Bayesian Modeling of Conditional Densities

Feng Li
<feng.li@cufe.edu.cn>

School of Statistics and Mathematics
Central University of Finance and Economics

Revision: December 11, 2013

Outline

The trend of statistical modeling

- In the 1950s, linear regression model was considered as very advanced which is now the standard course content for university students.
- The data are much more complicated nowadays we meet.
  - Numerical, categorical, brain image...
  - A few observations to millions by millions.
  - Very high-dimensional data are not rare anymore.

Density estimation

- **Density estimation** is the procedure of estimating an unknown density $p(y)$ from observed data
- Histogram, kernel methods, splines, wavelets are all density estimation methods.
- **Mixture models** (Jiang & Tanner, 1999) have become a popular alternative approach,

$$p(y|\theta) = \sum_{k=1}^{K} \omega_k p_k(y|\theta_k),$$

  where $\sum_{k=1}^{K} \omega_k = 1$ for non-negative mixture **weights** ω_k and $p_k(x|\theta_k)$ are the **component densities**.
- If $K = \infty$, it is called an **infinite mixture** (Escobar, 1994), the **Dirichlet process mixture** being the most prominent example.
- Mixture densities can be used to capture data characteristics such as multi-modality, fat tails.

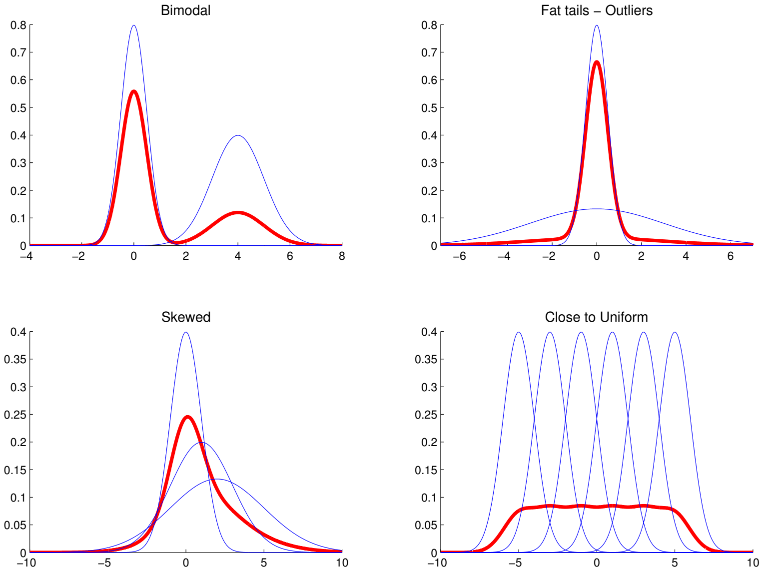


Figure: Using mixture of normal densities (thin lines) to mimic a flexible density (bold line).

Conditional density estimation

- The **conditional density estimation** concentrates on modeling the relationship between a response y and set of covariates x through a conditional density function $p(y|x)$
- Mixtures of conditional densities is the obvious extension of mixture models to the conditional density estimation problem:

$$p(y|x) = \sum_{k=1}^{K} \omega_k p_k(y|x)$$

  where $p_i(y|x)$ is the conditional density in i:th mixture component.
- A **smooth mixture** is a finite mixture density with weights that are smooth functions of the covariates

$$\omega_k(x) = \frac{\exp(x'\gamma_k)}{\sum_{i=1}^{K} \exp(x'\gamma_i)}.$$

Conditional density estimation

- In conditional density estimation, an important focus is modeling the regression mean $E(y|x)$.
- A **spline** is a popular approach for nonlinear regression that models the mean as a linear combination of a set of nonlinear basis functions of the original regressors (Holmes & Mallick, 2003),

$$y = f(x) + \epsilon = x'\beta + \sum_{i=1}^{k} x(\xi_i)'\beta_i + \epsilon$$

Multivariate density estimation with copulas

- The **multivariate density estimation** and conditional density estimation are analogues of their univariate cases except that the densities $p(\mathbf{Y})$ and $p(\mathbf{Y}|\mathbf{X})$ are multivariate.
- In addition to the methods mentioned above, a **copula function** separates the multivariate dependence from its marginal functions, and it is possible to use both continuous and discrete marginal models.
- Let $F(y_1, ..., y_M)$ be a multi-dimensional distribution function with marginal distribution functions $F_1(y_1), \cdots, F_M(y_M)$. Then there exists a copula function C (Sklar, 1959) such that

$$\begin{aligned} F(y_1, ..., y_M) =& C(F_1(y_1), ..., F_M(y_M)) \\ =& C\left(\int_{-\infty}^{y_1} f_1(z_1)dz_1, ..., \int_{-\infty}^{y_M} f_M(z_M)dz_M\right) = C(u_1, ..., u_M) \end{aligned}$$

Multivariate density estimation with copulas

- The **Kendall's τ correlation** between two marginal densities can be measured by Kendall's τ

$$\tau = 4\int\int F(y_1, y_2)dF(y_1, y_2) - 1 = 4\int\int C(u_1, u_2)dC(u_1, u_2) - 1.$$

- **Tail-dependence** measures the extent to which several variables simultaneously take on extreme values

$$\lambda_L = \lim_{u \to 0^+} \Pr(X_1 < F_1^{-1}(u) | X_2 < F_2^{-1}(u)) = \lim_{u \to 0^+} \frac{C(u, u)}{u},$$
$$\lambda_U = \lim_{u \to 1^-} \Pr(X_1 > F_1^{-1}(u) | X_2 > F_2^{-1}(u)) = \lim_{u \to 1^-} \frac{1 - C(u, u)}{1 - u}.$$

- Modeling tail-dependence is an very important topic in econometrics (Joe, 1997) (Patton, 2012).

The Bayesian approach for modeling density features

↪ A feature of a density

- We use the word **feature** to describe a characteristic of a density.
- In GLM or splines, $\mu = \eta(X\beta)$ is the feature that describes the **mean**.
- In mixtures contents, the **mean**, **variance**, **skewness** and **kurtosis** are features of each component density.
- In copula modeling, the **tail-dependence** and **correlation** are two features of interest.
- We allow each of the features are connected to covariates as

$$\begin{aligned}
\mu &= \beta_{\mu 0} + x_t'\beta_\mu \\
\ln \phi &= \beta_{\phi 0} + x_t'\beta_\phi \\
\ln \lambda &= \beta_{\lambda 0} + x_t'\beta_\lambda \\
\ln \nu &= \beta_{\nu 0} + x_t'\beta_\nu \\
\lambda_L &= \varphi_\lambda^{-1}(X\beta_\lambda) \\
\tau &= \varphi_\tau^{-1}(X\beta_\tau).
\end{aligned}$$

- This approach allows the feature to be dynamic and interpretable friendly.
- We only need to sample the posterior of $p(\beta|Data)$.

The Bayesian approach for modeling density features
$\looparrowright$ The general MCMC scheme

- The model settings are very complicated now.
- Sampling the posterior requires an efficient MCMC method.
- We update all the parameters jointly by using Metropolis-Hastings within Gibbs.
- The proposal density for each parameter vector β is a multivariate *t*-density with $\mathrm{df} > 2$,

$$\boldsymbol{\beta}_p|\boldsymbol{\beta}_c \sim \boldsymbol{MVT}\left[\hat{\boldsymbol{\beta}},\ -\left(\frac{\partial^2 \ln p(\boldsymbol{\beta}|\mathbf{Y})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right)^{-1}\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}},\ \mathrm{df}\right],$$

  where $\hat{\boldsymbol{\beta}}$ is obtained by R steps ($R \leqslant 3$) Newton's iterations during the proposal with analytical gradients.
- Variable selections are carried out simultaneously.
- **The key:** The analytical gradients require the derivative for the copula density and marginal densities.

Regularization via Bayesian variable selection

- **Variable selection** is commonly to select meaningful covariates that contributes to the model, inhibit ill-behaved design matrices, and to prevent model over-fitting.
- A standard Bayesian variable selection approach (Nott & Kohn, 2005) is to augment the regression model with a variable selection indicator $\mathcal{I}$ for each covariate

$$\mathcal{I}_j = \begin{cases} 1 & \text{if } \beta_j \neq 0 \\ 0 & \text{if } \beta_j = 0, \end{cases}$$

  where β_j is the jth covariate in the model.
- Variable selection is then obtained by sampling the posterior distribution of all regression coefficient jointly with the variable selection indicators, thereby yielding the marginal posterior probability of variable inclusion $p(\mathcal{I}|\text{Data})$.

Regularization via shrinkage estimator

- A **shrinkage estimator** shrinks the regression coefficients towards zero rather than eliminating the covariate completely.
- LASSO can be viewed as regression with a Laplace prior.
- One way to select a proper value of the shrinkage is by cross-validation, which is costly with big data and complicated models.
- In the Bayesian approach, the shrinkage parameter is usually automatically estimated together with other parameters in the posterior inference.
- Shrinkage and variable selection can be used **simultaneously**.

Bayesian predictive inference

- Assuming that the data observations are independent conditional on the model parameters θ, the **predictive density** can be written

$$p(Y_b|Y_{-b}) = \int \prod_{j=1}^{n} p(Y_{j,b}|\theta)p(\theta|Y_{-b})d\theta$$

- For a time series the forecast can instead be based on the decomposition

$$\begin{aligned} p(y_{T+1}, .., y_{T+T*}|y_1, .., y_T) =& p(y_{T+1}|y_1, .., y_T) \times \cdots \\ & \times p(y_{T+T*}|y_1, .., y_{T+T*-1}), \end{aligned}$$

with each term in the decomposition

$$p(y_t|y_1, .., y_{t-1}) = \int p(y_t|y_1, .., y_{t-1}, \theta)p(\theta|y_1, .., y_{t-1})d\theta,$$

- **The prediction error** at x_0 can be decomposed as three parts

$$\begin{aligned} EPE(x_0) &= E((Y - \hat{f}(x_0))^2|X = x_0) \\ &= \sigma^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0)) \end{aligned}$$

which is the so-called the **bias-variance trade-off**.

Bayesian model comparison

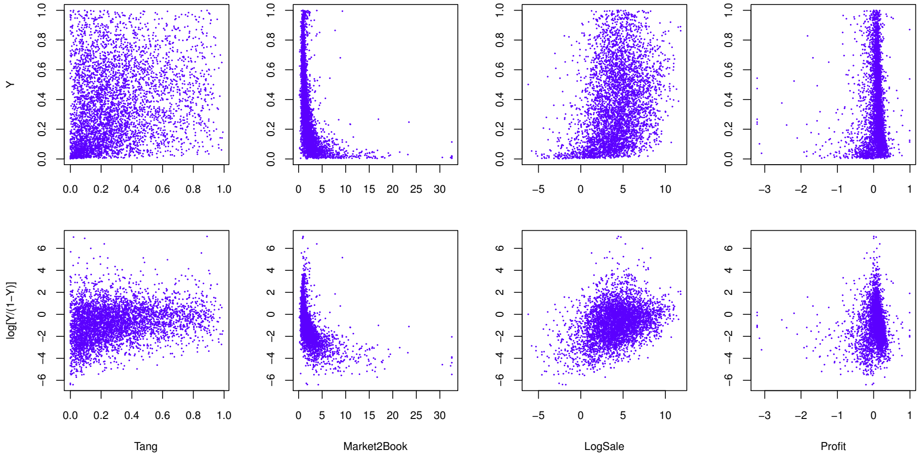
- Bayesian model comparison have historically been based on the marginal likelihood (Kass & Raftery, 1995).
- However,that the marginal likelihood is very sensitive to the specification of prior.
- A more prominent tool for model comparisons is based on the **log predictive density score** (LPDS)

$$\mathsf{LPDS} = \frac{1}{B}\sum_{i=1}^{B} \log p(Y_{b_i}|Y_{-b_i})$$

- In Bayesian framework, as the whole posterior of parameters can be obtained, model consistency evaluation does not rely one large sample properties.
- There are still consistency studies on issues like variable selections (Casella et al., 2009).

Modeling nonlinear mean with splines 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.



The multivariate surface model

$\looparrowright$ The model

- Splines are regression models with flexible **mean functions** by selecting and placing knots to covariates space.
- The multivariate surface spline model (Li & Villani, 2013) 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
$\hookrightarrow$ The prior

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

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

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

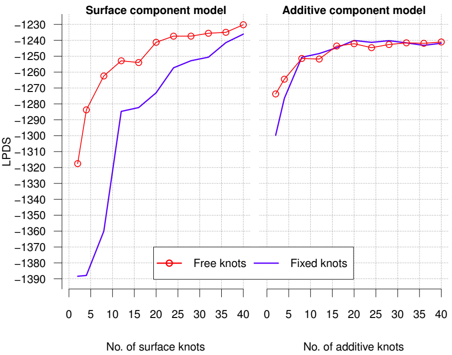
$$p(\mathbf{B}, \boldsymbol{\Sigma}, \xi, \lambda | \mathbf{Y}, \mathbf{X}) = p(\mathbf{B} | \boldsymbol{\Sigma}, \xi, \lambda, \mathbf{Y}, \mathbf{X}) p(\boldsymbol{\Sigma} | \xi, \lambda, \mathbf{Y}, \mathbf{X}) p(\xi, \lambda | \mathbf{Y}, \mathbf{X}).$$

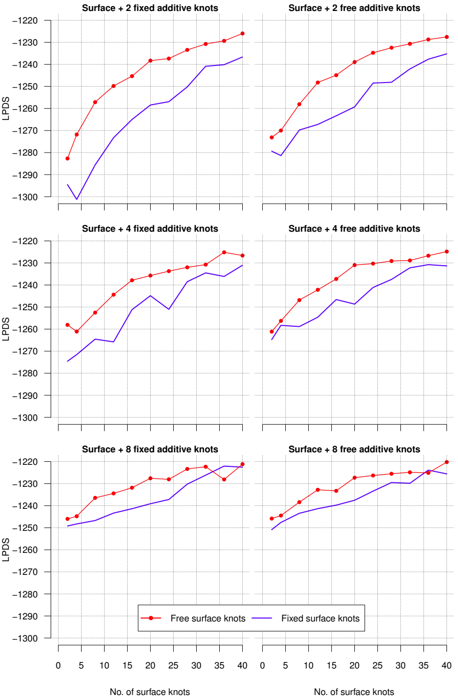
- Hence $p(\mathbf{B} | \boldsymbol{\Sigma}, \xi, \lambda, \mathbf{Y}, \mathbf{X})$ follows the multivariate normal distribution according to the conjugacy;
- When $p = 1$, $p(\boldsymbol{\Sigma} | \xi, \lambda, \mathbf{Y}, \mathbf{X})$ follows the inverse Wishart distribution

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

- 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

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





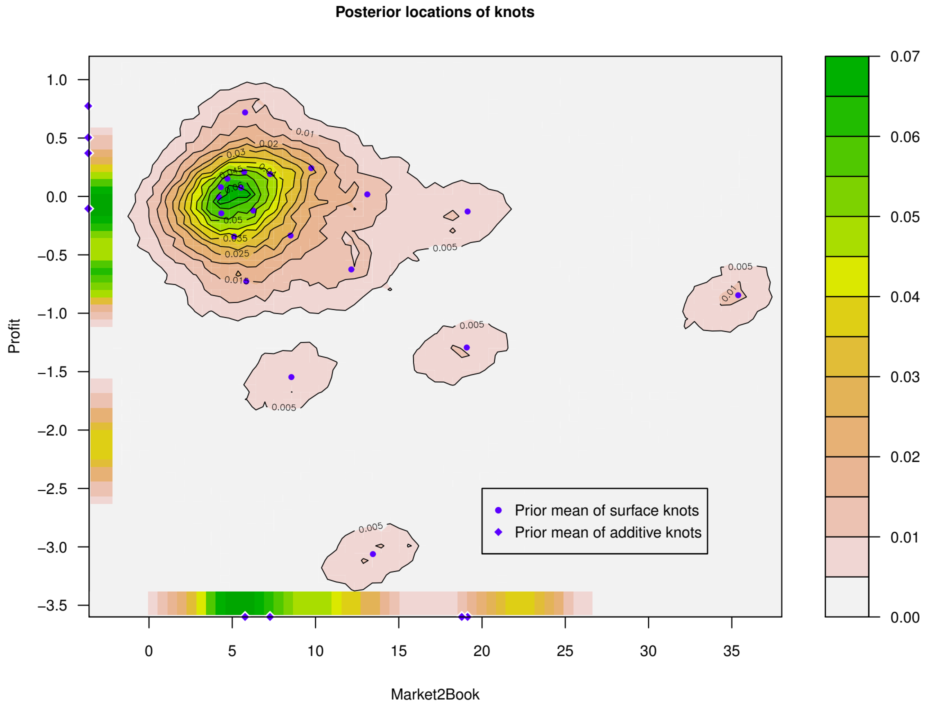
$\uparrow$ Models with only surface or additive components

$\rightarrow$ Model with both additive and surface components.

PDS Log predictive density score which is defined as

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

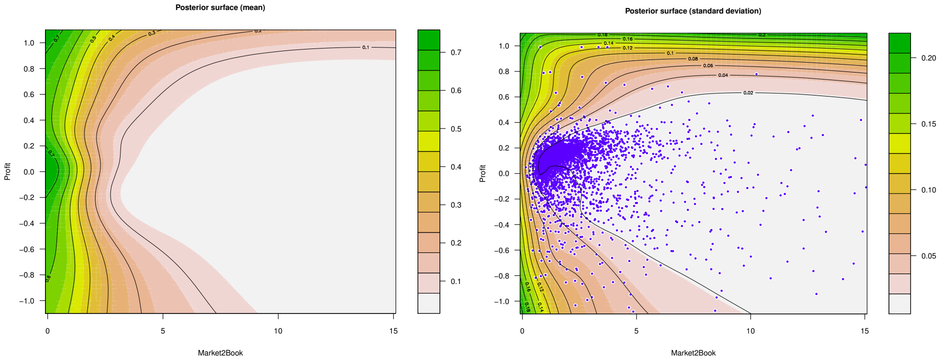
and $D = 5$ in the cross-validation.



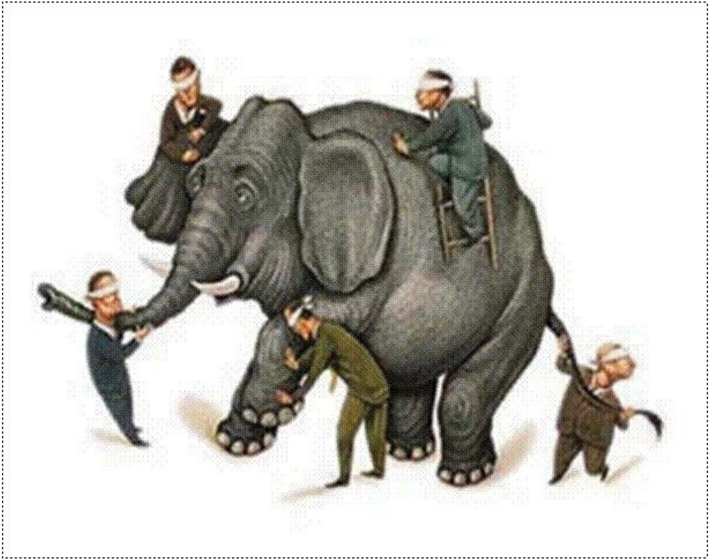
Posterior locations of knots
Profit
Market2Book
Prior mean of surface knots
Prior mean of additive knots
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
−3.5
0
5
10
15
20
25
30
35
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00

Modeling nonlinear mean with splines to firm leverage data

↬ Posterior mean surface(left) and standard deviation(right)



Can we have a model that is big like an elephant?



by John Godfrey Saxe (1816-1887)

Knowing the elephant

- Sophisticated models are essential for such situations.
- In principle, the complicated model should be able to capture more complicated data features.
- Estimating such model is not easy.
- There is huge space to explore.
    - The computer is still not fast enough.
    - Techniques like parallel computing should be used to speed up the computation.
    - Statistics with big data is the new challenge.

References

Casella, G., Girón, F. J., Martínez, M. L. & Moreno, E. (2009). Consistency of bayesian procedures for variable selection. *The Annals of Statistics* , 1207–1228.

Escobar, M. D. (1994). Estimating Normal Means with a Dirichlet Process Prior. *Journal of the American Statistical Association* **89**, 268.

Holmes, C. C. & Mallick, B. K. (2003). Generalized Nonlinear Modeling With Multivariate Free-Knot Regression Splines. *Journal of the American Statistical Association* **98**, 352–368.

Jiang, W. & Tanner, M. a. (1999). On the approximation rate of hierarchical mixtures-of-experts for generalized linear models. *Neural computation* **11**, 1183–98.

Joe, H. (1997). *Multivariate models and dependence concepts*. Chapman & Hall, London.

Kass, R. & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Li, F. & Villani, M. (2013). Efficient Bayesian multivariate surface regression. *Scandinavian Journal of Statistics* **in press**.

Nott, D. & Kohn, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika* **92**, 747–763.

Patton, A. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis* **110**, 4–18.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris* **8**, 229–231.

...essentially, all models are wrong, but some are useful

— George E. P. Box

Thank you!