Nonparametric approaches to describing heterogeneity

Mogens Fosgerau

Layout

- Motivation
- Kernels and regressions
- Series
- Summary

Setup

- Binomial and multinomial discrete choice models that contain a random preference parameter with an unknown distribution
- Unknown distribution is nonparametric. Combination with parametric model would then be called semiparametric
- In a discrete choice model, the random preference parameter may enter indirect utilities
- Model:

$$y \in \{1,...,J\}, P(y = j \mid x, \beta), \beta \sim F$$

Health warning: RE vs FE

- We shall maintain a random effect assumption: x and β are independent
- This is very convenient, but not always credible and by no means innocuous
- If, for example, the population is divided into men and women, distinguished by x=1 or x=2, then we have to be able to believe that

$$F(\beta) = F(\beta | x = 1) = F(\beta | x = 2)$$

- Sometimes possible to use a fixed effect assumption, under which some parameters can be random but not necessarily independent of the variables
- Fixed effects models are discussed in most econometrics textbooks but not here
- But they are used FAR TOO LITTLE in choice modelling circles

RE is useful!

- If RE assumption is accepted, then

$$P(y|x) = \int P(y|x,\beta) F(d\beta)$$

- If F is known, this integration can generally be carried out, either analytically or numerically
- This is routinely done in the many applications of the mixed logit model, where random parameters are given some distribution and the integration is carried out using simulation

But which distribution to use?

- Mostly we have very little idea what F should be
  - Possibly bounds, sign restriction
- Sometimes the precise form for F is not essential and then it may be unproblematic to impose a specific form
- But it is not desirable to impose a specific functional form on F when
  - The shape of F has significant impact on the object of interest for the investigation
  - When F itself is the object of interest
- Many applications of discrete choice models aim to estimate a WTP distribution
  - It is then highly desirable to be able to infer the functional form for the distribution of WTP from data

Layout

- Motivation
- Kernels and regressions
- Series
- Summary

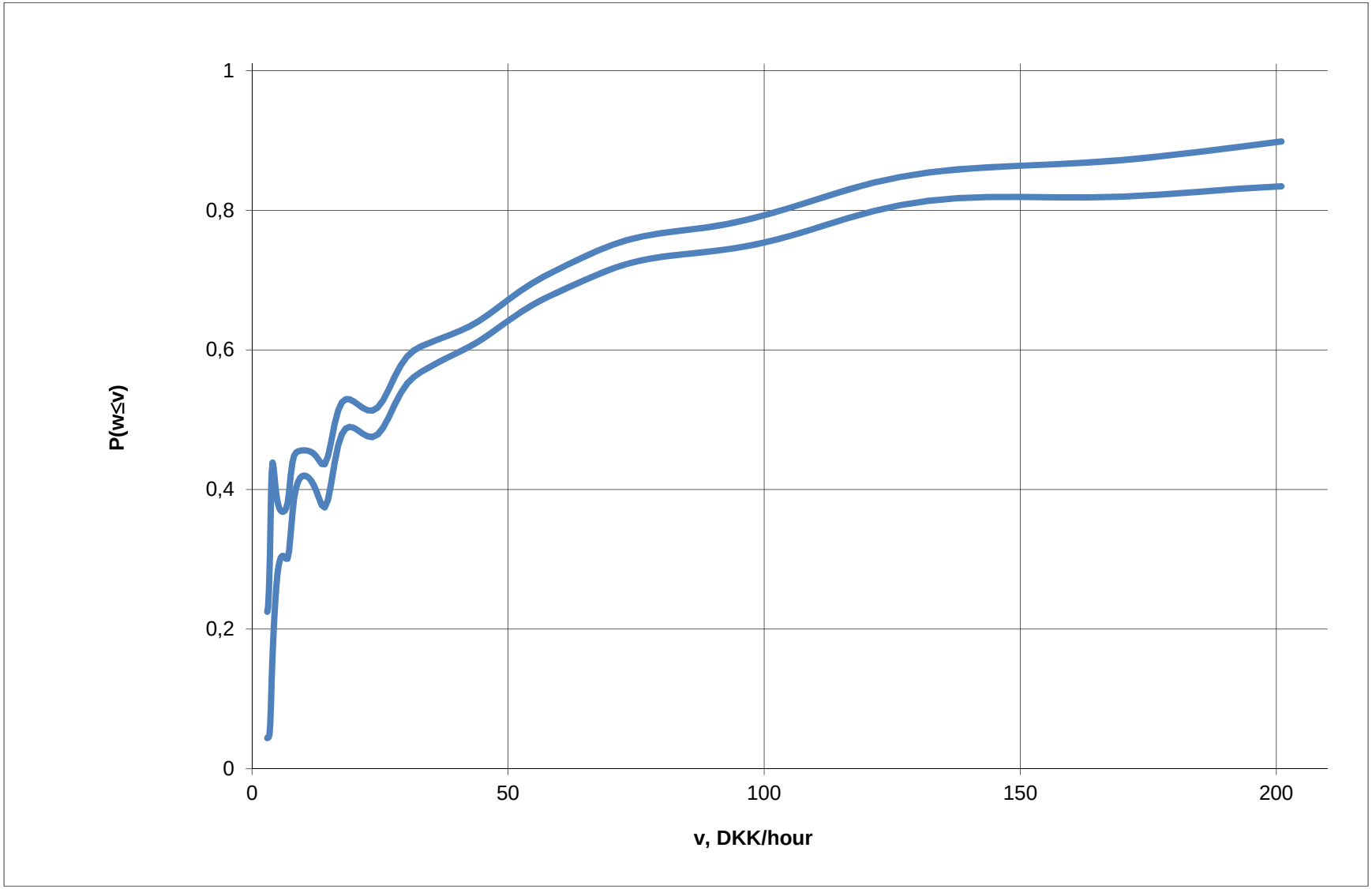
Regression based approaches – Binary choice and no covariates

- Observe $y=1\{ w<v\}$, observe (y,v) for a range of values of v
- Concerned with finding the CDF F of w.
  - Contingent valuation
  - Two factor binary choice
- $E(y|v) = P(w<v) = F(v)$
  - Mean y conditional on a value of v is an estimate of F at the point v
  - Might estimate F(v) as average y_i for (y_i,v_i) near v
- To estimate F we thus need to observe (y,v) many times for a range of values of v
- NEED TO OBSERVE AT ALL RELEVANT VALUES OF v!
- This problem is deadly serious when it is desired to estimate Ew
- Example
  - Know F for values of v up to 100, F(100)=0.9
  - The lower bound for the mean is reached if the residual mass is concentrated at 100
  - The upper bound for the mean is infinity
- Must verify that it is in fact possible to identify the distribution of interest from the data at hand
- Imposing parametric assumptions runs the risk of introducing errors that are extremely large

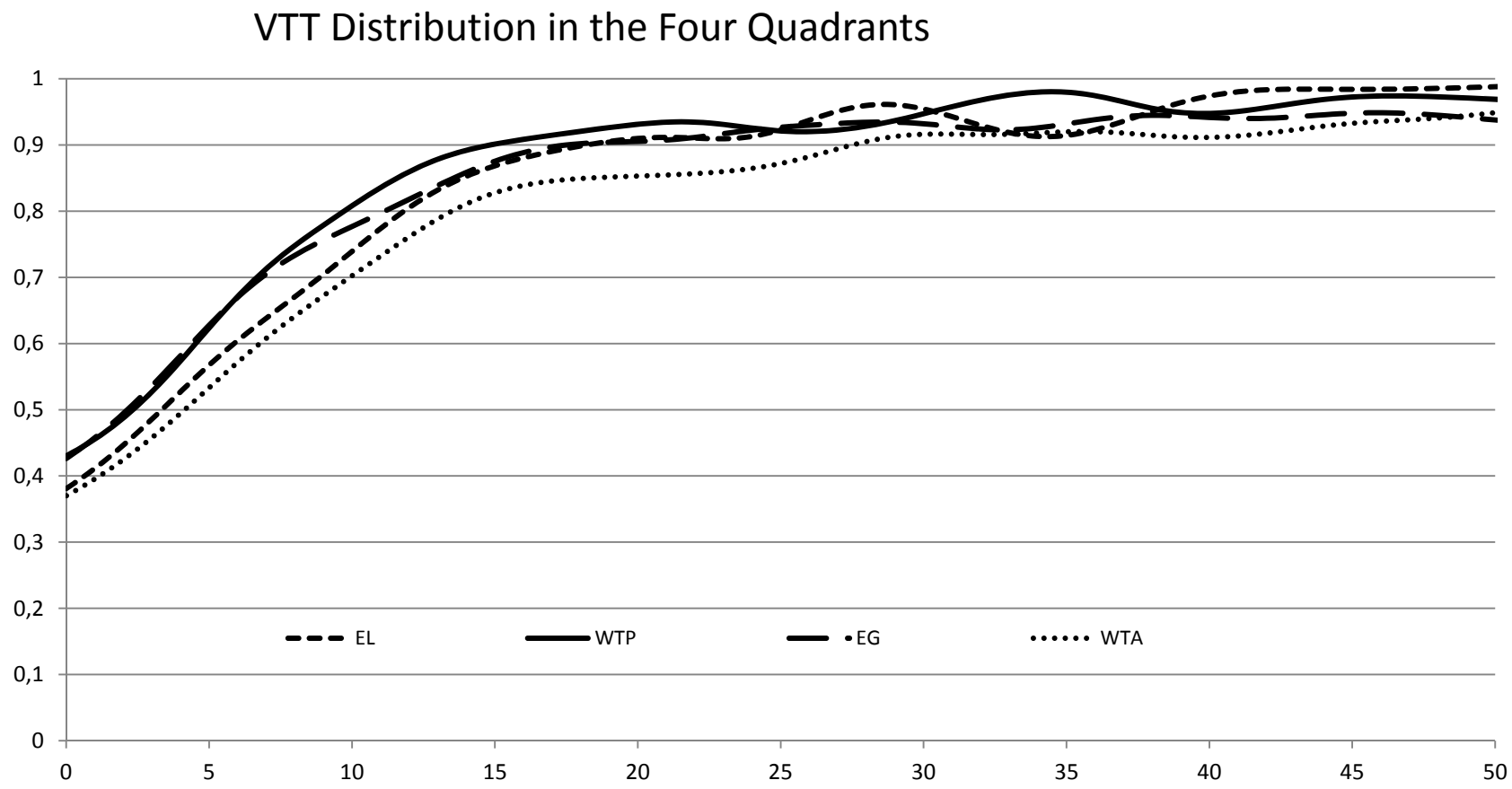
16 parametric distributions

Distribution F_w	(a) Min(supp)	(b) Max(supp)	(e) E(w*1{w>0})
Normal	-∞	∞	51.4
Lognormal	**0**	∞	250.8
Gamma	**0**	∞	70.9
Loggamma	**0**	∞	4.7E+06
Uniform	-190.1	211.0	55.4
Loguniform	**0.8547**	402.6	65.1
Triangular	-56.30	279.3	57.6
Logtriangular	**0.3690**	336.0	59.6
S_B	-14.00	103980	96.0
$LogS_B$	**0.0000**	1810	69.6
S_B1	**2.635**	**201**	54.4
$LogS_B1$	**2.746**	**201**	54.2
Beta	**0.3651**	71743	71.4
Logbeta	**0.0472**	380.3	61.4
Beta1	**2.978**	**201**	53.4
Logbeta1	**2.6021**	**201**	54.0

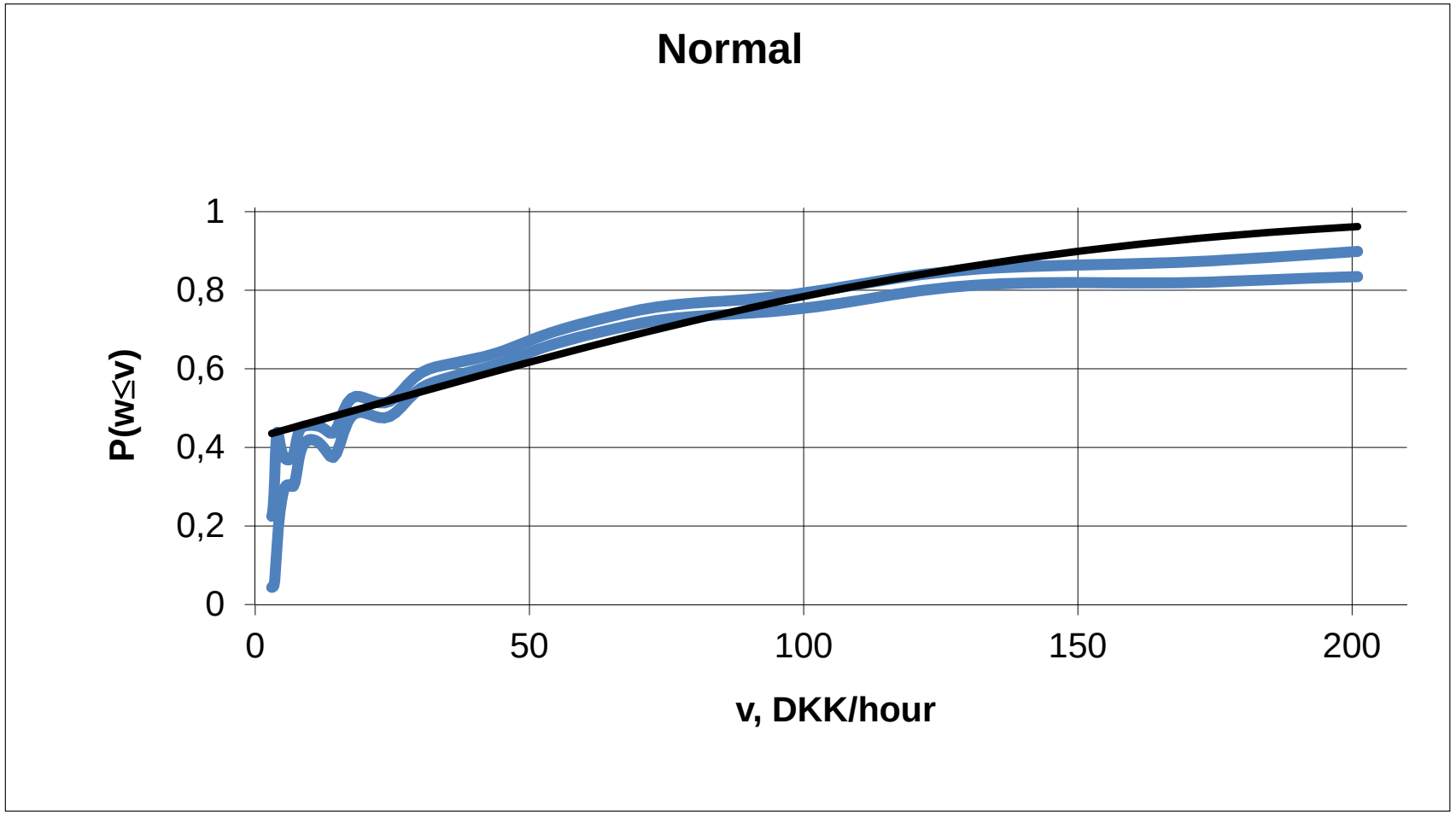
Lack of identification



Catching the tail



Misspecification



Questions for John

- What is an efficient design for nonparametric identification of WTP distribution?
- Will use of efficient designs (for less general models) help with nonparametric identification of WTP distribution?

Kernel regression

- Kernel that places a smooth bump of mass at the point x_0; the concentration of the mass is determined by the bandwidth parameter h $\frac{1}{h}\phi\left(\frac{x-x0}{h}\right)$

$$\widehat{F}(v_0) = \sum_i \overline{w}_i y_i = \frac{\sum_i y_i \phi\left(\frac{v_i - v_0}{h}\right)}{\sum_i \phi\left(\frac{v_i - v_0}{h}\right)}$$

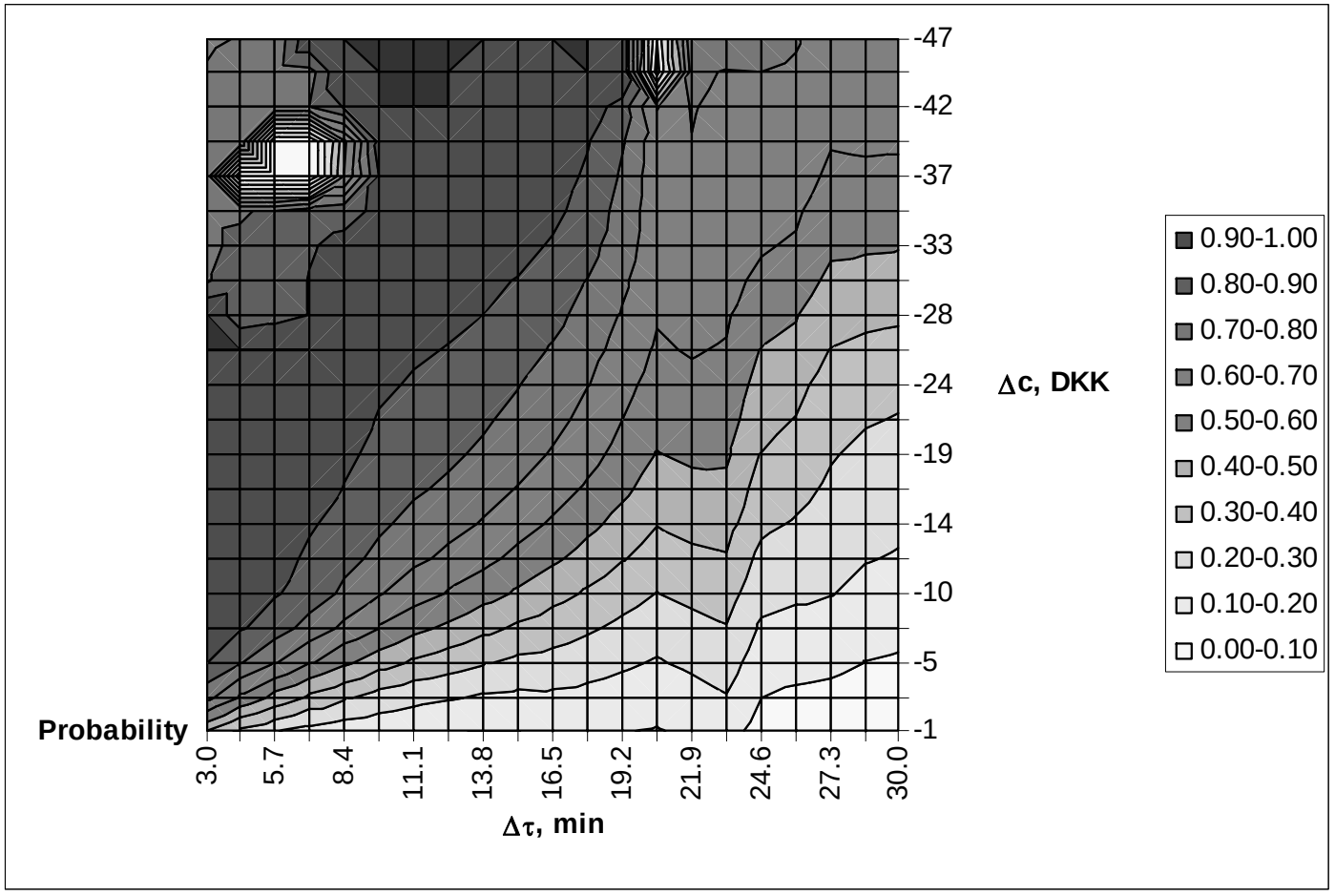
- Choice of kernel is less important
- Choice of bandwidth is very important
  - Mean-variance trade-off
  - Cross-validation
  - Plug-in bandwidth
  - Eye-balling
- Bandwidth will depend on the sample size
  - Optimal bandwidth is smaller for larger samples and in the limit the optimal bandwidth approaches zero

Binary choice including covariates

- Observe $y=1\{w+\beta x<v\}$
  - w independent of x and v
  - This is the same model as before, except now a term βx has been added to the unobserved w
  - x is observed and β must be estimated
  - Such a model arises, e.g., if v is the log of a bid and the willingness-to-pay is $\exp(w+\beta x)$

- If β was known, then we could just regress y against $v-\beta x$ in order to estimate F using kernel regression
- If F was known, then we could estimate β by maximum likelihood, since $P(y=1|v,x) =F(v-\beta x)$
- These observations are the basis for the Klein-Spady estimator.
- Iterate until convergence:
  - Estimate F given β
  - Estimating β given F,
- There are alternatives to Klein-Spady, e.g. Manski 1985, Horowitz 1992, Cosslett 1983
- Lee 1995 generalises to multinomial choice

Multivariate regression: $E(y|x_1,x_2)$



Layout

- Motivation
- Kernels and regressions
- Series
- Summary

Method of sieves

- Construct families of functions that may approximate an unknown function arbitrarily well
- Any (nice) real F may be written as a series in terms of basis functions via
  - $L_k()$ are known basis functions and γ are coefficients

$$F(\beta) = \sum_{k=0}^{\infty} \gamma_k L_k(\beta)$$

- A number of convenient bases exist
- While F has a representation in terms of coefficients, there are, in general, infinitely many coefficients
- F may be approximated by a truncated series

$$F_K(\beta) = \sum_{k=0}^{K} \gamma_k L_k(\beta)$$

- The choice of K determines the degree of flexibility in the approximating F_K
- The optimal K will depend on the shape of F and on the size of the available data set
  - Bias vs variance

- A good choice of leading term L_0 may economize on K

Fosgerau&Bierlaire

- F is now a univariate CDF with density f
- H is another CDF density h
- Use F as a base for estimating the true distribution H
  - F is a candidate for H
  - Require that support of F contains support of H

- Define $Q(u)=H(F^{-1}(u))$, then $Q(F(\beta))=H(\beta)$
- Q is a CDF for a random variable on the unit interval
  - q=Q' is the density
- $h(\beta) = q(F(\beta))f(\beta)$

The key idea

- DC model P(y|x,β) with β~H (true distribution). Then

$$P(y=j\mid x)=\int P(y=j\mid x,\beta)h(\beta)d\beta$$

$$=\int P\left(y=j\mid x,F^{-1}(u)\right)q(u)du$$

- Thus problem of finding unknown h is reduced to that of finding q, an unknown density on the unit interval
- The probability may be simulated using R standard uniform draws u_r and computing

- $$P(y=j\mid x)\approx\frac{1}{R}\sum_r P\left(y=j\mid x,F^{-1}(u_r)\right)q(u_r)$$

Approximating q

- Let L_k be the k'th Legendre polynomial on the unit interval
- Easily computed, orthonormal basis
- Define the density

$$q(u) = \frac{\left(1 + \sum_k \gamma_k L_k(u)\right)^2}{1 + \sum_k \gamma_k^2}$$

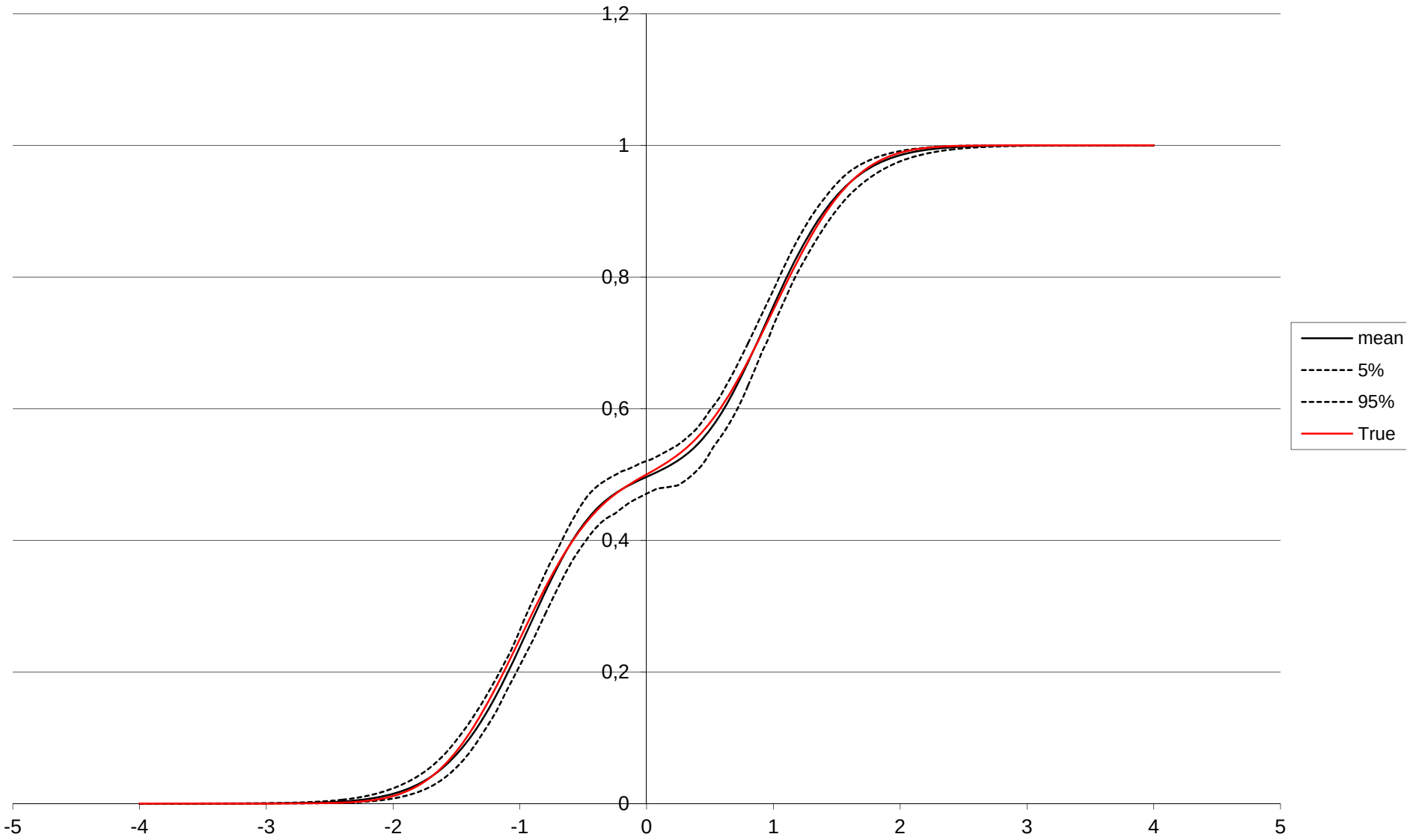
- Any density on the unit interval can be written in this way.
- One way to use this setup is to test the hypothesis that $(\gamma_1,...,\gamma_K)=0$
  - Then q=1 so this amounts to testing whether Q_K is different from the uniform distribution
  - or equivalently whether H=F
- Alternatively, it is possible just to use the flexibility such that the random parameter has distribution $Q_K(F(\beta))$

Example

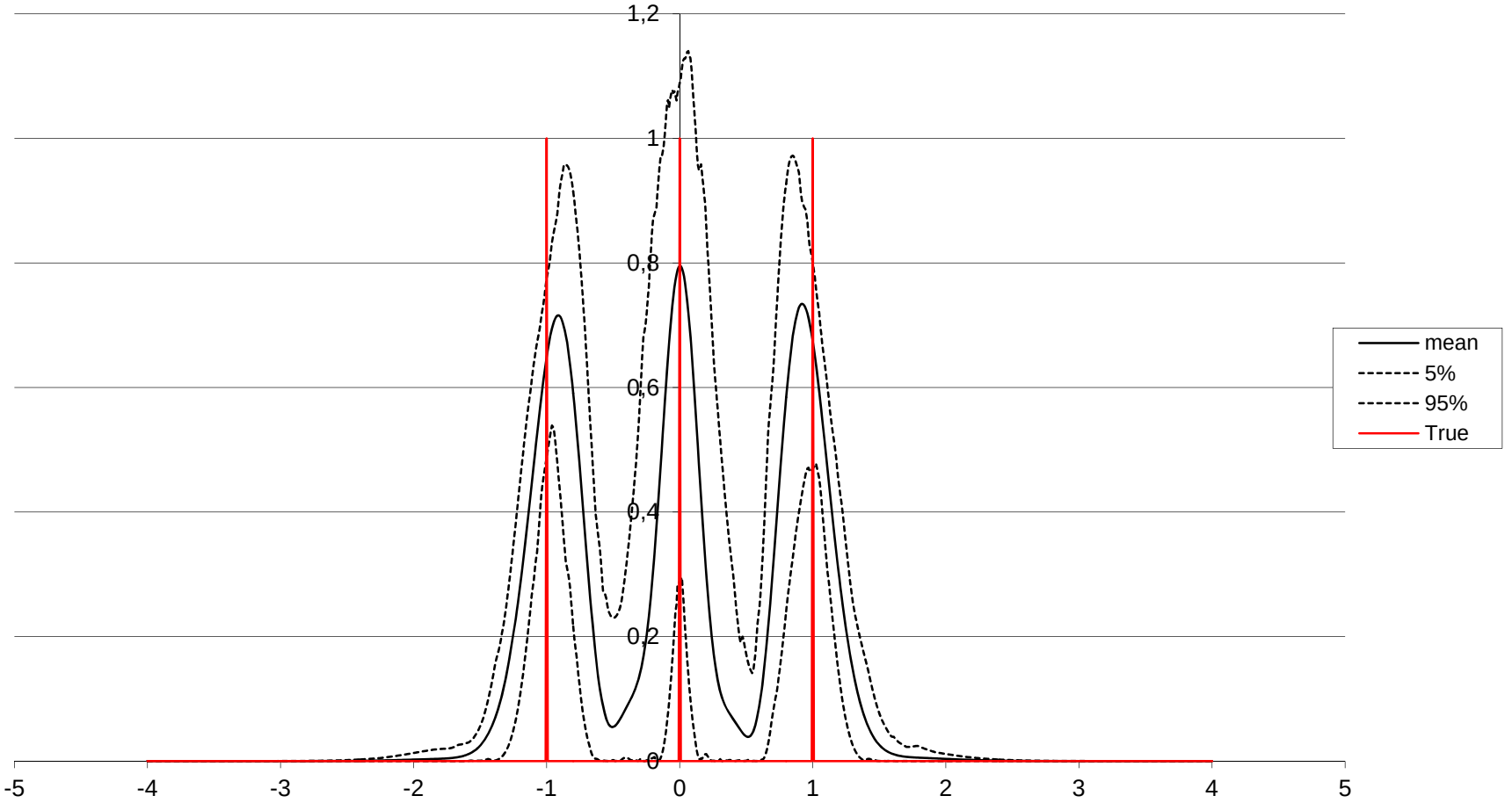
$$P\left(y=j|\mathbf{x},\beta\right)=\frac{\exp\left(\boldsymbol{\alpha}\mathbf{x}_j+\beta x_j^0\right)}{\sum_{j'}\exp\left(\boldsymbol{\alpha}\mathbf{x}_{j'}+\beta x_{j'}^0\right)}$$

$$P(y=j|\mathbf{x})\simeq\frac{1}{R}\sum_r\frac{\exp\left(\boldsymbol{\alpha}\mathbf{x}_j+F^{-1}\left(u_r\right)x_j^0\right)}{\sum_{j'}\exp\left(\boldsymbol{\alpha}\mathbf{x}_{j'}+F^{-1}\left(u_r\right)x_{j'}^0\right)}\frac{\left(1+\sum_{k=1}^K\gamma_kL_k\left(u_r\right)\right)^2}{1+\sum_{k=1}^K\gamma_k^2}$$

FB approximation



FB9 on three mass points



Tricks with series

i. A series
ii. Squared series is always positive
iii. If L is an ONB
iv. Always increasing
v. Convex

$$(i): f(x) = \sum_{k=0}^{K} \alpha_k L_k(x)$$

$$(ii): f^2(x) = \sum_{k,l=0}^{K} \alpha_k \alpha_l L_k(x) L_l(x)$$

$$(iii): \int f^2(s)\,ds = \sum_{k=0}^{K} \alpha_k^2$$

$$(iv): \int^x f^2(s)\,ds$$

$$(v): \int^x \int^t f^2(s)\,ds\,dt$$

Combining sieves with a copula

- Multivariate distributions lead to the curse of dimensionality
- Use a parametric form to describe a dependence structure using a small number of parameters, while allowing marginal distributions to be arbitrary. This is achieved through the use of copula
- Consider a random vector ($X_1,...,X_D$)~F, with continuous marginal distributions F_d
- Write $F(\beta) = C (F_1(\beta),...,F_D(\beta))$
  - Where C is the CDF of the random vector $(F_1(X_1),...,F_D(X_D))$
  - Such a C is called a copula
  - It is a CDF on the unit cube with univariate marginal distributions being uniform
  - Any such CDF is a copula
  - The copula in captures precisely the dependence structure of F and does not depend on the marginal distributions
  - The simplest copula is the independence copula, which is the product $C(u) = u_1 * \ldots * u_D$

Creating copula

- Use known multivariate CDF, e.g. multivariate normal, completely defined in terms of the correlation matrix. In D dimensions, this has D (D-1)/2 parameters
- Archimedian copula have the form $C(u) = \psi(\psi^{-1}(u_1)+...+ \psi^{-1}(u_D))$,
- Any multivariate extreme value distribution with EV1 marginals has the form $\exp(-G(\exp(-\beta_1),...,\exp(-\beta_D)))$
  - G is a choice probability generating function with certain properties
  - Such CPGF may be viewed as generalisations of summation
  - Replacing the sum in the Archimedian copula leads to a generalised Archimedian copula
- Then complex dependence structures may be handled using nesting as in the nested or cross-nested logit models

Copula and simulation

- Copula C has density c:

$$P(y=j|\mathbf{x}) = \int_{\mathbf{u}\in[0,1]^D} P\left(y=j|\mathbf{x}, \left(F_1^{-1}(u_1), \ldots, F_D^{-1}(u_D)\right)\right) c(\mathbf{u})\, d\mathbf{u},$$

Mixtures of distributions – a sum of bumps

- Define pairs (μ_k, σ_k) of means and standard deviations
- Corresponding weights π_k, that are positive and sum to 1
- Approximate unknown CDF as a discrete mixture of smooth distributions (e.g. standard normals) using

$$F(\boldsymbol{\beta}) = \sum_{k=1}^{K} \pi_k \Phi\left(\frac{\boldsymbol{\beta} - \boldsymbol{\mu}_k}{\sigma_k}\right)$$

Layout

- Motivation
- Kernels and regressions
- Series
- Summary

Summary

- Why/when is it important to use nonparametric distributions for random parameters?
- Regression based methods
  - wysiwyg
- Series methods
  - Always applicable
- Things to develop
  - Identification tests
  - Multivariate stuff, copula
- Never forget
  - Identification?
  - Possibility of using FE?
  - Specification testing