A Multivariate Likelihood-tuned Density Estimator and Modal Inference

Yeojin Chung and Bruce G. Lindsay
Department of Statistics, The Pennsylvania State University

PENNSTATE 1855

Summary

We consider an improved multivariate nonparametric density estimator which arises from treating the kernel density estimator as an element of the model that consists of all mixtures of the kernel, continuous or discrete. One can obtain the kernel density estimator with "likelihood-tuning" by using the uniform density as the starting value in an EM algorithm. The second tuning leads to a fitted density with higher likelihood than the kernel density estimator. The two-step likelihood-tuned density estimator reduces asymptotic bias. The new density estimator captures interesting signals better than the others but does not have clear improvement in root MSE. We compare the performance of the new density estimator with other modified density estimators in higher dimensions.

Introduction

Let $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n \in \mathbb{R}^d$ be a random sample from $f(\boldsymbol{x})$ and $K_H(\cdot,\cdot)$ be a kernel function with a bandwidth matrix $H = h^2AA^T$ $(A \in \mathbb{R}^{d\times d})$.

Multivariate kernel density estimator

$$\hat{f}_{KER}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} K_H(\boldsymbol{x}, \boldsymbol{X}_i) \qquad (1)$$

- Bias of order $O\left(h^2\right)$, variance of order $O\left(n^{-1}h^{-d}\right)$.

Several modifications of kernel density estimators have been proposed to reduce the bias of $\hat{f}_{KER}$.

Adaptive bandwidth density estimator (Hall and Marron (1988))

$$\hat{f}_{ABW}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} K_{h^2 f(\boldsymbol{X}_i)^{-1}I}(\boldsymbol{x}, \boldsymbol{X}_i)$$

Multiplicative bias correction density estimator (Jones et al. (1995))

$$\hat{f}_{MBC}(\boldsymbol{x}) = \hat{f}_{KER}(\boldsymbol{x})\frac{1}{n}\sum_{i=1}^{n}\frac{K_H(\boldsymbol{x}, \boldsymbol{X}_i)}{\hat{f}_{KER}(\boldsymbol{X}_i)}$$

- Both have a bias of order $O(h^4)$ and a variance of order $O(n^{-1}h^{-d})$

In this poster,

- **we propose a new density estimator with higher likelihood than the kernel density estimator.**
- **The new density estimator has a bias of order $O(h^4)$ and a variance of order $O(n^{-1}h^{-d})$ as $\hat{f}_{ABW}$ and $\hat{f}_{MBC}$.**
- **The new density estimator captures interesting signals better than the others but does not have clear improvement in root MSE.**

Nonparametric Maximum Likelihood Estimator

Nonparametric mixture model

$$f(\boldsymbol{x}; \Pi) = \int K(\boldsymbol{x}, \boldsymbol{\phi})d\Pi(\boldsymbol{\phi})$$

$\boldsymbol{\phi}$: a latent variable
$\Pi(\boldsymbol{\phi})$: a mixing (latent) distribution
$K(\cdot, \boldsymbol{\phi})$: a known density function

The likelihood function on an observation: $L_i(\Pi) = \int K(\boldsymbol{X}_i, \boldsymbol{\phi})d\Pi(\boldsymbol{\phi})$
The log-likelihood function with multiple observations : $l(\Pi) = \sum_{i=1}^{n}\ln(L_i(\Pi))$.

Nonparametric maximum likelihood estimator (NPMLE) of Π: $\hat{\Pi} = \arg\max_{\Pi} l(\Pi)$

Lindsay (1995) shows that there exists a maximum likelihood estimator $\hat{\Pi}$ that is a discrete distribution with no more than n distinct points of support.

Likelihood-tuning Procedure

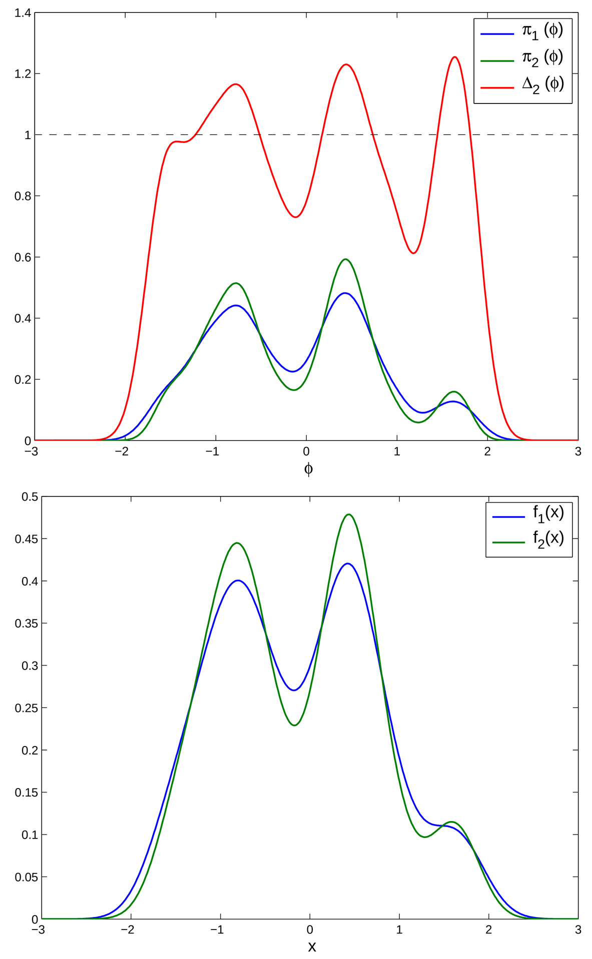


Figure 1: Illustration of a likelihood-tuning procedure

IDEA

1. **Start from the uniform initial $\pi(\phi)$ (no prior information)**
2. **Do EM steps to update π (give information about where we need more mass)**
3. **Fit a density $f(x)$ with updated $\pi(\phi)$.**

Let π be a density function corresponding to Π. Given an initial estimate π_0,

1. the continuous EM algorithm (Vardi and Lee(1993)) updates π by

$$\pi_{(k+1)}(\boldsymbol{\phi}) = \pi_{(k)}(\boldsymbol{\phi})\Delta_{(k)}(\boldsymbol{\phi})$$

where $\Delta_{(k)}(\boldsymbol{\phi}) = n^{-1}\sum_{i=1}^{n}\frac{K(\boldsymbol{X}_i,\boldsymbol{\phi})}{f_{(k)}(\boldsymbol{x}_i)}$.

2. Update the density estimator by

$$\hat{f}_{(k+1)}(\boldsymbol{x}) = \int K(\boldsymbol{x}, \boldsymbol{\phi})\pi_{(k+1)}(\boldsymbol{\phi})\, d\boldsymbol{\phi}.$$

Consider $K_H(\boldsymbol{x}, \boldsymbol{X}_i) = H^{-1/2}\varphi\left\{H^{-1/2}(\boldsymbol{x} - \boldsymbol{X}_i)\right\}$ where $\varphi(\cdot)$ is the independent standard normal density.

- Initial estimate of π: $\hat{\pi}(\boldsymbol{\phi}) = 1$ on $\mathbb{R}^d$ - "improper prior"
- The first tuning step gives $\hat{f}_{EM1}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} K_{2H}(\boldsymbol{X}_i, \boldsymbol{x}) \Rightarrow \hat{f}_{KER}$ with a bandwidth $2H$
- The second tuning step gives

$$\hat{f}_{EM2}(\boldsymbol{x}) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}K_{\frac{3}{2}H}(\boldsymbol{x}, \bar{\boldsymbol{X}}_{ij})$$

where $w_{ij} = K_{2H}(\boldsymbol{X}_i, \boldsymbol{X}_j)/\hat{f}_1(\boldsymbol{X}_i)$ and $\bar{\boldsymbol{X}}_{ij} = \frac{1}{2}(\boldsymbol{X}_i + \boldsymbol{X}_j)$.
Define the second step estimator $\hat{f}_{EM2}(\boldsymbol{x})$ to be **the (second-step) likelihood-tuned density estimator**.

Asymptotic Properties

Theorem 1 *Let $K_H(\boldsymbol{x}, \boldsymbol{\phi}) = |H|^{-1/2}\varphi\left(|H|^{-1/2}(\boldsymbol{x} - \boldsymbol{\phi})\right)$ where $H = h^2AA^T$. Under regularity conditions,*

$$E\left[\hat{f}_{EM2}(\boldsymbol{x})\right] = f(\boldsymbol{x}) + g_1(A)h^4 + o(h^4)$$

and

$$Var\left[\hat{f}_{EM2}(\boldsymbol{x})\right] = f(\boldsymbol{x})\pi^{-\frac{d}{2}}\left(2^{2-\frac{3}{2}d} + 2^{-2d} - 2^{2-d}3^{-\frac{d}{2}}\right)n^{-1}h^{-d} + o\left(n^{-1}h^{-d}\right).$$

as $h \to 0$ and $nh^d \to \infty$.

- $\hat{f}_{EM2}$ has a bias of order $O(h^4)$ and variance of order $O(n^{-1}h^{-d})$ as the other improved density estimators $\hat{f}_{ABW}$ and $\hat{f}_{MBC}$.
- $\hat{f}_{EM2}$ and $\hat{f}_{MBC}$ share the same asymptotic variance.

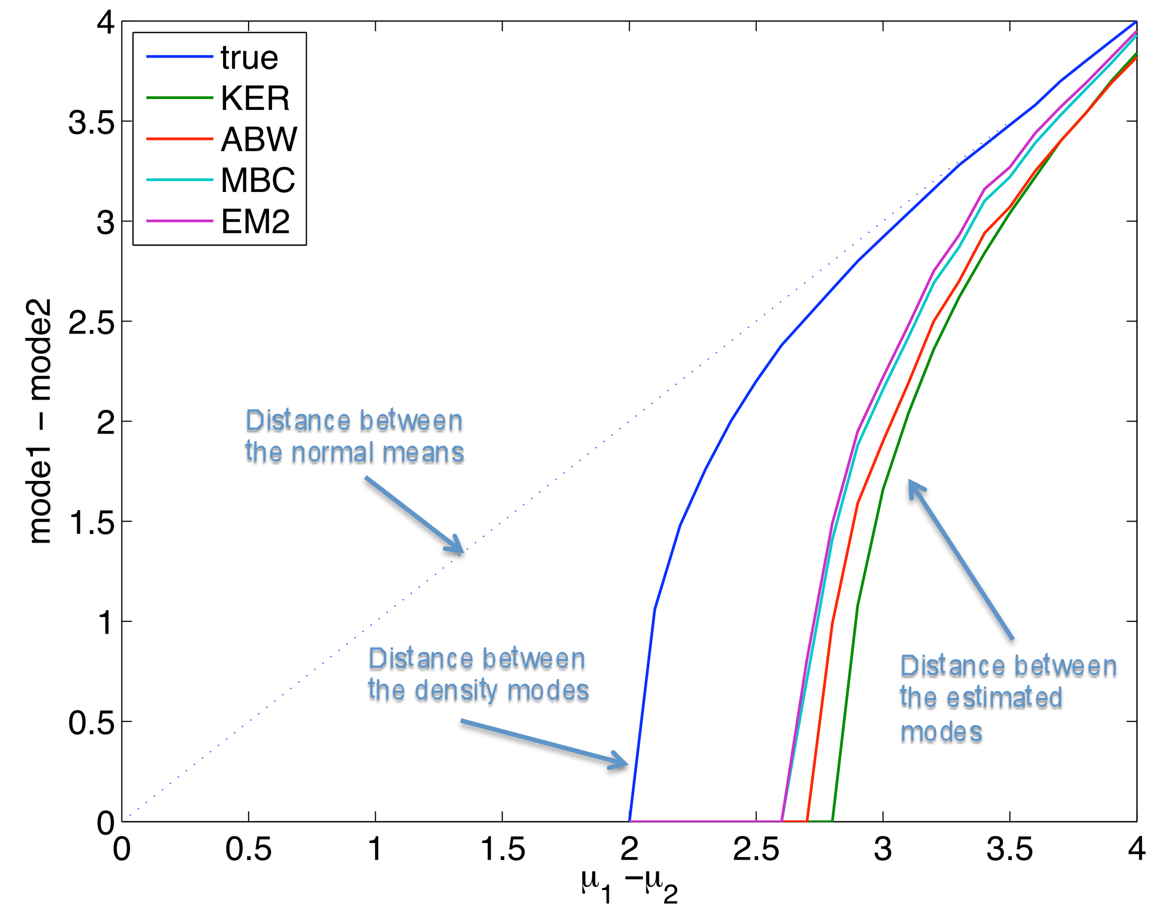
Modal Inference

Q: Do interesting signals persist through smoothing?

Let $X_1 \sim f(x) = 0.5N(\mu_1, 1) + 0.5N(\mu_2, 1)$.

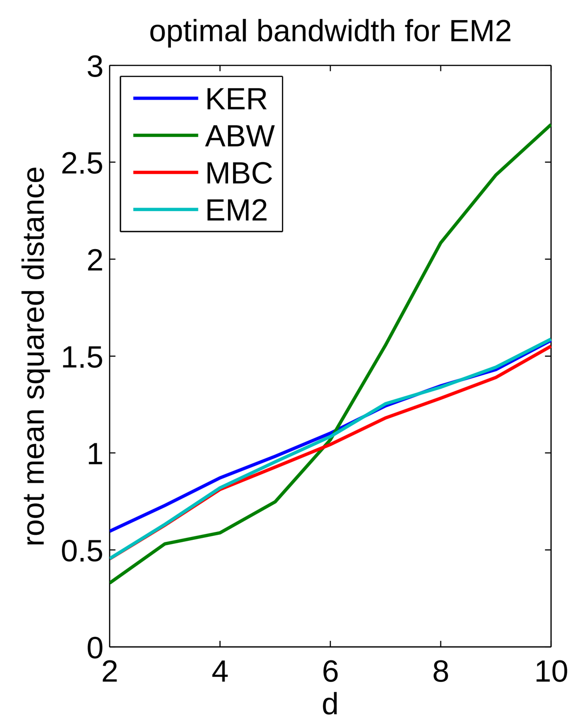
- Modes of the true $f(x)$ separates when $|\mu_1 - \mu_2| > 2$.
- Add coordinates of white noise $(X_1, Z_2, \ldots, Z_d)$.
- How strong signal is needed for the density estimators to capture it?

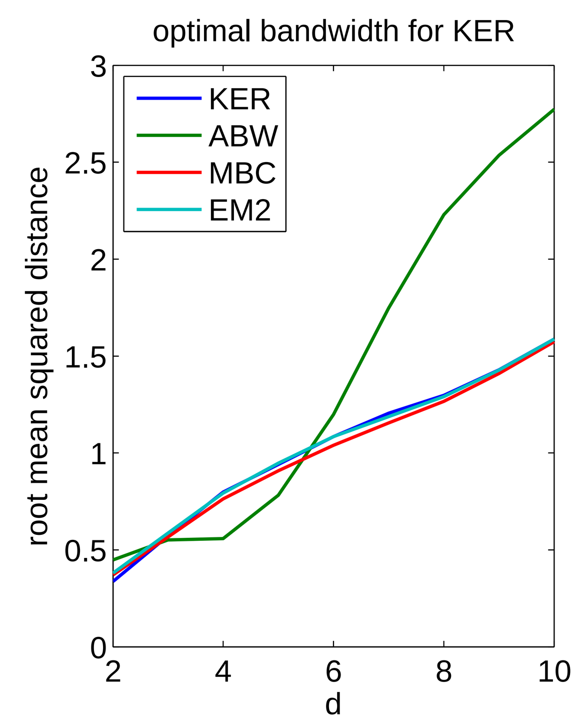
Data n=100, replicate=1000 and dimension=3. The bandwidth $h = 1$.



- $\hat{f}_{MBC}$ and $\hat{f}_{EM2}$ tend to separate the modes more clearly than $\hat{f}_{KER}$ and $\hat{f}_{ABW}$.
- At four standard deviations, all density estimators capture two modes clearly.
- Dimension does not have much influence on separating the modes.

Q: What is the effect of increasing dimension on ability to find modes?
Data n = 100, replication=500, bimodal normal mixture.





- No clear improvement on $\hat{f}_{KER}$ by $\hat{f}_{EM2}$ or others.

References

Hall, P. and Marron, J. "Variable window width kernel estimates of probability densities." *Probability Theory and Related Fields*, 80(1):37–49 (1988).

Jones, M., Linton, O., and Nielsen, J. "A simple bias reduction method for density estimation." *Biometrika*, 82(2):327–338 (1995).

Lindsay, B. *Mixture Models: Theory, Geometry, and Applications*. Ims (1995).

Marron, J. and Wand, M. "Exact mean integrated squared error." *Ann. Statist*, 20(2):712–736 (1992).

Vardi, Y. and Lee, D. "From image deblurring to optimal investments: Maximum likelihood solutions for positive linear inverse problems." *J. Roy. Statist. Soc. B*, 55(3):569–612 (1993).