ECON6300/7320/8300 Advanced Microeconometrics Finite Mixtures

Christiern Rose [1]

[1] University of Queensland

Review: Series Estimation

- Series estimation is widely used for nonparametric analysis.
- Let $\phi_1(\cdot), \phi_2(\cdot), \dots$ be a sequence of (basis) functions such that

$$\sum_{j=1}^{\infty} \theta_j \phi_j(\cdot)$$

approximates any function on the same domain with some $\theta_1, \theta_2, \dots$.

- Then, if $\phi_1(\cdot), \phi_2(\cdot), \dots$ are densities, by restricting θ_j to be positive and sum to one, we have a fully flexible density specification with infinite dimensional θ.
- If $\phi_1(\cdot), \phi_2(\cdot), \dots$ are not densities, by normalising

$$\exp\left(\sum_{j=1}^{\infty} \theta_j \phi_j(\cdot)\right)$$

we may approximate any density.

Series Estimation

- In practice, we cannot have infinitely many components. So, we often use a finite number K of components.
- K determines the smoothness of the estimate like h of KDE.

- If K is too small, the density estimate would not be informative.
  If K is too large, the density estimate would be too noisy (bumpy).

- The optimal number of components K should increase as the sample size increases.

Series Estimation

- How to choose K?
- Frequentist:
  - often use BIC or AIC, or some data driven method such as cross-validation, or even use prior information informally.
  - There is no universally accepted rule for choosing K.
  - In any case, computation of parameter estimates is for each given fixed K, but inference is nonparametric (slow-convergence).
- Bayesian:
  - regard K as a latent variable (parameter), put a prior on K with a full support $\mathbb{N}$, and obtain the posterior of K as well as other parameters using an MCMC method.
  - Do not choose an arbitrary K as its distribution is determined.

Series Estimation

- There are many functions that can be used as basis functions $\{\phi_j(\cdot)\}$
- Examples include

  - Polynomials: Legendre polynomials, Bernstein polynomials, etc.
  - Splines: piecewise linear splines, *B*-splines, etc.
  - Densities: Beta densities (Bernstein polynomials), normal densities, Gamma densities, etc.

Series Estimation: BPD

- For example, the Bernstein polynomial density of order K is given as

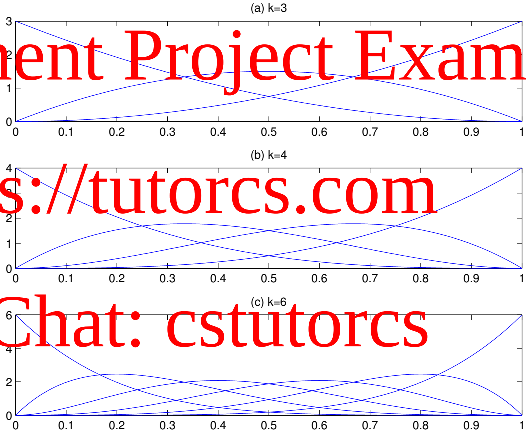
$$f(y|\theta_1, \ldots, \theta_K) := \sum_{j=1}^{K} \theta_j \mathrm{Beta}(j, K - j + 1)$$

where θ_j are all positive and sum to 1 and $\mathrm{Beta}(a, b)$ denotes the Beta density with parameters a and b, i.e., its mean is $ab/(a + b)$.

- When $K \to \infty$, the BPD approximates any absolutely continuous density on $[0, 1]$; see Petrone (1999) for Bayesian nonparametric method using BPD.

Series Estimation: BPD

- The Basis functions are plotted below



- The BPD is a histogram smoothing; see Petrone (1999)

Series Estimation: normal mixture

- Another example: normal densities can be used as basis functions.

$$f(y|\{\pi_j, \mu_j, \sigma_j^2\}) = \sum_{j=1}^{\infty} \pi_j \left\{ \frac{1}{\sigma_j} \phi\left(\frac{y - \mu_j}{\sigma_j} \right) \right\}$$

where $\phi(\cdot)$ is the PDF of $\mathcal{N}(0, 1)$.

- The normal mixture approximates any absolutely continuous density.

Series Estimation: normal mixture

- The normal mixture has been widely used with Dirichlet process prior in nonparametric Bayesian literature; Ferguson (1973), Escobar and West (1995), Walker (2006), etc.

- Note that Dirichlet process can be viewed as a probability distribution over the space of density functions.
- Walker (2006) developed an efficient Gibbs sampler to sample densities from the posterior.

Series Estimation → FMM

- For the rest of the lecture, we consider the normal mixture with a fixed K.
- The mixture model is fairly flexible, but it is a parametric model since K is fixed as a finite number, i.e., K does not increase as N grows.
- When we use a fixed K, the parametric model is called the finite mixture model (FMM).
- Note that frequentiest nonparametric methods use a fixed K for estimation purpose, in which case the nonparametric estimate is numerically the same as the estimate of FMM with the same K.
- However, nonparametric inference is different from parametric inference, e.g., hypothesis testing, confidence intervals, etc.

Series Estimation → FMM

- It is natural to consider FMM as a convenient approximation of a nonparametric model: nonparametric analysis is more complicated.

- However, FMM itself can be a reasonable specification for certain empirical problems.
- Suppose the population can be partitioned into two sub-classes. Within the class, individuals are relatively homogeneous, but between the class, individuals are more heterogeneous.

Series Estimation $\rightarrow$ FMM

- Consider for individual i belongs to class 1 with probability $\pi_1 \in (0, 1)$ and otherwise belongs to class 0 with probability $\pi_0 = 1 - \pi_1$.

- Then, the following finite mixture may be reasonable;

$$f(y|\{\pi_j, \mu_j, \sigma_j^2\}_{j=0}^1) = \pi_0 \left\{ \frac{1}{\sigma_0} \phi \left(\frac{y - \mu_0}{\sigma_0} \right) \right\} + \pi_1 \left\{ \frac{1}{\sigma_1} \phi \left(\frac{y - \mu_1}{\sigma_1} \right) \right\}$$

- It is straightforward to extend to the many class case.

FMM, an example

- Estimating parameters of the distribution of lengths of halibut.

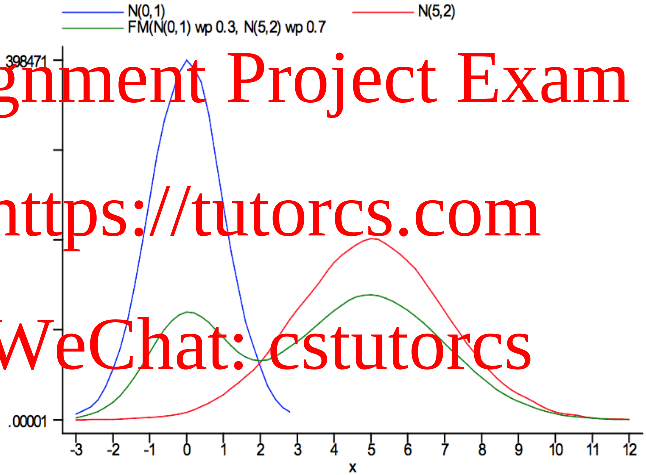
FMM, an example

- Some are small, but some are very large!

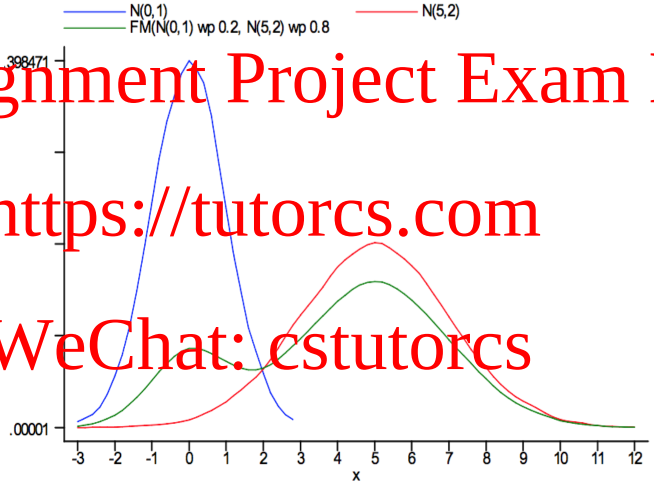
FMM, an example

- It is known that female halibut is longer, on average, than the male and that the distribution of lengths is normal
- Gender cannot be determined at measurement
- Then distribution is a 2-component finite mixture of normals. Latent "types" correspond to gender
- A finite mixture model allows one to estimate:
  - mean/variance of lengths of male and female halibut
  - mixing probability (proportions)
- Other examples: Stock returns in "typical" and "crisis" regimes, GDP growth, Insurance with "risk loving" and "risk averse"

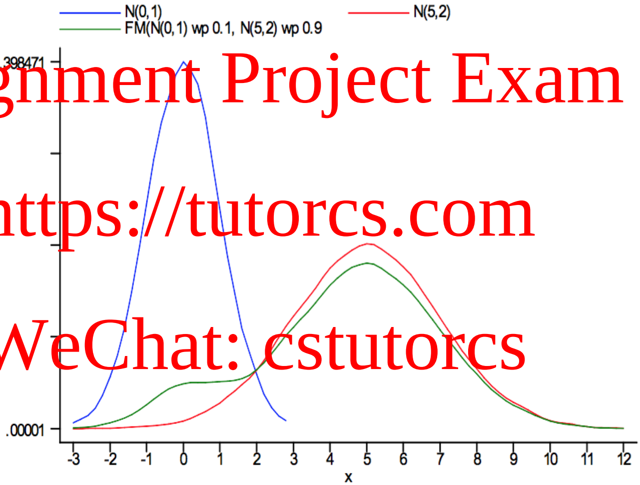
Normal mixture with two components



Normal mixture with two components



Normal mixture with two components



FMM, covariates

- Suppose we observe $\{(y_i, x_i)\}_{i=1}^N$ and there are two latent classes, $\{0, 1\}$ and define the membership indicator $d_i := 1(i \text{ belongs class } 1)$.

- Moreover, we assume

$$y_i | x_i, d_i \sim \mathcal{N}(x_i \beta_{d_i}, \sigma_{d_i}^2)$$

- If we observed d_i, the likelihood would be

$$\prod_{i=1}^{N} \pi_{d_i} \left\{ \frac{1}{\sigma_{d_i}} \phi\left(\frac{y_i - x_i' \beta_{d_i}}{\sigma_{d_i}} \right) \right\}$$

  where $\pi_j = \Pr(d_i = j)$ for $j \in \{0, 1\}$.

- We will briefly discuss about how to obtain the estimates.

FMM, generalisation

- For each class, more generally, we could assume

$$y_i|x_i, d_i \sim f(y_i|x_i, \theta_{d_i})$$

where $f(y_i|x_i, \theta_j)$ is a parametric density function with parameter θ_j, $j \in \{0, 1\}$.

- Then, if we observed d_i, the likelihood would be

$$\prod_{i=1}^{N} \pi_{d_i} f(y_i|x_i, \theta_{d_i})$$

- A problem is that $d := (d_1, \ldots, d_N)$ is not observed.

FMM, estimation

- Frequentist:
  - EM algorithm can be used, which iterates between
    1. Computing the expectation of the log-likelihood as a function of the parameter values from the previous iteration:

$$\mathbb{E}_{d|data,\theta^{(s-1)}}[\ln L(\theta)]$$

    2. Maximizing with respect to θ to obtain $\theta^{(s)}$.

  - For FMM, it is likely that the likelihood has multiple local maxima. So, many initial values have to be considered to be sure of a global maximum.
  - We will see how to estimate a FMM using Stata.

- Bayesian:
  - We obtain the posterior of $(d, \pi_1, \theta_0, \theta_1)$ using an MCMC method.
  - In the Bayesian framework, there is no distinction between missing data d and parameters $(\pi_1, \theta_0, \theta_1)$.
  - Especially, the method to handle the missing data in an MCMC method is called the Data Augmentation.

FMM, some properties

- After integrating d_i out, the model can be written as

$$y_i|x_i,\pi_1,\theta_0,\theta_1 \sim \pi_1 f(y_i|x_i,\theta_1) + (1-\pi_1) f(y_i|x_i,\theta_0)$$

- Hence, the regression function is given as

$$E[y_i|x_i,\pi_1,\theta_0,\theta_1] = \pi_1 E[y_i|x_i,\theta_1] + (1-\pi_1) E[y_i|x_i,\theta_0]$$

- Moreover, the marginal effect is

$$\frac{\partial}{\partial x_i} E[y_i|x_i,\pi_1,\theta_0,\theta_1] = \pi_1 \frac{\partial}{\partial x_i} E[y_i|x_i,\theta_1] + (1-\pi_1) \frac{\partial}{\partial x_i} E[y_i|x_i,\theta_0]$$

- Extension to many classes is straightforward.