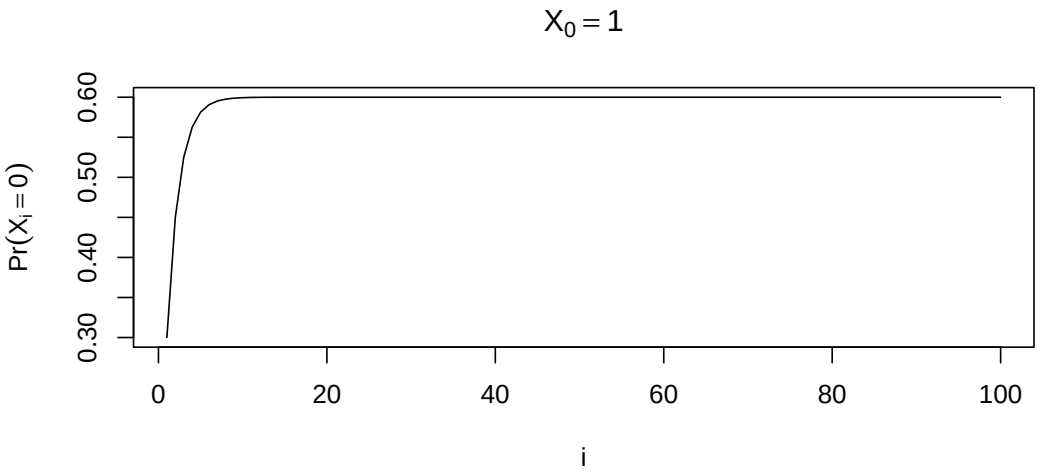
- Now code this up in R.
- What is $\Pr(X_n = 0 | X_0 = 0)$ when
  - $n = 3$
  - $n = 5$
  - $n = 100$?
- Do the same when $X_0 = 1$

# Convergence?

- For $n = 3$ and $n = 5$, the starting point made a big difference.
- For $n = 100$ it did not make a big difference.
- Could this Markov chain be converging to something?
- Now write code to keep the values of $\pi_i$ for $i = 1, 2, \ldots, 100$.
- Then plot the values of $\pi_{i1}$ against $i$

# Convergence



# More Questions

- What is $\Pr(X_{100} = 0|X_0 = 0)$?
- What is $\Pr(X_{100} = 0|X_0 = 1)$?
- What is $\Pr(X_{1000} = 0|X_0 = 0)$?
- What is $\Pr(X_{1000} = 0|X_0 = 1)$?
- The answer to all of these is 0.6.
- The $X$ do not converge. They keep changing from 0 to 1. The Markov chain however converges to a **stationary distribution**.

# Simulation with a Markov chain

- Go back to your code for generating a Markov chain and generate a chain with $n = 110000$
- Exclude the first 10000 values of $X_i$ and keep the remaining 100000 values.
- How many $X_i = 0$? How many $X_i = 1$
- We have discovered a new way to simulate from a distribution with $Pr(X_i = 0) = 0.6$ and $Pr(X_i = 1) = 0.4$

# Markov Chains

- Sometimes two different Markov chains converge to the same stationary distribution. See what happens when

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.15 & 0.85 \end{pmatrix} \tag{18}$$

- Sometimes Markov chains do not converge to a stationary distribution at all.
- Some Markov chains can get stuck in an **absorbing state**. For example what would the simple example look like if $Pr(X_{i+1} = 0|X_i = 0) = 1$?
- Markov chains can be defined on continuous support as well, $X_i$ can be continuous.

# Some important points

- This is a very complicated way to generate from a simple distribution.
- For the binary example the direct method would be better.
- However for other examples, either the direct method or accept/reject algorithm do not work.
- In these cases we can construct a Markov chain that has a stationary distribution that is our **target distribution**.
- All we need is the kernel of the density function, and an algorithm called the **Metropolis Algorithm**

# The Metropolis algorithm

- The Metropolis algorithm was developed in a 1953 paper by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller.
- The aim is to simulate $x \sim p(x)$ where $p(x)$ is called the **target density**.
- We will need a **proposal density** $q(x^{[old]} \rightarrow x^{[new]})$
- For example one choice of $q$ is

$$x^{[new]} \sim N(x^{[old]}, 1) \tag{20}$$

- This is called a **Random Walk proposal**

# Symmetric proposal

- An important property of $q$ in the Metropolis algorithm is symmetry of the proposal

$$q(x^{[old]} \to x^{[new]}) = q(x^{[new]} \to x^{[old]}) \tag{21}$$

- Later we will not need this assumption
- Can you confirm this is true for $x^{[new]} \sim N(x^{[old]}, 1)$?
- Can you simulate from this random walk (use $x_0 = 0$ as a starting value)?

# Proof of symmetry of random walk

The proposal

$$q(x^{[old]} \to x^{[new]}) = (2\pi)^{-1/2} exp\left\{-\frac{1}{2}\left(x^{[new]} - x^{[old]}\right)^2\right\} \tag{22}$$

$$= (2\pi)^{-1/2} exp\left\{-\frac{1}{2}\left[-1\left(x^{[new]} - x^{[old]}\right)\right]^2\right\} \tag{23}$$

$$= (2\pi)^{-1/2} exp\left\{-\frac{1}{2}\left(x^{[old]} - x^{[new]}\right)^2\right\} \tag{24}$$

$$= q(x^{[new]} \to x^{[old]}) \tag{25}$$

# Accept and reject

- By itself the random walk will not converge to anything.
- To make sure this Markov chain converges to our target, we need to include the following.
- At step $i+1$ set $x^{[old]} = x^{[i]}$.
- Generate $x^{[new]} \sim N(x^{[old]}, 1)$ and compute

$$\alpha = \min\left(1, \frac{p(x^{[new]})}{p(x^{[old]})}\right) \tag{26}$$

- Then
  - Set $x^{[i+1]}$ to $x^{[new]}$ with probability $\alpha$ (accept)
  - Set $x^{[i+1]}$ to $x^{[old]}$ with probability $1 - \alpha$ (reject)

# Non-Symmetric proposal

- In 1970, Hastings proposed an extension to the Metropolis Hastings algorithm.
- This allows for the case when

$$q(x^{[old]} \rightarrow x^{[new]}) \neq q(x^{[new]} \rightarrow x^{[old]}) \tag{28}$$

- The only thing that changes is the acceptance probability

$$\alpha = \min\left(1, \frac{p(x^{[new]})q(x^{[new]} \rightarrow x^{[old]})}{p(x^{[old]})q(x^{[old]} \rightarrow x^{[new]})}\right) \tag{29}$$

- This is called the **Metropolis-Hastings algorithm**

# An interesting proposal

- Suppose we use the proposal:

$$x^{new} \sim N(0, \sqrt{(5/3)}) \tag{30}$$

- What is $q(x^{[old]} \rightarrow x^{[new]})$?
- It is $q(x^{[new]})$ where $q(.)$ is the density of a $N(0, \sqrt{(5/3)})$.
- Is this symmetric?
- No, since generally $q(x^{[new]}) \neq q(x^{[old]})$

# Metropolis Hastings

- Code this where $p(.)$ is the standard t density with 5 d.f, and $q(.)$ is normal with mean 0 and standard deviation $5/3$.
- Inside a loop
  - Generate $x^{[new]} \sim N(0, \sqrt{(5/3)})$
  - Set $x^{old} = x^{[i]}$ and compute

$$\alpha = \min\left(1, \frac{p(x^{[new]})q(x^{[old]})}{p(x^{[old]})q(x^{[new]})}\right) \tag{31}$$

  - Set $x^{[i+1]}$ to $x^{[new]}$ with probability $\alpha$ (accept)
  - Set $x^{[i+1]}$ to $x^{[old]}$ with probability $1 - \alpha$ (reject)
- Try it

# Comparison

- The Effective Sample Size of this proposal is about 43000 much higher than the best random walk proposal.
- Why does it work so well?
- The standard t distribution with 5 df has a mean of 0 and a standard deviation of $\sqrt{(5/3)}$
- So the $\mathrm{N}(0, \sqrt{(5/3)})$ is a good approximation to the standard student t with 5 df.

# Laplace Approximation

Using a Taylor expansion of $\ln p(x)$ around the point $a$

$$\ln p(x) \approx \ln p(a) + \left.\frac{\partial \ln p(x)}{\partial x}\right|_{x=a} (x-a) + \frac{1}{2} \left.\frac{\partial^2 \ln p(x)}{\partial x^2}\right|_{x=a} (x-a)^2$$

Let $a$ be the point that maximises $\ln p(x)$ and let

$$b = -\left( \left.\frac{\partial^2 \ln p(x)}{\partial x^2}\right|_{x=a} \right)^{-1} \tag{32}$$

The approximation is

$$\ln p(x) \approx \ln p(a) - \frac{1}{2b}(x-a)^2$$

Taking exponential of both sides

$$p(x) \approx k \times exp\left[-\frac{(x-a)^2}{2b}\right]$$

Any distribution can be approximated by a normal distribution with mean $a$ and variance $b$ where $a$ and $b$ values can be found numerically if needed.

# Multiple variables

- Suppose we now want to sample from a bivariate distribution $\mathrm{p}(\mathrm{x}, z)$
- The ideas involved in this section work for more than two variables.
- It is possible to do a 2-dimensional random walk proposal. However as the number of variables goes up the acceptance rate becomes lower.
- Also the Laplace approximation does not work as well in high dimensions.
- Indirect methods of simulation suffer from the same problem.
- We need a way to break the problem down.

# Method of composition

- Markov chain methods, allow us to break multivariate distributions down.
- If it is easy to generate from $\mathrm{p}(\mathrm{x})$ then the best way is **Method of composition**. Generate
  - $\mathrm{x}^{[\mathrm{i}]} \sim \mathrm{p}(\mathrm{x})$
  - $z^{[\mathrm{i}]} \sim \mathrm{p}(z|\mathrm{x} = \mathrm{x}^{[\mathrm{i}]})$
- Sometimes $\mathrm{p}(\mathrm{x})$ is difficult to get

# Gibbs Sampler

- If it is easy to simulate from the conditional distribution $f(x|z)$ then that can be used as a proposal
- What is the acceptance ratio?

$$\begin{aligned}\alpha &= \left(1, \frac{p(x^{new}, z)p(x^{old}|z)}{p(x^{old}, z)p(x^{new}|z)}\right) \\ &= \left(1, \frac{p(x^{new}|z)p(z)p(x^{old}|z)}{p(x^{old}|z)p(z)p(x^{new}|z)}\right) \\ &= 1\end{aligned}$$

# Gibbs Sampler

- This gives the Gibbs Sampler
  - Generate $x^{[i+1]} \sim p(x^{[i+1]}|z^{[i]})$
  - Generate $z^{[i+1]} \sim p(z^{[i+1]}|x^{[i+1]})$
  - Repeat
- $x$ and $z$ can be swapped around.
- It works for more than two variables.
- Always make sure the conditioning variables are at the current state.

# Metropolis within Gibbs

- Even if the individual conditional distributions are not easy to simulate from, Metropolis Hastings can be used *within* each Gibbs step.
- This works very well because it breaks down a multivariate problem into smaller univariate problems.
- We will practice some of these algorithms in the context of *Bayesian Inference*

# Summary

- You should be familiar with a *Markov chain*
- You should understand this can have a *stationary distribution*
- You should have a basic understanding of the Metropolis Hastings and the special cases
    - Random Walk Metropolis
    - Laplace approximation
    - Gibbs Sampler

# 5 The posterior inference

## 5.1 The prior

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

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

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

# 5 The posterior inference
## ↛ 5.2 The Bayesian posterior

- The posterior distribution is conveniently decomposed as

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

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

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

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

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

# 5 The posterior inference
## ↛ 5.3 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 $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$ is a multivariate $t$-density with $\nu > 2$ df,

$$\theta_p|\theta_c \sim \mathbf{t}\left[\hat{\theta},\ -\left(\frac{\partial^2 \ln p(\theta|\mathbf{Y})}{\partial\theta\partial\theta'}\right)^{-1}\Bigg|_{\theta=\hat{\theta}},\ \nu\right],$$

    - ⋆ where $\hat{\theta}$ is obtained by $R$ steps $(R \leqslant 3)$ Newton's iterations during the proposal with analytical gradients for matrices.
  - The proposal density for $\boldsymbol{\Sigma}$ is the inverse Wishart density on previous slide.

# 5 The posterior inference
## ↛ 5.4 Computational remarks

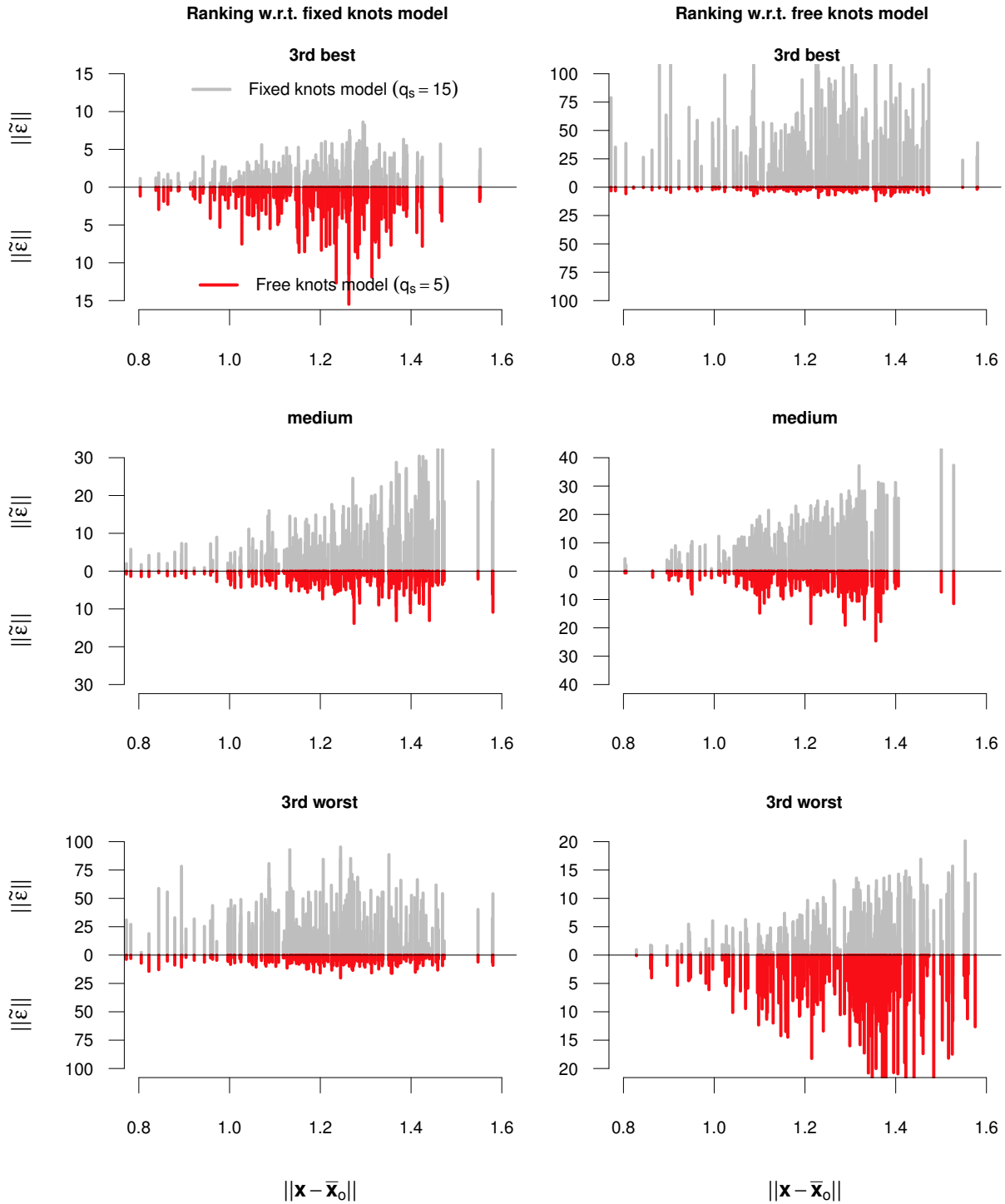
- The MCMC implementations are straightforward.
- But there are lots of parameters to estimate:
  - e.g., a 10 covariates, univariate response with 50 knots in both additive and surface components need $10 \times 50 + 1 \times 50 + 110 + 1 = 661$ parameters.
- We allow the parameters to be updated via:
  - parallel mode for small datasets,
  - batched mode for big datasets.
- We derive the analytical gradients during R-step Newton's iterations which are very complicated. We have implemented it in an efficient way.

# 6 Simulation study

- We randomly generate 100 datasets with various degrees of nonlinearity ($p = 1,\ 2,\ n = 200,\ 1000$). Covariates: Mixture of multivariate normal with five components; Bases: Five true bases from the covariates.
- For each dataset we fit with the fixed knots model for 5, 10, 15, 20, 25 and 50 surface knots, and also the free knots model for 5, 10, and 15 surface knots.
- The results show the free knots model outperforms the fixed knots model in the large majority of the datasets. This is particularly true when the data are strongly nonlinear.

---

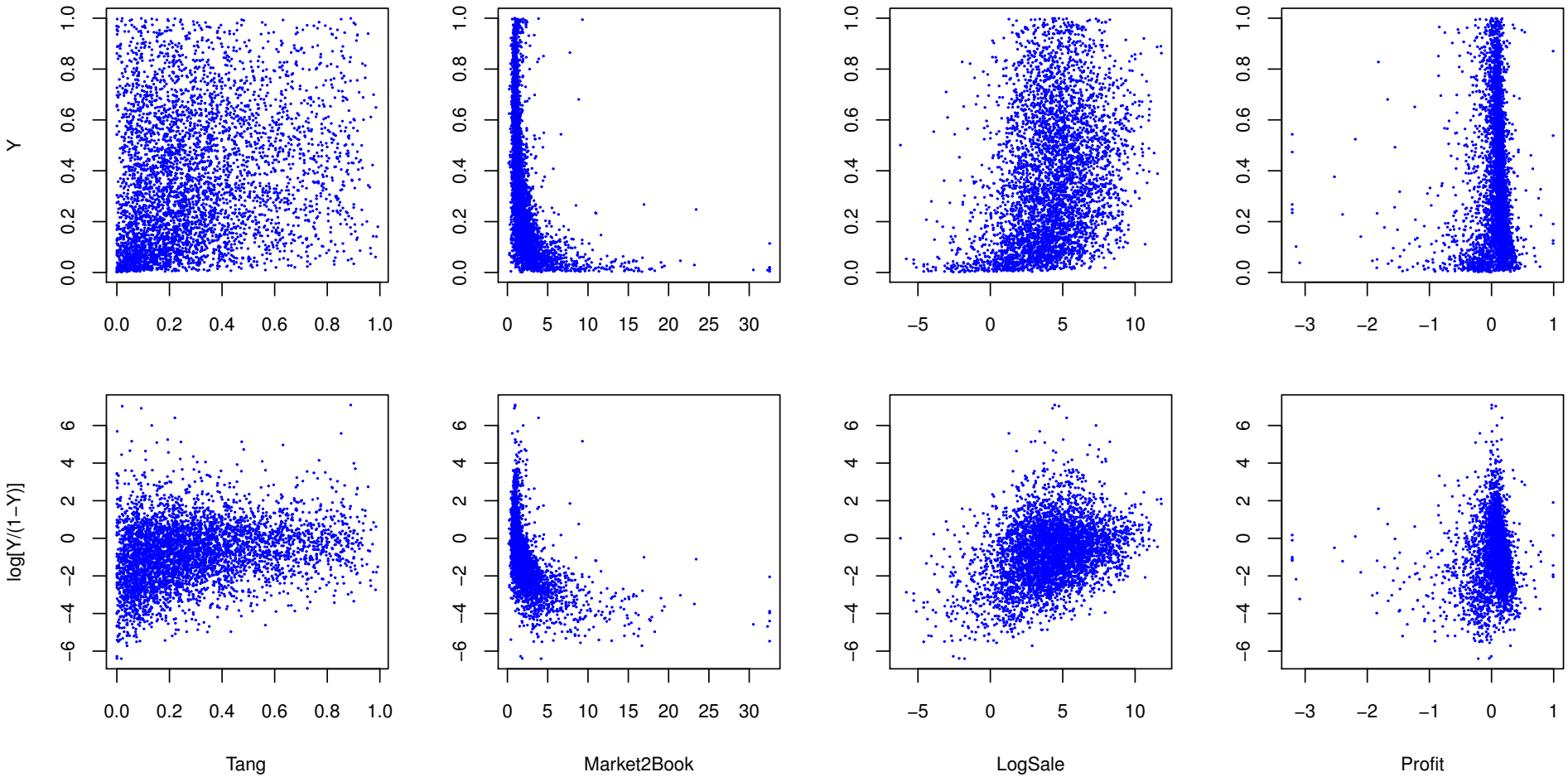
- Some results from the simulation →
- The norm of the predictive multivariate residuals ($p = 2$) against the distance between the testing covariates $x$ (randomly selected covariates space, therefore out-of-sample) and DGP sample mean.
- The residuals from vertical bars above the zero line is from the model with 15 fixed surface knots.
- The residuals from vertical bars below the zero line is from the model with 5 free surface knots.

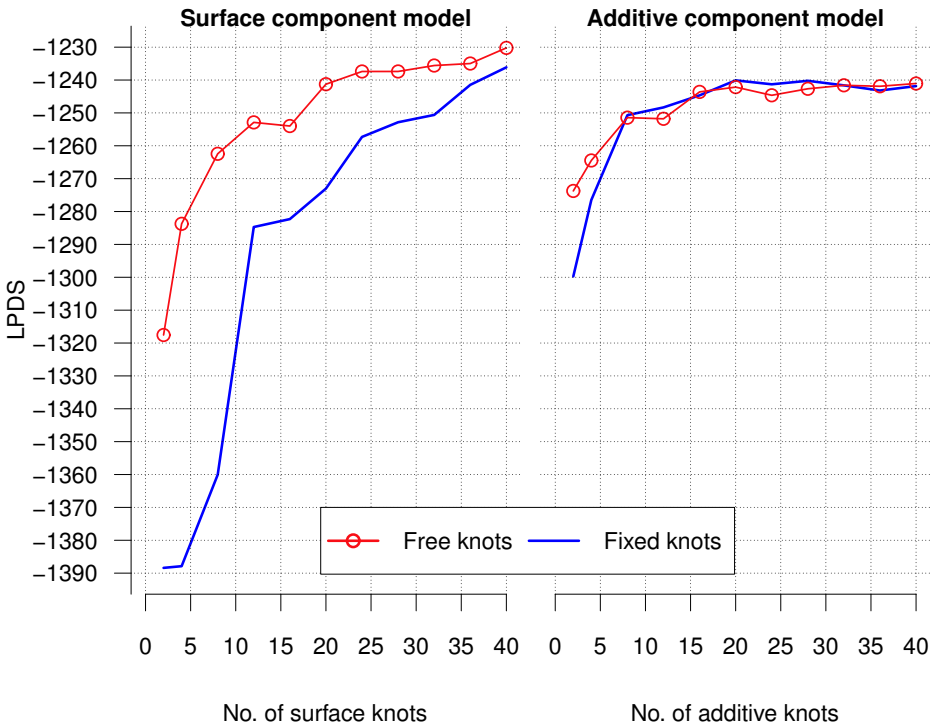


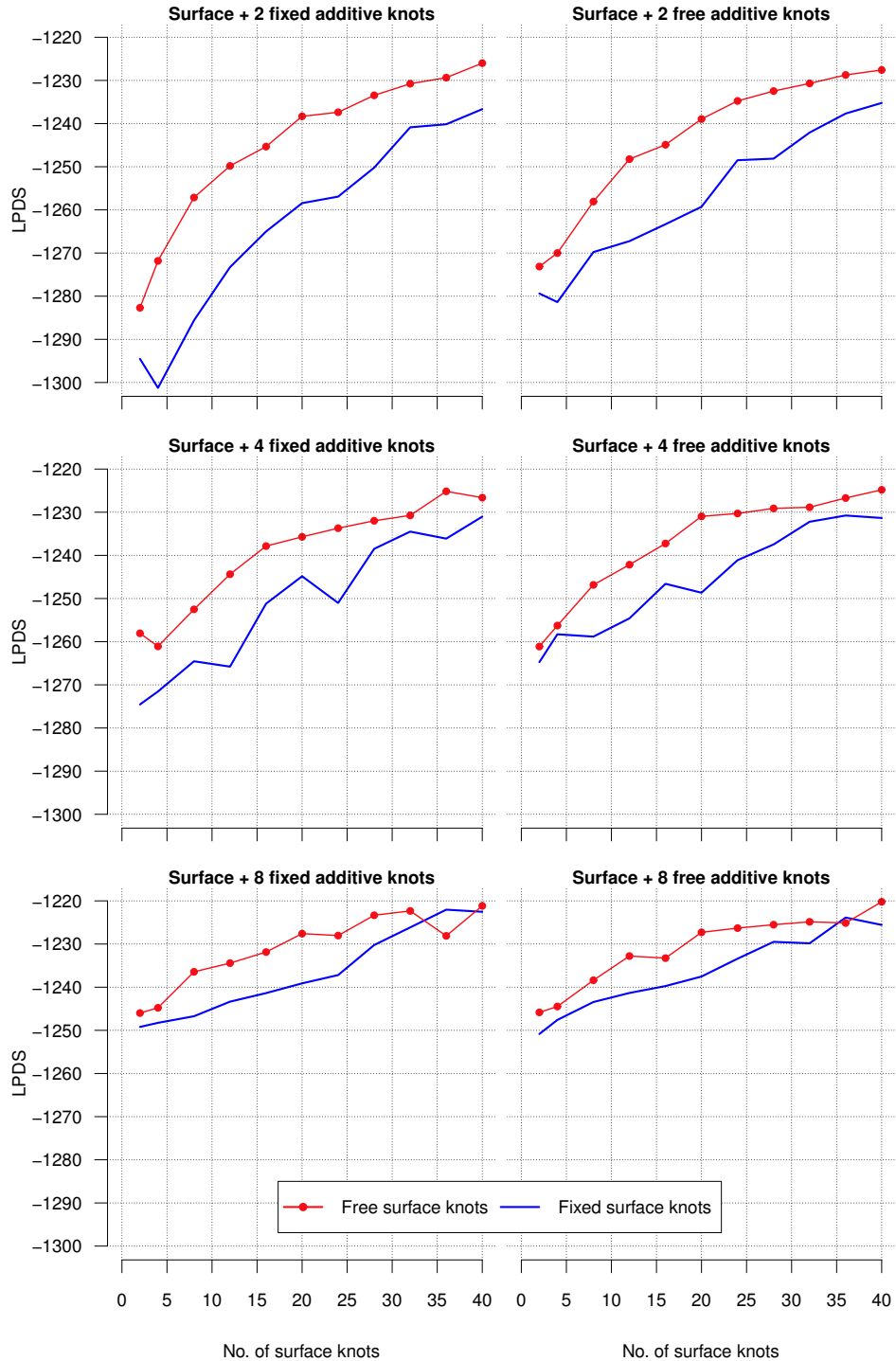
# 7 Application to firm leverage data

## ↛ 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.





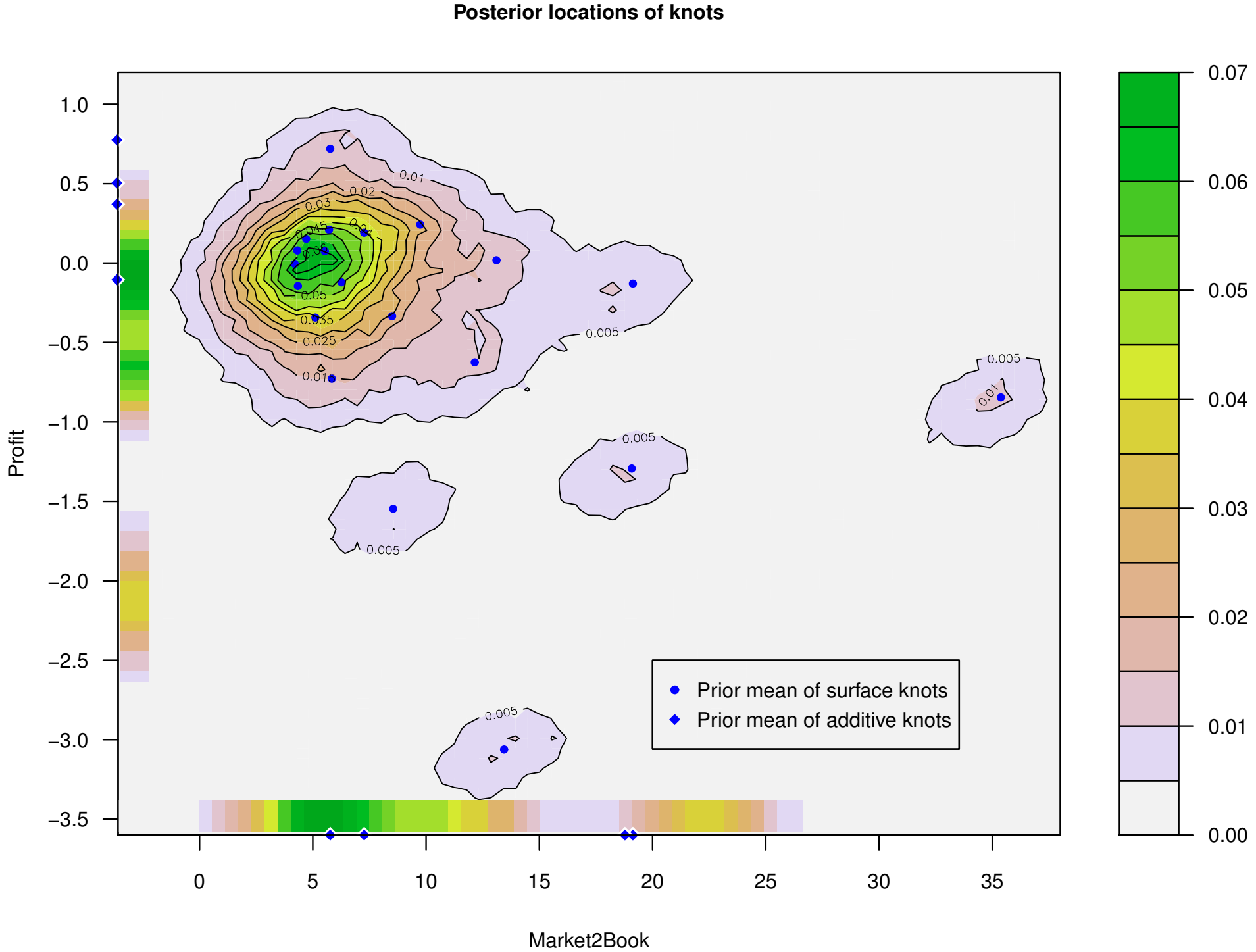

↑ Models with only surface or additive components

→ Model with both additive and surface components.

**LPDS** Log predictive density score which is defined as

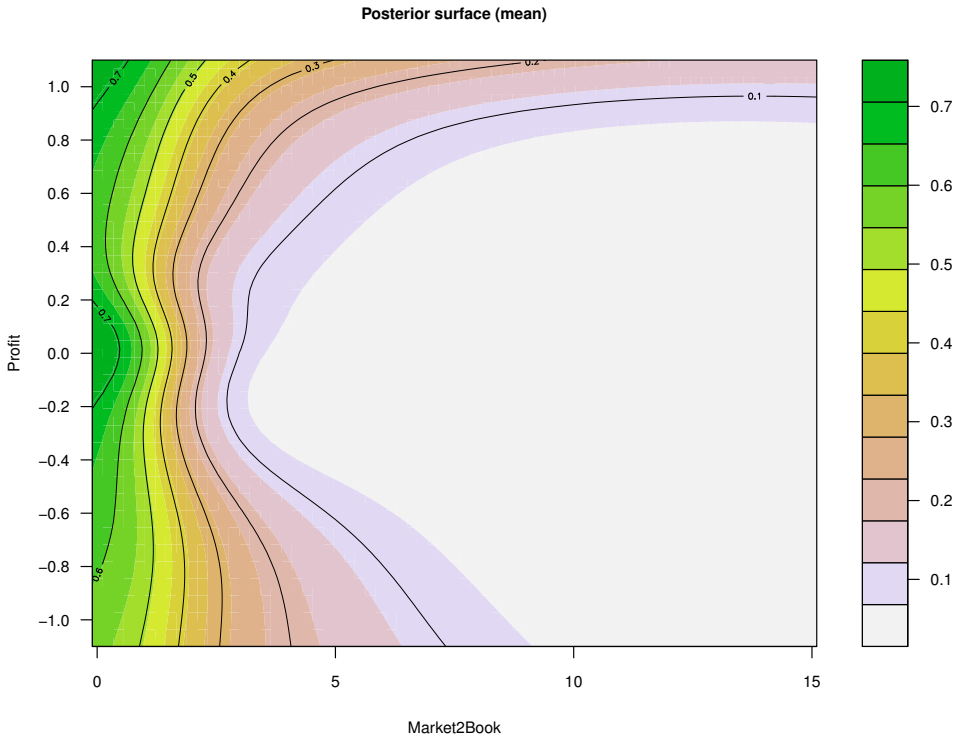
$$\frac{1}{D}\sum_{d=1}^{D} \ln p\left(\tilde{Y}_d \mid \tilde{Y}_{-d}, X\right),$$

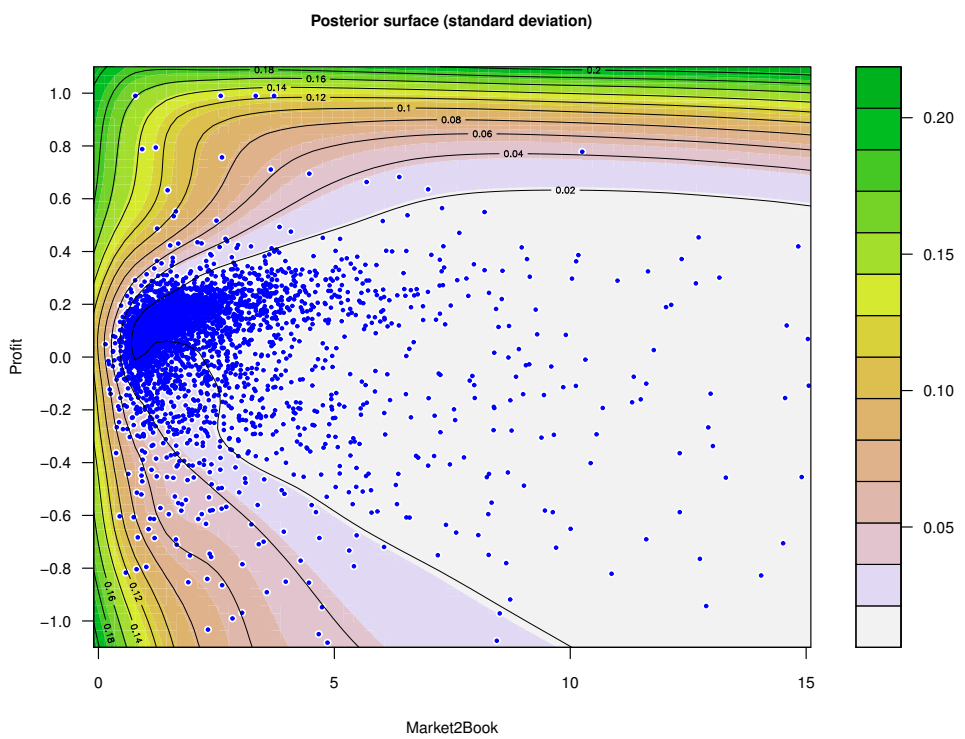
and $D = 5$ in the cross-validation.


Posterior locations of knots
Profit
Market2Book
Prior mean of surface knots
Prior mean of additive knots

# 7 Application to firm leverage data

## ↛ Posterior mean surface(left) and standard deviation(right)



# 8 Conclusions

- We have presented a general Bayesian approach for fitting a flexible surface model for a continuous multivariate response using a radial basis spline with freely estimated knot locations.
- Our approach uses shrinkage priors to avoid overfitting.
- All knot locations are sampled jointly using a Metropolis-Hastings proposal density tailored to the conditional posterior.
- Both a simulation study and a real application on firm leverage data show that models with free knots have a better out-of-sample predictive performance than models with fixed knots.
- models that mix surface and additive spline basis functions in the same model perform better than models with only one of the two basis types.

Thank you!