Bayesian Modeling of Conditional Densities
—Presentation on the Cramér Society 2014 Annual Meeting

Feng Li

<feng.li@cufe.edu.cn>

BEFORE:
Department of Statistics, Stockholm University

NOW:
School of Statistics and Mathematics
Central University of Finance and Economics, Beijing

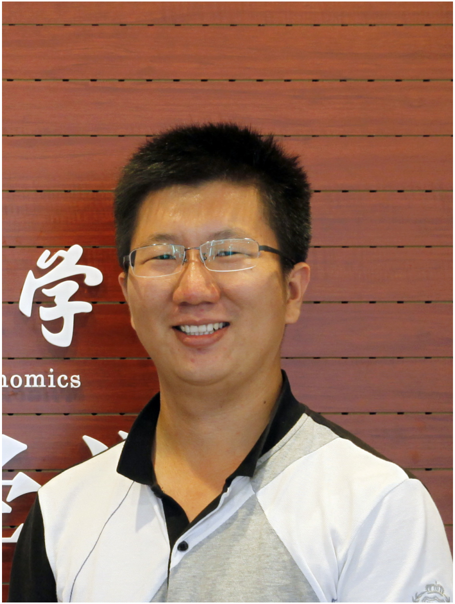


Figure: This is how I look now.

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, texts, brain image...
  - Data volume from 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 and 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.

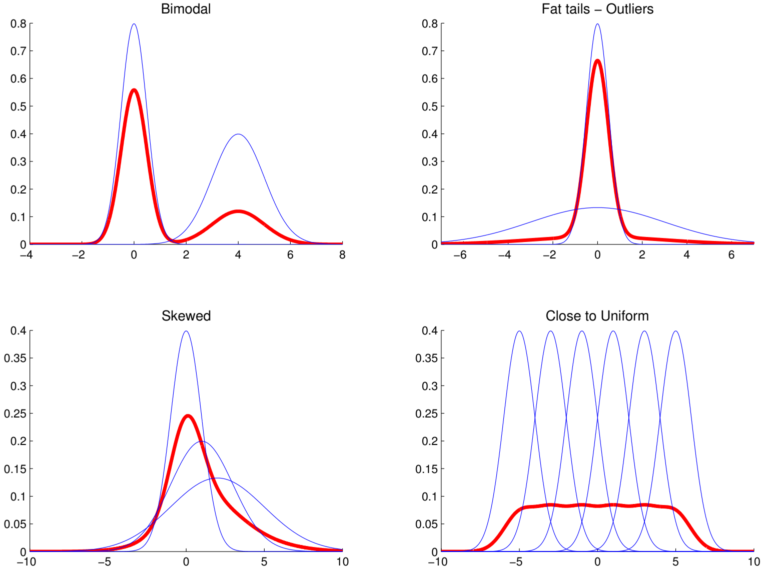


Figure: Using mixture of normal densities (thin lines) to mimic a flexible density (bold line).

One-dimensional conditional density estimation with mixtures

- 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

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

One-dimensional conditional density estimation with splines

- 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 and 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 ↪ The efficient 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 $\mathrm{df} > 2$,

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

  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 and 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

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

Bayesian model comparison

- Bayesian model comparison have historically been based on the marginal likelihood, e.g. **Bayes factor** (Kass and Raftery, 1995).
- However,that the marginal likelihood is very sensitive to the specification of prior.
- The marginal likelihood is also difficult to compute for complicated models.
- A more prominent tool for model comparisons is based on the **log predictive density score** (LPDS)

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

- The predictive density eliminates the inference from prior by integrating out the posterior.

The multivariate surface model ↪ The model

- Splines are regression models with flexible **mean functions** by selecting and placing knots to covariates space.
- The multivariate surface spline model (Li and 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.

The multivariate surface model ↪ The prior

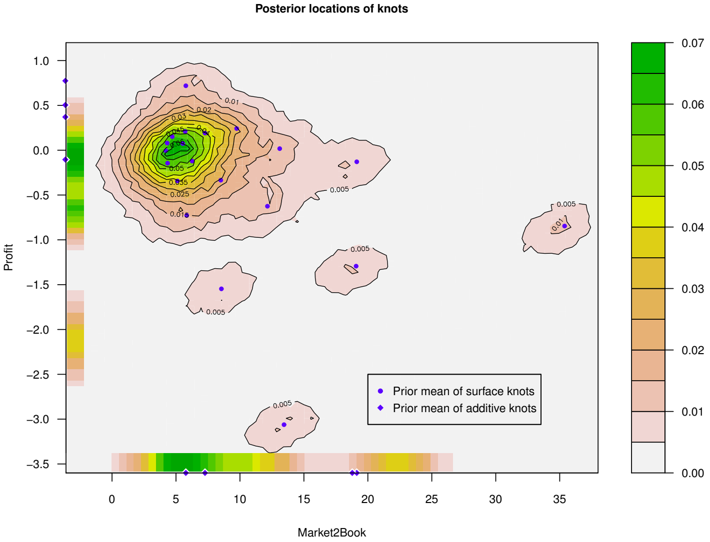
- Conditional on the knots, the prior for $\mathbf{B}$ and $\boldsymbol{\Sigma}$ are set as

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

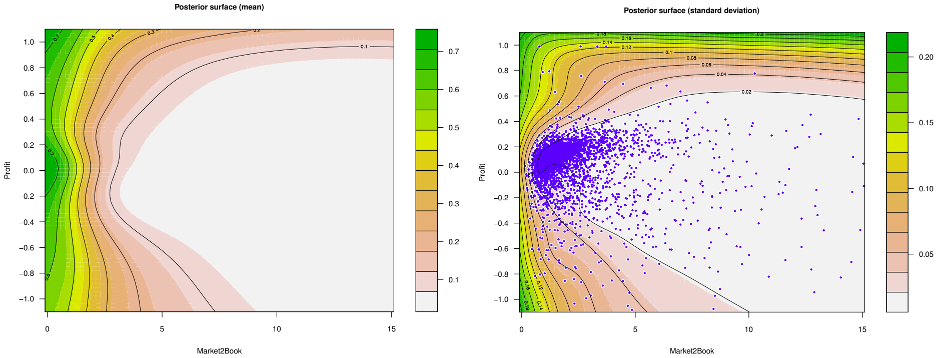
  - $\boldsymbol{\Lambda}_i = diag(\boldsymbol{\lambda}_i)$ are called the shrinkage parameters, which is used for overcome overfitting through the prior.
  - A small $\boldsymbol{\lambda}_i$ shrinks the variance of the conditional posterior for $\mathbf{B}_i$
  - It is another approach to selection important variables (knots) and components.
- The shrinkage parameters are estimated in MCMC
- 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.

Modeling nonlinear mean with splines to firm leverage data

↪ The posterior locations for knots



Modeling nonlinear mean with splines to firm leverage data
↪ Posterior mean surface(left) and standard deviation(right)



Dependence for high-dimensional density with continuous and discrete margins

- In principle, a high dimensional density can be construed via bivariate copulas and their margins.

$$\prod_{k=1}^{M} f_k(x_k) \times$$
$$\prod_{i=1}^{M-1} \prod_{j=1}^{M-i} c_{i,i+j|1:(i-1)}(F(x_i)|x_1, ..., x_{i-1}, F(x_{i+j}|)x_1, ..., x_{i-1})$$

- However this construction depends on the order of the margins.
- The reversible jump MCMC used is not efficient.
- Estimate high-dimensional tail-dependencies are more complicated.

Surface maximization

- The predictive density can be viewed as a **dynamic probability surface** conditional on X

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

- Where is the maximum point of the surface?

$$x_{best} = \text{argmin}_x \int a(f, x) dF(x) \quad (1)$$

where $a(f, x)$ is called the **acquisition function**.

- This approach is called **Bayesian Global Optimization**.
- Used mostly in engineering but not in statistics.

References I

Casella, G., Girón, F. J., Martínez, M. L., and 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. and Mallick, B. K. (2003), "Generalized Nonlinear Modeling With Multivariate Free-Knot Regression Splines," *Journal of the American Statistical Association*, 98, 352–368.

Jiang, W. and 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. and Raftery, A. (1995), "Bayes factors," *Journal of the American Statistical Association*, 90, 773–795.

References II

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

Nott, D. and 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.

Thank you!