Bayesian Multivariate Density Estimation

— with the approach of copula methods

Feng Li

Department of Statistics, Stockholm University

April, 2013

Outline of the talk

1. Flexible density estimation
2. The challenge in multivariate density estimation
3. The multivariate density estimation with copulas

Flexible density estimation

↛ Introduction

- Density estimation consecrates on modeling the relationship between the response $\mathbf{y}$ with covariates $\mathbf{x}$ with flexible density function $f(\cdot)$

$$\mathbf{y} = f(\mathbf{x}, \theta)$$

  flexible: the density feature θ are modeled in a flexible way.

- An example: GLM: density estimation with flexible mean function $\eta(\mu) = \mathbf{X}\boldsymbol{\beta}$ via the linkage.
- Two main factors that influence the efficiency of the density estimation,
  (1) choice of flexible densities, and
  (2) ways of constructing densities features.

Flexible density estimation

↛ Univariate or multivariate

- (Relevantly) simpler in univariate response
  - Mixture of experts
  - Nonparametric methods: kernel regression, splines...
- More tricky in the multivariate case
  - Flexible multivariate density is difficult to construct *per se*
  - Not only modeling the density features in each marginal model
  - But also multivariate correlations and other dependences need to take into account.

Features of interest in multivariate densities

Besides the features of interest in each marginal density, dependence is the never ending story in multivariate densities.

- The general measure of correlation: **Kendall's** τ

$$\tau = 4 \int \int F(x_1, x_2) dF(x_1, x_2) - 1$$

- The dependence in the tail

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

However estimating these features are very difficult in standard multivariate density settings. No general approach until the next slide.

Copulas

- **Sklar's theorem** Let H be a multi-dimensional distribution function with marginal distribution functions $F_1(x_1), ..., F_m(x_m)$. Then there exists a function C (**copula function**) such that

$$\begin{aligned} H(x_1, ..., x_m) =& C(F_1(x_1), ..., F_m(x_m)) \\ =& C\left(\int_{-\infty}^{x_1} f(z_1)dz_1, ..., \int_{-\infty}^{x_m} f(z_m)dz_m\right) = C(u_1, ..., u_m). \end{aligned}$$

- Here is the magic

$$\tau = 4\int\int F(x_1, x_2)dF(x_1, x_2) - 1 = 4\int\int C(u_1, u_2)dC(u_1, u_2) - 1.$$

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

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

The covariate-dependent copula model

- The multivariate density (in terms of copulas) features to be connected with observed data information. In particular, correlation and tail-dependence are the two concepts of interest in various of situation,

$$\tau = \eta_\tau^{-1}(\mathbf{X}\boldsymbol{\beta}_\tau), \text{ and } \lambda = \eta_\lambda^{-1}(\mathbf{X}\boldsymbol{\beta}_\lambda)$$

- The marginal models are constructed via the usual approach in univariate density (mixtures, splines, etc)
- The Bayesian approach

$$\begin{aligned}\log p(\{\boldsymbol{\beta}, \mathfrak{I}\}|\mathbf{Y}, \mathbf{X}) =& \text{constant} + \sum_{j=1}^{M} \log p(\mathbf{Y}_{\cdot j}|\{\boldsymbol{\beta}, \mathfrak{I}\}_j, \mathbf{X}_j) \\ &+ \log \mathcal{L}_C(\mathbf{u}|\{\boldsymbol{\beta}, \mathfrak{I}\}_C, \mathbf{Y}, \mathbf{X}) + \log p(\{\boldsymbol{\beta}, \mathfrak{I}\})\end{aligned}$$

  - Efficient MCMC to sample the posterior (tailored Metropolis-Hastings)
  - Bayesian variable selection is integrate seamlessly.
  - Model comparison and copula density selection via predictive likelihood.

Thank you!

feng.li@stat.su.se