Nonparametric Multivariate Density Estimation Using Mixtures

Xuxu Wang and Yong Wang

Department of Statistics, The University of Auckland, New Zealand

Sep 12, 2013

Nonparametric Density Estimation and Its Applications

- Estimation of a density function nonparametrically from multivariate observations
- Applications
  - Biostatistics (e.g., Duong and Hazelton (2005))
  - Medicine (e.g., Hastie et al. (2009))
  - Finance (e.g., Yuan (2009))

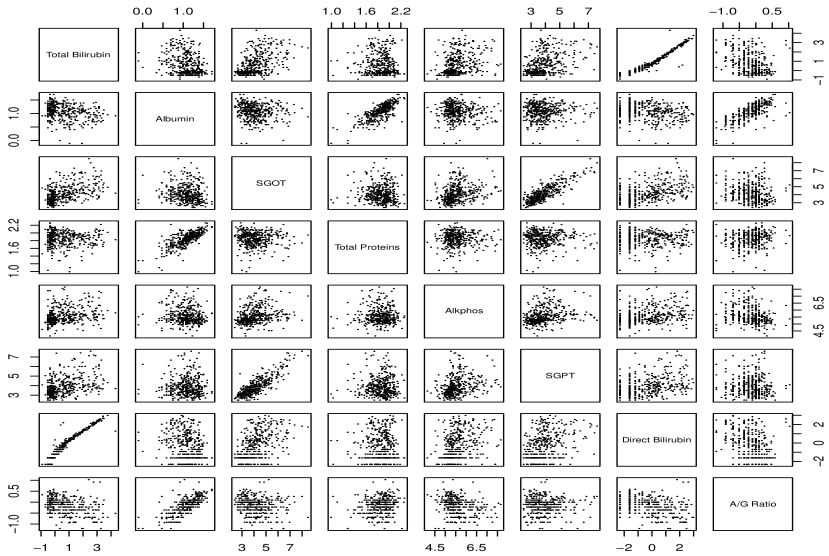
Existing Approaches

- Kernel-based density estimation (KDE)
  - Bandwidth selection
  - Two multivariate selectors with full bandwidth matrices
    - The plug-in (PI) selector of Duong and Hazelton (2003)
    - The smooth cross-validation (SCV) selector of Duong and Hazelton (2005)
- Mixture-based Density Estimation (MDE)
  - Advantages of the MDE
  - Univariate MDE (Wang and Chee, 2012; Chee and Wang, 2012, 2013)
  - Difficulties in fitting multivariate nonparametric or semiparametric mixtures

Practical Problems

- Patients with liver disease continuously increase.
- An early diagnosis can increase patients survival rate.
- The Indian Liver Patient data (416 patients with 8 variables)
- Our interest: estimating the distribution for the patient data accurately

The Scatterplot for The Indian Liver Patient Data



Nonparametric and Semiparametric Mixture Models

- A nonparametric mixture model has a density of the form

$$f(x; G) = \int_{\Omega} f(x; \theta)\, \mathrm{d}G(\theta),$$

where $f(x;\theta)$, $x \in \chi$, $\theta \in \Omega \subset \Re$, is the component density, and $G(\theta)$ the mixing distribution function.

- A nonparametric mixture can be extended to a semiparametric mixture by incluing a finite-dimensional parameter β, which is of the form

$$f(x; G, \beta) = \int_{\Omega} f(x; \theta, \beta)\, \mathrm{d}G(\theta),$$

where $\beta \in \Re^r$ is common to all components.

Maximum Likelihood Estimate

- There always exists a discrete nonparametric maximum likelihood estimate (NPMLE) $\hat{G}$ (Lindsay, 1983a,b).
- For a discrete G,

$$G(\theta) = \sum_{j=1}^{m} \pi_j \delta_{\theta_j},$$

  where $\theta_j \in \Omega$ and $\pi_j > 0$ for $j = 1, \ldots, m$, $\sum_{j=1}^{m} \pi_j = 1$, and δ_{θ_j} puts mass 1 at θ_j.
- The log-likelihood function of G and that of (G, β) can be written as

$$l(G) = \sum_{i=1}^{n} \log\{f(x; G)\},$$

$$\text{and } l(G, \beta) = \sum_{i=1}^{n} \log\{f(x; G, \beta)\}.$$

Computation of Maximum Likelihood Estimate

- For computing the NPMLE, the gradient function plays a critical role, which has a form

$$d(\theta; G) = \sum_{i=1}^{n} \frac{f(x_i; \theta)}{f(x_i; G)} - n.$$

- The general equivalent theorem:
  $\hat{G}$ maximizes $l(G) \Leftrightarrow \hat{G}$ minimizes $\sup_\theta \{d(\theta; G)\} \Leftrightarrow \sup_\theta \{d(\theta; \hat{G})\} = 0$.
- For computing the semiparametric MLE, another condition

$$\frac{\partial l(\hat{G}, \beta)}{\partial \beta} = 0,$$

  needs to be met.

Algorithms

- Wang (2007) proposed the constrained Newton method with multiple support points (CNM) for computing a NPMLE. At each iteration, it
  - adds all local maxima of the gradient function to support points set;
  - updates all mixing proportions via a quadratically convergent method;
  - gets new support points by discarding support points with mass 0.
- For computing a semiparametric MLE, Wang (2010) proposed three general algorithms by combining the CNM with an optimization algorithm for computing $\hat{\beta}$.

Density Estimation

- An univariate nonparametric MDE is defined by

$$f_h(x; \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{j=1}^{m} \pi_j \, K_h(x - \theta_j),$$

  where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\mathsf{T} \in \mathbf{R}^m$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)^\mathsf{T}$, with $\pi_j > 0$ for $j = 1, \ldots, m$, $\sum_{j=1}^m \pi_j = 1$.
- Likelihood maximization
- Bandwidth selection

Nonparametric MDE

- The extension of the univariate MDE to vectorized data, $x \in \Re^d$, is straight forward for the multivariate MDE (H-fixed MDE),

$$f_{\mathbf{H}}(x; G) = \int_{\Omega} f_{\mathbf{H}}(x; \boldsymbol{\theta}) \, \mathrm{d}G(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} \in \Omega \subset \Re^d$ and **H** is the bandwidth matrix, being symmetric, positive-definite and common to all the components.

- The log-likelihood function of G with **H** fixed is given by

$$l_{\mathbf{H}}(G) = \sum_{i=1}^{n} \log \left\{ \int_{\Omega} f_{\mathbf{H}}(x_i; \boldsymbol{\theta}) \, \mathrm{d}G(\boldsymbol{\theta}) \right\}.$$

- For any fixed **H**, a discrete multivariate NPMLE $\hat{G}$ alway exists among all G (Lindsay, 1983a,b, 1995).

Difficulties in Estimating Nonparametric MDE

- When maximizing the likelihood function with **H** simply treated as an argument, the bandwidth matrix becomes singular and the likelihood approaches infinity.
- Difficulty to estimate the nonparametric MDE directly, as **H** has $(d^2 + d)/2$ unknown elements
- Computational unfeasibility to select the entire **H** via cross-validation or model selection when d increases

Decomposition of the **H**

- A decomposition of the **H** is considered,

$$\mathbf{H} = h^2 B, \text{ subject to } |B| = 1,$$

where $h = |\mathbf{H}|^{\frac{1}{2d}}$ and B is a symmetric and positive-definite matrix.

  - h determines the volume of **H** and B determines its shape and orientations.
  - h controlls the smoothness of the density.
  - h is similar to the bandwidth scalar in the univariate case.

Semiparametric MDE

- With h fixed, the mixture density with a discrete G becomes

$$f_h(x; G, B) = \sum_{j=1}^{m} \pi_j \, f_h(x; \boldsymbol{\theta}_j, B).$$

- The log-likelihood function of (G, B) with h fixed is given by

$$l_h(G, B) = \sum_{i=1}^{n} \log \left\{ f_h(x; G, B) \right\}.$$

- With any fixed h, the log-likelihood of (G, B) is bounded by

$$l_h(G, B) \leq -\frac{nd}{2} \log(2\pi h^2).$$

- The estimation procedure consists of two steps.

Volume Selection

- A sequence of semiparametric mixtures is defined through controlling h-value and profiling the likelihood function.
- How to choose an appropriate h-value?
- The information-theoretic model selection criteria

$$\text{AIC}(h) = -2\tilde{l}(h) + 2p,$$

where $\tilde{l}(h) \equiv \max_{G,B} l_h(G, B)$ is the profile log-likelihood function of h and p the number of free parameters including h.

A Hybrid Approach

- Difficulty to use a single optimization algorithm to find the MLE $(\hat{G}_h, \hat{\mathbf{B}}_h)$
- The hybrid algorithm
  - (i) The expectation-maximisation (EM) algorithm for updating $\boldsymbol{\pi}$, $\boldsymbol{\Theta}$ and $\mathbf{B}$ (Dempster et al., 1977)
  - (ii) The constrained-Newton method (CNM) for updating G nonparametrically (Wang, 2007)

Updating $\boldsymbol{\pi}$, $\boldsymbol{\Theta}$ and $\mathbf{B}$

- The log-likelihood function for a homoscedastic finite mixture is given by

$$l(\boldsymbol{\pi}, \boldsymbol{\Theta}, \mathbf{H}) = \sum_{i=1}^{n} \log \left\{ f(x_i; \boldsymbol{\pi}, \boldsymbol{\Theta}, \mathbf{H}) \right\}.$$

- The EM iteration formulae under the restriction $|\mathbf{B}| = 1$ are given by

$$\pi_j' = \frac{1}{n} \sum_{i=1}^{n} p_{ij}, \quad \boldsymbol{\theta}_j' = \frac{\sum_{i=1}^{n} p_{ij} x_i}{\sum_{i=1}^{n} p_{ij}},$$

$$\mathbf{H}' = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}(x_i - \boldsymbol{\theta}_j)(x_i - \boldsymbol{\theta}_j)^\top, \quad \mathbf{B}' = \mathbf{H}'/|\mathbf{H}'|^{\frac{1}{d}},$$

where $p_{ij} = \frac{\pi_j f_{ij}}{\sum_{l=1}^{m} \pi_l f_{il}}$ and $f_{ij} = f(x_i; \boldsymbol{\theta}_j, \mathbf{H})$.

- $(\boldsymbol{\pi}, \boldsymbol{\Theta}, \mathbf{H})$ is updated by one iteration to $(\boldsymbol{\pi}', \boldsymbol{\Theta}', \mathbf{H}')$.

Updating G

- Roughly locating new support points using a random grid for gradient valuation
- How to generate such a random grid?

$$d_{\mathbf{H}}^*(\boldsymbol{\theta}; G_t) = C^{-1} \sum_{i=1}^{n} w_i f(x_i; \boldsymbol{\theta}, \mathbf{H}),$$

where $w_i = f_{\mathbf{H}}(x_i; G_t)^{-1}$ and $C = \sum_{i=1}^{n} w_i \int_\Omega f(x_i; \boldsymbol{\theta}, \mathbf{H}) \, \mathrm{d}\boldsymbol{\theta}$.

- With a given $\boldsymbol{\Theta}$, the mixing proportion vector $\boldsymbol{\pi}$ is updated by maximizing $l_h(\boldsymbol{\pi}', \boldsymbol{\Theta}, \mathbf{B})$ according to the second-order Taylor series expansion about $\boldsymbol{\pi}$,

$$l_h(\boldsymbol{\pi}', \boldsymbol{\Theta}, \mathbf{B}) \approx l_h(\boldsymbol{\pi}, \boldsymbol{\Theta}, \mathbf{B}) - \frac{1}{2}\|\mathbf{S}\boldsymbol{\pi}' - \mathbf{2}\|^2 + \frac{n}{2}.$$

Five Estimatorss

- Four bandwidth matrix selection methods of the KDE
  - Two with diagonal matrices: the product kernel estimator (PD) (Scott, 1992) and the adaptive kernel estimator (AD) (Silverman, 1986)
  - Two with full matrices: the plug-in selector (PI) (Duong and Hazelton, 2003) and the smooth cross-validation selector (SCV) (Duong and Hazelton, 2005)
- MDE
  - A grid of 10 potential h-values that were evenly distributed from $\sqrt[d]{0.1}s$ to s
  - d is the dimensionality and s the volume parameter value of the sample covariance matrix.

Two Performance Measures

- The mean integrated square error and the mean Kullback-Leibler divergence:

$$\text{ISE}(\hat{f}_m, \hat{f}) = \int_{\mathbb{R}} \{\hat{f}(x)\}^2 \mathrm{d}x - \frac{2}{m}\sum_{i=1}^{m} \hat{f}(x_i)$$

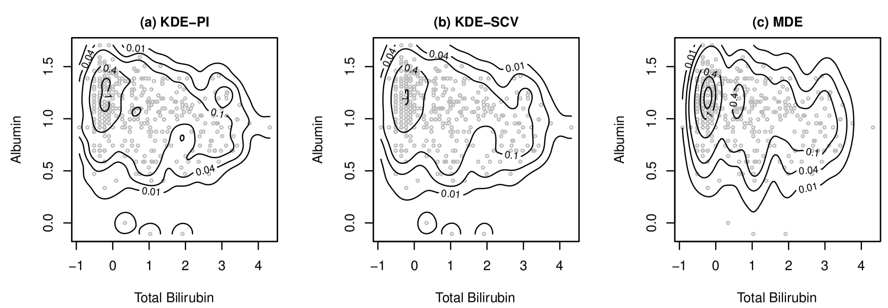
$$\text{KL}(\hat{f}_m, \hat{f}) = -\frac{1}{m}\sum_{i=1}^{m} \log\{\hat{f}(x_i)\}$$

- $\hat{f}$ denotes a density estimate from a training set and $\hat{f}_m$ the empirical mass function based on a test set of size m.
- 10-fold cross-validation with 10 repetitions

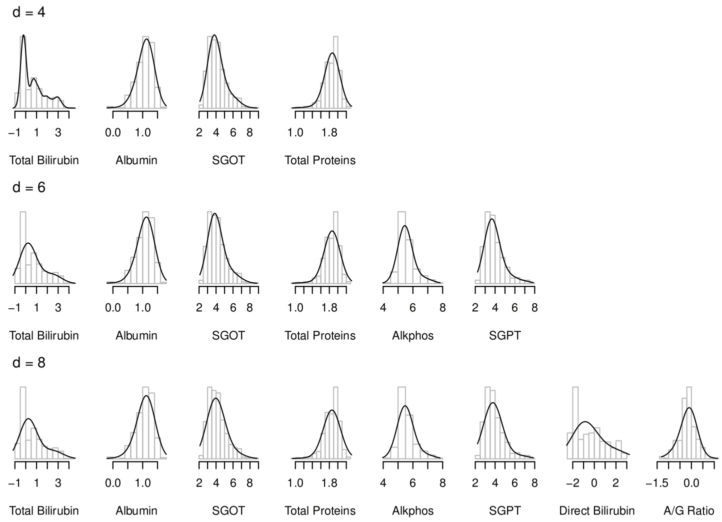
Statisitical Results

	KDE-AD	KDE-AD	KDE-PI	KDE-SCV	MDE
			$d = 2$		
MISE	−0.538 (0.006)	−0.626 (0.009)	−0.659 (0.010)	−0.634 (0.009)	−**0.724** (0.012)
MKL	1.383 (0.014)	1.348 (0.017)	1.331 (0.021)	1.330 (0.019)	**1.298** (0.020)
			$d = 4$		
MISE	−0.484 (0.008)	−0.710 (0.026)	−0.801 (0.012)	−0.700 (0.010)	−**0.866** (0.042)
MKL	2.254 (0.033)	1.927 (0.042)	2.187 (0.050)	2.025 (0.038)	**1.775** (0.090)
			$d = 6$		
MISE	−0.123 (0.005)	−0.419 (0.033)	—	—	−**0.458** (0.021)
MKL	3.820 (0.051)	3.636 (0.073)	—	—	**3.441** (0.082)
			$d = 8$		
MISE	−0.053 (0.001)	−0.392 (0.033)	—	—	−**0.615** (0.032)
MKL	4.807 (0.064)	4.493 (0.089)	—	—	**3.286** (0.190)

Contour Plots of Density Estimates for The Bivariate Indian Liver Patient Data

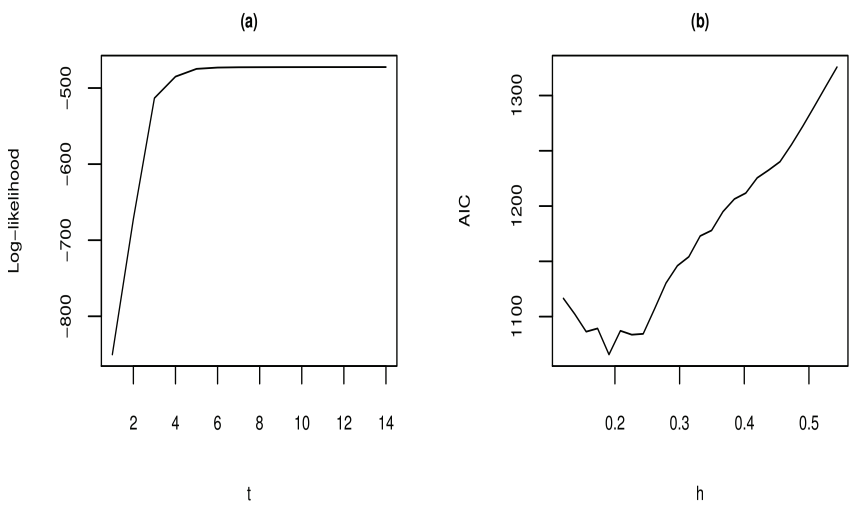


Univariate Marginal Densities by The MDE for The Indian Liver Patient Data



Log-likelihood Path and The AIC Path

Log-likelihood path for the bivariate Indian Liver Patient data, and the AIC path for various choices of h-values.



Computation Times (mm:ss)

d	KDE-PI	KDE-SCV	MDE
2	0:01	0:02	1:31
4	12:28	13:50	4:46
6	—	—	9:13
8	—	—	9:04

- For the MDE, it includes the time needed for all 10 h-values defined by the grid evenly spaced from $\sqrt[d]{0.1}s$ to s.
- When $d = 2$, the MDE procedure was much slower.
- when $d = 4$, the time needed by a KDE increases dramatically and the MDE requires only about a third of the time needed by a KDE.
- When $d = 6$ or 8, it becomes computationally too costly for the KDE's to produce solutions in a reasonable time (more than 2 hours).

Future Work

More types of h-fixed MDE

- In future research, by considering specific assumptions on the orientations and the shape of the bandwidth matrix, more types of h-fixed MDE can be obtained.

Volume selection methods

- No reliable theories are established for applying the AIC to mixtures.
- Developing efficient model selection criteria to mixtures represents a key direction for future research.

Heteroscedastic mixtures

- In future research, mixtures with heteroscedastic components will be investigated due to its obvious merits for irregular multivariate data sets.

Summary

- Outline the multivariate nonparametric mixture-based density estimator and its attributes.
- Propose the *h*-fixed semiparametric mixtures for density estimation as an alternative to the kernel-based nonparametric approaches.
- A general methodology for using the *h*-fixed semiparametric mixtures in multivariate density estimation has been investigated.
- The information-theoretic model selection criteria
- Satisfactory performance in real-world examples

Chee, C.-S. and Y. Wang (2012). Estimation of finite mixtures with symmetric components. *Statistics and Computing 23*, 233–249.

Chee, C.-S. and Y. Wang (2013). Minimum quadratic distance density estimation using nonparametric mixtures. *Computational Statistics and Data Analysis 57*, 1–16.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B 39*, 1–38.

Duong, T. and M. L. Hazelton (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics 15*, 17–30.

Duong, T. and M. L. Hazelton (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics 32*, 485–506.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Lindsay, B. G. (1983a). The geometry of mixture likelihoods: A general theory. *Annals of Statistics 11*, 86–94.

Lindsay, B. G. (1983b). The geometry of mixture likelihoods, part II: The exponential family. *Annals of Statistics 11*, 783–792.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, Volume 5, Hayward. Institute for Mathematical Statistics.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Wang, Y. (2007). On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. *Journal of the Royal Statistical Society, Ser. B 69*, 185–198.

Wang, Y. (2010). Maximum likelihood computation for fitting semiparametric mixture models. *Statistics and Computing 20*, 75–86.

Wang, Y. and C.-S. Chee (2012). Density estimation using nonparametric and semiparametric mixtures. *Statistical Modelling 12*, 67–92.

Yuan, M. (2009). State price density estimation via nonparametric mixtures. *Annals of Applied Statistics 3*, 963–984.