Estimation of Finite Mixtures with Nonparametric Components

C. S. Chee

Department of Statistics
The University of Auckland
chee@stat.auckland.ac.nz
Joint work with Dr Yong Wang

Part I

General Idea

Consider a mixture distribution having density $g(x)$:

$$g(x) = \sum_{k=1}^{K} \lambda_k f(x - \mu_k)$$

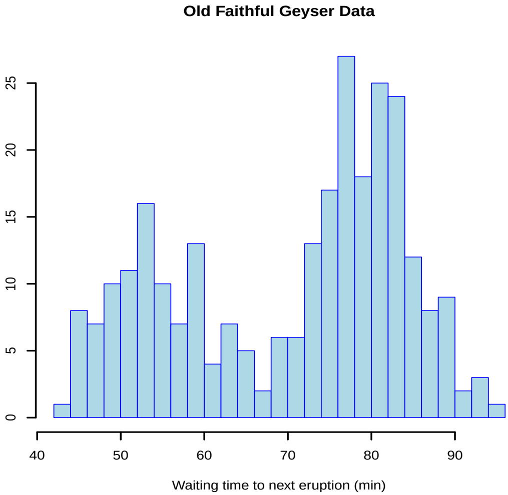
Focus: A two-component mixture of location-shifted distributions ($K = 2$).

Goal: Estimate λ_1, μ_1, μ_2 and f given an iid sample $\{x_i\}_{1 \leq i \leq n}$ from $g(x)$.

Approach: Use a nonparametric mixture for f.

A Running Example

- This dataset contains the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser.
- A two-component mixture model is reasonable.



Identifiability Result

Model

$$g(x) = \lambda f(x - \mu_1) + (1 - \lambda) f(x - \mu_2)$$

- Do not want to assume f belongs to a parametric family.
- **Question**: Is $g(x)$ identifiable?
- $g(x)$ is identifiable if it is unique as a function of λ, μ_1, μ_2 and f.

Identifiability Result

Model

$$g(x) = \lambda f(x - \mu_1) + (1 - \lambda) f(x - \mu_2)$$

- Do not want to assume f belongs to a parametric family.
- **Question**: Is $g(x)$ identifiable?
- $g(x)$ is identifiable if it is unique as a function of λ, μ_1, μ_2 and f.
- Bordes *et al.* (2006) and Hunter *et al.* (2007) showed that $g(x)$ is identifiable if f is a symmetric density about zero when $\lambda \neq \frac{1}{2}$ and $\mu_1 \neq \mu_2$.

The Kernel-based Semiparametric Model

Bordes *et al.* (2007)

$$g(x) = \lambda f(x-\mu_1) + (1-\lambda) f(x-\mu_2)$$

with f being the symmetrized nonparametric kernel given by

$$f(y) = \sum_{i=1}^{n} \frac{1}{2nh} \left\{ K\left(\frac{y - x_i + \mu_{z_i}}{h}\right) + K\left(\frac{y + x_i - \mu_{z_i}}{h}\right) \right\},$$

where

- $K(\cdot)$ — a kernel density function,
- $h > 0$ — the bandwidth,
- $z_i \in \{1,2\}$ — the component label of x_i.

Construction of g

$$g(x) = \lambda f(x - \mu_1) + (1 - \lambda) f(x - \mu_2)$$

$$f(y) = \sum_{i=1}^{n} \frac{1}{2nh} \left\{ K\left(\frac{y - (x_i - \mu_{z_i})}{h} \right) + K\left(\frac{y - [-(x_i - \mu_{z_i})]}{h} \right) \right\}$$

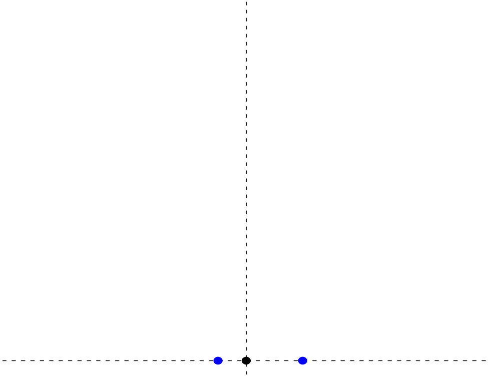
For illustration, consider $h = 1$, $n = 2$ and $K(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$.

Assume that z_i and μ_{z_i} are known.

Let $x_i^\star = x_i - \mu_{z_i}$.

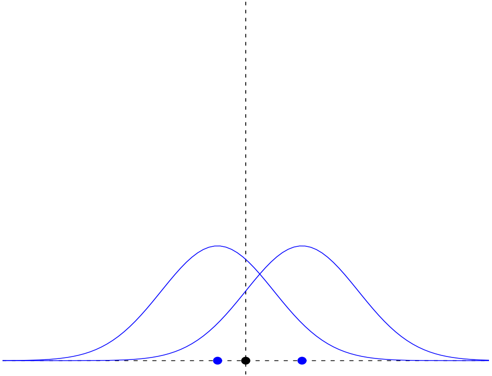
Construction of g

$$f(y) = \sum_{i=1}^{n} \frac{1}{2nh} \left\{ K\left(\frac{y - x_i^\star}{h}\right) + K\left(\frac{y - (-x_i^\star)}{h}\right) \right\}$$



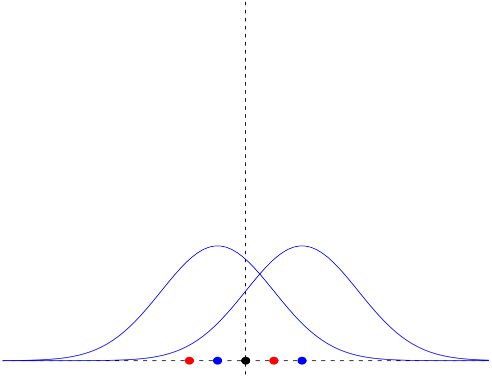
Construction of g

$$f(y) = \sum_{i=1}^{n} \frac{1}{2nh} \left\{ K\left(\frac{y - x_i^\star}{h}\right) + K\left(\frac{y - (-x_i^\star)}{h}\right) \right\}$$



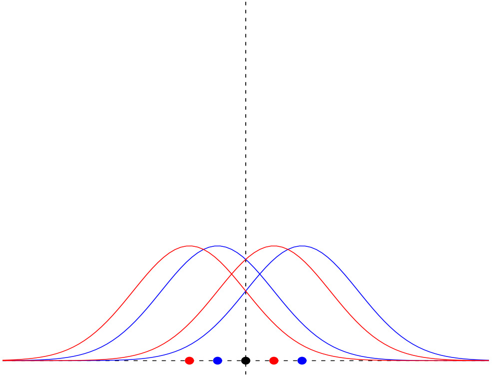
Construction of g

$$f(y) = \sum_{i=1}^{n} \frac{1}{2nh} \left\{ K\left(\frac{y - x_i^{\star}}{h}\right) + K\left(\frac{y - (-x_i^{\star})}{h}\right) \right\}$$



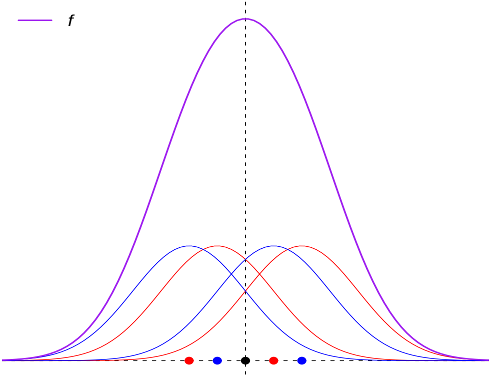
Construction of g

$$f(y) = \sum_{i=1}^{n} \frac{1}{2nh} \left\{ K\left(\frac{y - x_i^\star}{h}\right) + K\left(\frac{y - (-x_i^\star)}{h}\right) \right\}$$



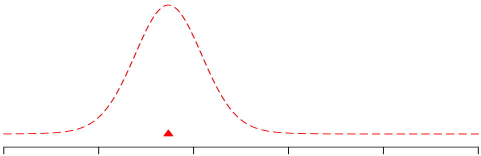
Construction of g

$$f(y) = \sum_{i=1}^{n} \frac{1}{2nh} \left\{ K\left(\frac{y - x_i^\star}{h}\right) + K\left(\frac{y - (-x_i^\star)}{h}\right) \right\}$$



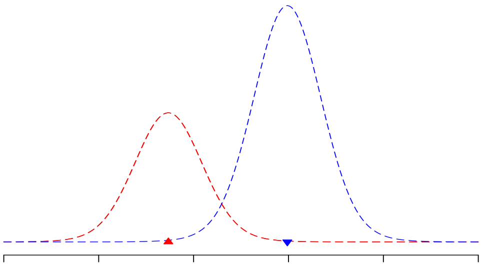
Construction of g

$$g(x) = \lambda f(x - \mu_1) + (1 - \lambda) f(x - \mu_2)$$



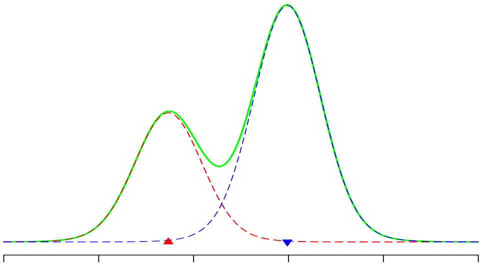
Construction of g

$$g(x) = \lambda f(x - \mu_1) + (1 - \lambda) f(x - \mu_2)$$

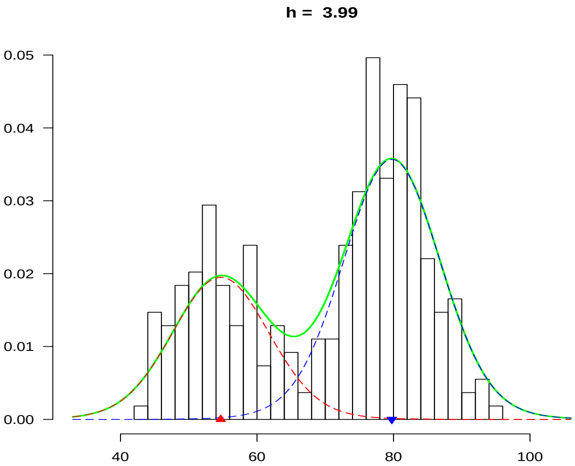


Construction of g

$$g(x) = \lambda f(x - \mu_1) + (1 - \lambda) f(x - \mu_2)$$



Old Faithful Geyser Data



Old Faithful Geyser Data

$\frac{\widehat{\lambda}}{2nh}\times$

Left component

$$\Bigg\{\Bigg\{\widehat{p}_{11}K\left(\frac{x-\widehat{\mu}_1-x_1+\widehat{\mu}_1}{h}\right)+\widehat{p}_{11}K\left(\frac{x-\widehat{\mu}_1+x_1-\widehat{\mu}_1}{h}\right)+\widehat{p}_{12}K\left(\frac{x-\widehat{\mu}_1-x_1+\widehat{\mu}_2}{h}\right)+\widehat{p}_{12}K\left(\frac{x-\widehat{\mu}_1+x_1-\widehat{\mu}_2}{h}\right)\Bigg\}$$

$$+\widehat{p}_{21}K\left(\frac{x-\widehat{\mu}_1-x_2+\widehat{\mu}_1}{h}\right)+\widehat{p}_{21}K\left(\frac{x-\widehat{\mu}_1+x_2-\widehat{\mu}_1}{h}\right)+\widehat{p}_{22}K\left(\frac{x-\widehat{\mu}_1-x_2+\widehat{\mu}_2}{h}\right)+\widehat{p}_{22}K\left(\frac{x-\widehat{\mu}_1+x_2-\widehat{\mu}_2}{h}\right)\Bigg\}$$

$$+\cdots$$

$$+\widehat{p}_{n1}K\left(\frac{x-\widehat{\mu}_1-x_n+\widehat{\mu}_1}{h}\right)+\widehat{p}_{n1}K\left(\frac{x-\widehat{\mu}_1+x_n-\widehat{\mu}_1}{h}\right)+\widehat{p}_{n2}K\left(\frac{x-\widehat{\mu}_1-x_n+\widehat{\mu}_2}{h}\right)+\widehat{p}_{n2}K\left(\frac{x-\widehat{\mu}_1+x_n-\widehat{\mu}_2}{h}\right)\Bigg\}\Bigg\}$$

$\frac{(1-\widehat{\lambda})}{2nh}\times$

Right component

$$\Bigg\{\Bigg\{\widehat{p}_{11}K\left(\frac{x-\widehat{\mu}_2-x_1+\widehat{\mu}_1}{h}\right)+\widehat{p}_{11}K\left(\frac{x-\widehat{\mu}_2+x_1-\widehat{\mu}_1}{h}\right)+\widehat{p}_{12}K\left(\frac{x-\widehat{\mu}_2-x_1+\widehat{\mu}_2}{h}\right)+\widehat{p}_{12}K\left(\frac{x-\widehat{\mu}_2+x_1-\widehat{\mu}_2}{h}\right)\Bigg\}$$

$$+\widehat{p}_{21}K\left(\frac{x-\widehat{\mu}_2-x_2+\widehat{\mu}_1}{h}\right)+\widehat{p}_{21}K\left(\frac{x-\widehat{\mu}_2+x_2-\widehat{\mu}_1}{h}\right)+\widehat{p}_{22}K\left(\frac{x-\widehat{\mu}_2-x_2+\widehat{\mu}_2}{h}\right)+\widehat{p}_{22}K\left(\frac{x-\widehat{\mu}_2+x_2-\widehat{\mu}_2}{h}\right)\Bigg\}$$

$$+\cdots$$

$$+\widehat{p}_{n1}K\left(\frac{x-\widehat{\mu}_2-x_n+\widehat{\mu}_1}{h}\right)+\widehat{p}_{n1}K\left(\frac{x-\widehat{\mu}_2+x_n-\widehat{\mu}_1}{h}\right)+\widehat{p}_{n2}K\left(\frac{x-\widehat{\mu}_2-x_n+\widehat{\mu}_2}{h}\right)+\widehat{p}_{n2}K\left(\frac{x-\widehat{\mu}_2+x_n-\widehat{\mu}_2}{h}\right)\Bigg\}\Bigg\}$$

The Mixture-based Semiparametric Model

Our proposal

$$g(x) = \lambda f(x - \mu_1) + (1 - \lambda) f(x - \mu_2)$$

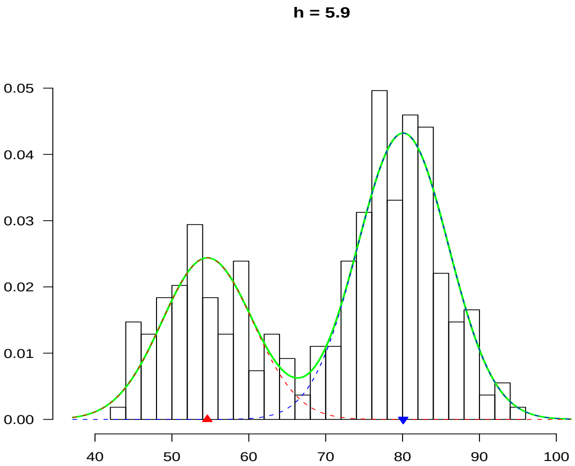
with f being the symmetrized nonparametric mixture given by

$$f(y; \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{j=1}^{m} \frac{\pi_j}{2h} \left\{ \phi\left(\frac{y - \theta_j}{h}\right) + \phi\left(\frac{y + \theta_j}{h}\right) \right\},$$

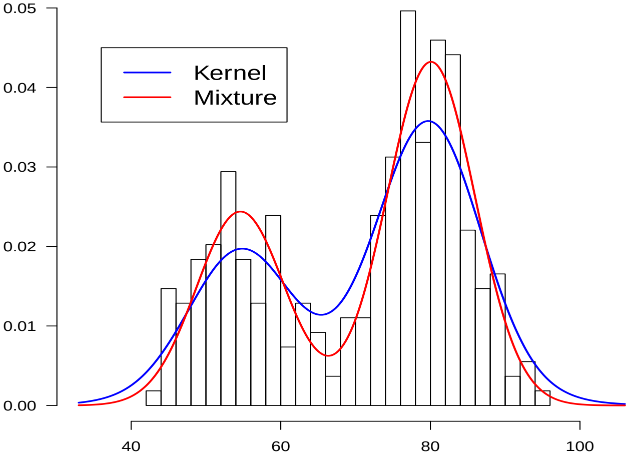
where

- $\phi(\cdot)$ — a known unimodal density that is symmetric about zero,
- $h > 0$ — the (known or fixed) tuning parameter,
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top$ — a support point vector,
- $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)^\top$ — the corresponding probability mass vector.

Old Faithful Geyser Data



A Comparison



A Comparison

Kernel (left component)

$$\frac{\widehat{\lambda}}{2nh}\times$$

$$\Bigg\{\Bigg\{\widehat{p}_{11}K\left(\frac{x-\widehat{\mu}_1-x_1+\widehat{\mu}_1}{h}\right)+\widehat{p}_{11}K\left(\frac{x-\widehat{\mu}_1+x_1-\widehat{\mu}_1}{h}\right)+\widehat{p}_{12}K\left(\frac{x-\widehat{\mu}_1-x_1+\widehat{\mu}_2}{h}\right)+\widehat{p}_{12}K\left(\frac{x-\widehat{\mu}_1+x_1-\widehat{\mu}_2}{h}\right)\Bigg\}$$

$$+\widehat{p}_{21}K\left(\frac{x-\widehat{\mu}_1-x_2+\widehat{\mu}_1}{h}\right)+\widehat{p}_{21}K\left(\frac{x-\widehat{\mu}_1+x_2-\widehat{\mu}_1}{h}\right)+\widehat{p}_{22}K\left(\frac{x-\widehat{\mu}_1-x_2+\widehat{\mu}_2}{h}\right)+\widehat{p}_{22}K\left(\frac{x-\widehat{\mu}_1+x_2-\widehat{\mu}_2}{h}\right)\Bigg\}$$

$$+\cdots$$

$$+\widehat{p}_{n1}K\left(\frac{x-\widehat{\mu}_1-x_n+\widehat{\mu}_1}{h}\right)+\widehat{p}_{n1}K\left(\frac{x-\widehat{\mu}_1+x_n-\widehat{\mu}_1}{h}\right)+\widehat{p}_{n2}K\left(\frac{x-\widehat{\mu}_1-x_n+\widehat{\mu}_2}{h}\right)+\widehat{p}_{n2}K\left(\frac{x-\widehat{\mu}_1+x_n-\widehat{\mu}_2}{h}\right)\Bigg\}\Bigg\}$$

Mixture (left component)

$$\frac{\widehat{\lambda}}{2h}\times\left\{\phi\left(\frac{x-\widehat{\mu}_1-\widehat{\theta}}{h}\right)+\phi\left(\frac{x-\widehat{\mu}_1+\widehat{\theta}}{h}\right)\right\}$$

Part II

Model Parameter Estimation

- Assume that h is KNOWN.
- Employ maximum likelihood estimation of the model parameters.
- Denote by G a discrete distribution formed by the m points of support $\boldsymbol{\theta}$ with corresponding masses $\boldsymbol{\pi}$.
- Let $\boldsymbol{\beta} = (\lambda, \mu_1, \mu_2)^\top$.
- The log-likelihood function: $\ell_h(G, \boldsymbol{\beta}) = \sum_{i=1}^n \log g_h(x_i; G, \boldsymbol{\beta})$
- $\widehat{G}$ and $\widehat{\boldsymbol{\beta}}$ (estimates of G and $\boldsymbol{\beta}$) can be found by the algorithm of Wang (2009).

Model Parameter Estimation

- Assume that h is KNOWN.
- Employ maximum likelihood estimation of the model parameters.
- Denote by G a discrete distribution formed by the m points of support $\boldsymbol{\theta}$ with corresponding masses $\boldsymbol{\pi}$.
- Let $\boldsymbol{\beta} = (\lambda, \mu_1, \mu_2)^\top$.
- The log-likelihood function: $\ell_h(G, \boldsymbol{\beta}) = \sum_{i=1}^{n} \log g_h(x_i; G, \boldsymbol{\beta})$
- $\widehat{G}$ and $\widehat{\boldsymbol{\beta}}$ (estimates of G and $\boldsymbol{\beta}$) can be found by the algorithm of Wang (2009).
- What if h is UNKNOWN?

Tuning Parameter Selection

- Select a "good" h from a set of predetermined candidates.
- General approaches to model selection can be used:
  - cross-validation (CV)
  - information criteria
- General strategy for automatic selection of tuning parameter:
  - *Step 1*: Choose a selection criterion.
  - *Step 2*: Compute the value of the specified selection criterion over a grid of tuning parameters.
  - *Step 3*: Select the h that has the minimum value of the selection criterion.

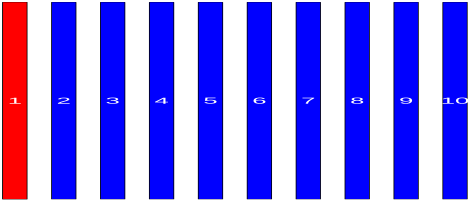
Cross-Validation

- Two CV-based criteria:
  - LSCV — the least-squares cross-validation criterion
  - LCV — the likelihood cross-validation criterion
- One of the commonly used CV methodologies is the V-fold CV.
- We set $V = 10$.

Illustration: V-fold CV

$\{x_i\}_{1 \leq i \leq n}$ is split into V roughly equal-sized and non-overlapping subsets $S_1, \ldots, S_V$.

- Select S_1 as the test set and the remaining folds as the training set.
- Model parameters are estimated based on the training set.
- Fitted model is evaluated on the test set.
- Repeat for $S_2, \ldots, S_V$.



CV-based Criteria

The LSCV and LCV criteria implemented via the V-fold CV procedure are respectively defined by

$$\mathrm{LSCV}(h) = \frac{1}{V}\sum_{v=1}^{V}\int\left\{\widehat{g}_{-v}(x;\widehat{G},\widehat{\beta},h)\right\}^2 dx - \frac{2}{V}\sum_{v=1}^{V}\sum_{x_j\in S_v}\frac{1}{|S_v|}\widehat{g}_{-v}(x_j;\widehat{G},\widehat{\beta},h)$$

and

$$\mathrm{LCV}(h) = -\frac{1}{V}\sum_{v=1}^{V}\sum_{x_j\in S_v}\frac{1}{|S_v|}\log\widehat{g}_{-v}(x_j;\widehat{G},\widehat{\beta},h),$$

where $|S_v|$ denotes the cardinality of S_v and $\widehat{g}_{-v}(x;\widehat{G},\widehat{\beta},h)$ is the fitted model based on all the data points except the observations belonging to the subset S_v.

Information Criteria

- Two popular information criteria:

$$\begin{aligned} \mathrm{AIC}(h) &= -2\ell_h(\widehat{G}, \widehat{\beta}) + 2p \\ \mathrm{BIC}(h) &= -2\ell_h(\widehat{G}, \widehat{\beta}) + p\log(n) \end{aligned}$$

  where p is the number of free parameters.

- We can also use a small sample version of AIC, called $\mathrm{AIC_c}$ (see Burnham and Anderson, 2002):

$$\mathrm{AIC_c}(h) = \mathrm{AIC}(h) + \frac{2p(p+1)}{n-p-1}$$

Part III

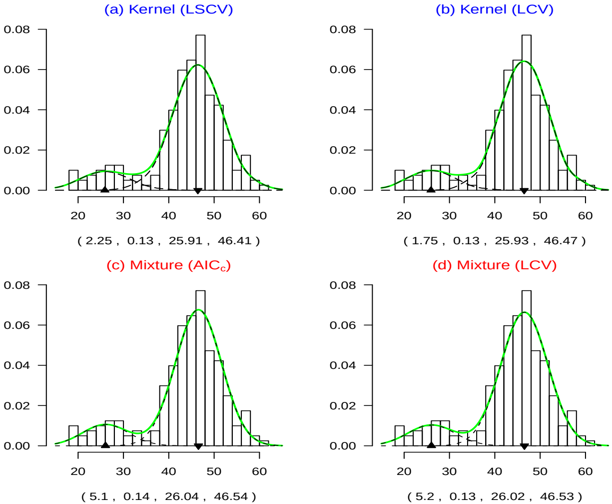
Real Data Examples

- Compare our mixture-based semiparametric model against the kernel-based semiparametric model.
- Used the algorithm of Benaglia *et al.* (2009) for fitting the kernel-based model.
- $K(\cdot)$ and $\phi(\cdot)$ were taken to be the standard Gaussian density.
  - *Example 1*: 2008 World Fly Fishing Championships Data
  - *Example 2*: Exploring Relationships in Body Dimensions
  - *Example 3*: Australian Athletes Data

Example 1: 2008 World Fly Fishing Championships Data

- The 2008 WFFC was held in the Taupo-Rotorua regions; details may be obtained at Yee (2009).
- Considered the length of fish caught in Lake Rotoaira ($n = 201$).

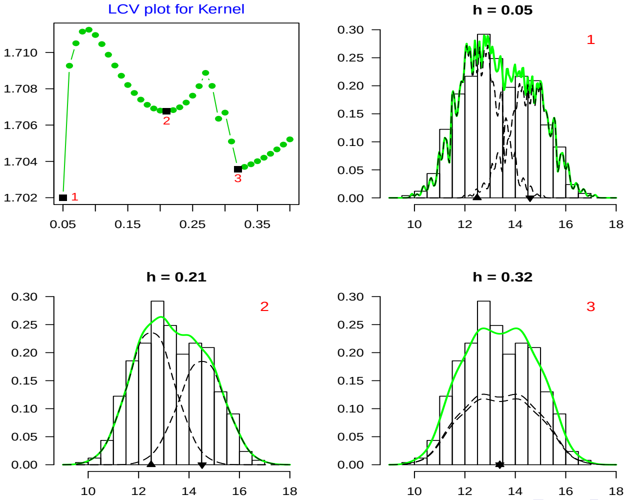
Result 1: Fish Length Data



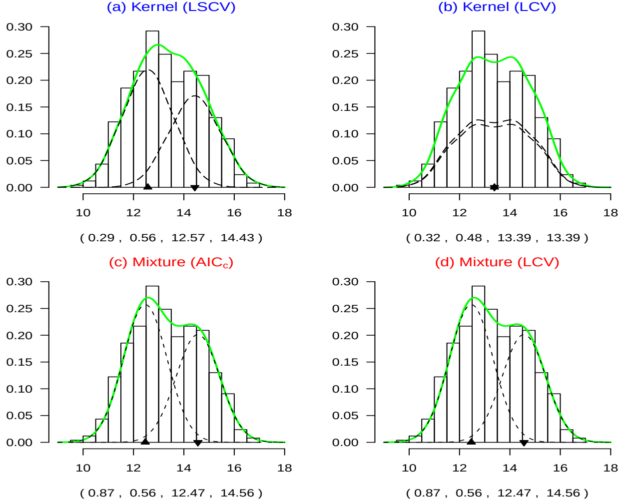
Example 2: Exploring Relationships in Body Dimensions

- 21 body dimension measurements as well as age, weight, height and gender on 247 men and 260 women (Heinz *et al.*, 2003).
- Used the variable elbow diameter.

Result 2: Elbow Diameter Data



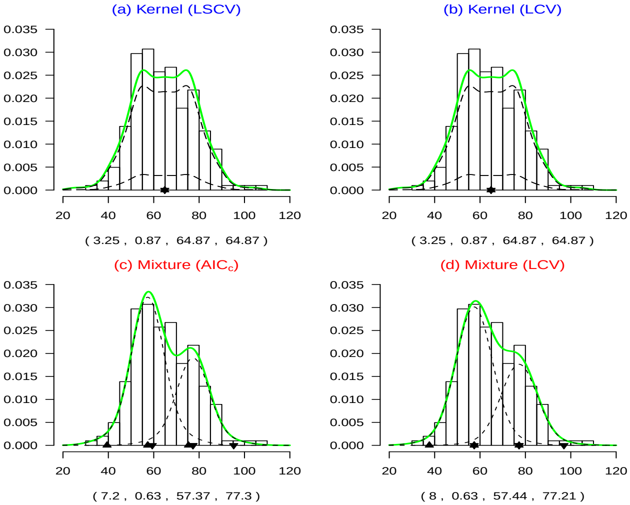
Result 2: Elbow Diameter Data



Example 3: Australian Athletes Data

- Data on 102 male and 100 female Australian athletes collected at the Australian Institute of Sport (Cook and Weisberg, 1994).
- Used the variable lean body mass (LBM).

Result 3: LBM Data

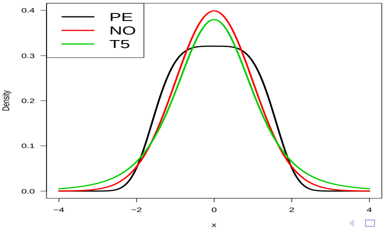


Part IV

Data

- Consider three two-component mixtures (NOmix, T5mix, PEmix) of these components:
  - standard normal (NO)
  - t-distribution with 5 dof (T5)
  - standard power exponential with shape parameter $\nu = 4$ (PE)

  with $\lambda = 0.3$, $\mu_1 = -2$, $\mu_2 = 2$ and $\sigma = 1$.
- In each simulation example, 100 replications with $n = 200$ were generated.



Performance Measures

- To select h:

LSCV and LCV

- To assess parameter estimates:

$$\mathrm{SE}(\alpha, \widehat{\alpha}) = (\widehat{\alpha} - \alpha)^2$$

- To assess density estimates:

$$\begin{aligned} \mathrm{ISE}(g, \widehat{g}) &= \int [\widehat{g}(x) - g(x)]^2 \, dx \\ \mathrm{KLL}(g, \widehat{g}) &= \int g(x) \log \frac{g(x)}{\widehat{g}(x)} \, dx \end{aligned}$$

Result (a): Parameter Estimates

Model	Selection Method	PEmix	NOmix	T5mix
(i)	$\widehat{\text{MSE}}$ ($\times 10^{-3}$) of λ estimates			
Kernel	LSCV	0.96 (0.15)	1.26 (0.17)	1.58 (0.23)
Kernel	LCV	0.95 (0.15)	1.27 (0.17)	1.65 (0.23)
Mixture	LSCV	0.92 (0.14)	1.23 (0.17)	1.43 (0.21)
Mixture	LCV	0.92 (0.14)	1.22 (0.17)	1.45 (0.21)
(ii)	$\widehat{\text{MSE}}$ ($\times 10^{-2}$) of μ_1 estimates			
Kernel	LSCV	2.55 (0.28)	2.32 (0.36)	5.02 (0.61)
Kernel	LCV	2.50 (0.28)	2.30 (0.36)	4.82 (0.59)
Mixture	LSCV	2.30 (0.29)	2.31 (0.35)	3.97 (0.53)
Mixture	LCV	2.10 (0.27)	2.33 (0.34)	4.00 (0.56)
(iii)	$\widehat{\text{MSE}}$ ($\times 10^{-2}$) of μ_2 estimates			
Kernel	LSCV	1.11 (0.16)	0.87 (0.12)	1.30 (0.20)
Kernel	LCV	1.03 (0.15)	0.88 (0.12)	1.62 (0.28)
Mixture	LSCV	1.03 (0.18)	0.87 (0.11)	1.13 (0.18)
Mixture	LCV	0.86 (0.13)	0.87 (0.12)	1.06 (0.17)

NOTE: The values in parentheses are the corresponding standard errors.

Result (b): Density Estimates

Model	Selection Method	PEmix	NOmix	T5mix
(i)	$\widehat{\text{MISE}}$ $(\times 10^{-3})$			
Kernel	LSCV	3.08 (0.18)	3.10 (0.20)	3.87 (0.24)
Kernel	LCV	3.22 (0.19)	2.87 (0.18)	4.04 (0.27)
Mixture	LSCV	2.96 (0.20)	2.56 (0.18)	3.27 (0.22)
Mixture	LCV	2.73 (0.18)	2.24 (0.16)	3.11 (0.21)
(ii)	$\widehat{\text{EKLL}}$ $(\times 10^{-2})$			
Kernel	LSCV	2.13 (0.09)	1.84 (0.13)	5.96 (0.73)
Kernel	LCV	1.91 (0.09)	1.70 (0.11)	5.03 (0.73)
Mixture	LSCV	1.88 (0.11)	1.40 (0.10)	3.96 (0.38)
Mixture	LCV	1.56 (0.09)	1.28 (0.10)	3.49 (0.32)

NOTE: The values in parentheses are the corresponding standard errors.

Part V

Summary

- A methodology for a two-component mixture model with symmetrized nonparametric components is proposed.
- Simulation results and real examples show that the mixture-based methods are more appealing than the kernel-based methods.
- With advantages such as:
  - greater flexibility
  - simpler final model
  - better performance

  the mixture-based methods are competitive for practical applications.

Acknowledgements

- MOHE and UMT – sponsoring my study
- Dept. of Stats., The UoA – covering conference-related costs

References

Cook, R. D. and Weisberg, S.
An Introduction to Regression Graphics
New York: Wiley, 1994.

Burnham, K. P. and Anderson, D. R.
Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Second Edition
New York: Springer-Verlag, 2002.

Heinz, G., Peterson, L. J., Johnson, R. W. and Kerk, C. J.
Exploring Relationships in Body Dimensions
J. Stat. Educ., 11(2):online, 2003.

Bordes, L., Mottelet, S. and Vandekerkhove, P.
Semiparametric Estimation of a Two-Component Mixture Model
Ann. Statist., 34(3):1204–1232, 2006.

Hunter, D. R., Wang, S. and Hettmansperger, T. P.
Inference for Mixtures of Symmetric Distributions
Ann. Statist., 35(1):224–251, 2007.

References

Bordes, L., Chauveau, D. and Vandekerkhove, P.
A stochastic EM algorithm for a semiparametric mixture model
Comput. Statist. Data Anal., 51(11):5429–5443, 2007.

Benaglia, T., Chauveau, D. and Hunter, D. R.
An EM-Like Algorithm for Semi- and Nonparametric Estimation in Multivariate Mixtures
J. Comput. Graph. Stat., 18(2):505–526, 2009.

Wang, Y.
Maximum Likelihood Computation for Fitting Semiparametric Mixture Models
Stat. Comput., 2009, To appear.

Yee, T. W.
VGLMs and VGAMs: an Overview for Applications in Fisheries Research
Fish. Res., 2009, Accepted.