Varying Random Coefficient Models

Christoph Breunig
Humboldt-Universität zu Berlin

Workshop on Inverse Problems

Heidelberg
October 28th, 2016

Introduction

Heterogeneity of individuals is a common feature in microeconometric applications.

Therefore, recent interest of econometrics in models that allow for marginal effects to vary across individuals

Causes for individual heterogeneity are not only observed but also unobserved individual characteristics

This paper investigates the impact of both, observed and unobserved characteristics on individual heterogeneity

Models with heterogeneous Parameters

In this talk, consider linear models of the form

$$Y = B_0 + B_1' X$$

where Y denotes random scalar, X covariates, and $B = (B_0, B_1')'$ captures individual heterogeneity.

A common assumption for identification is

$$X \perp B$$

enables identification the distribution of B.

Can be relaxed to conditional independence. But still difficult to justify in some applications!

1st Example: Engel Curves

- Demand for a good Y is modeled through

$$Y = B_0 + B_1 X_1 + B_2' X_{-1}$$

where X_1 denotes log total expenditure and X_{-1} are other household characteristics (abstract from prices).

- If $X_1 \perp B_1$ then B_1 captures effect of a marginal change of X_1
- B_1 might be generated by heterogeneity in preferences
- Assumption of independence of preferences and total expenditure seems implausible
  $\Rightarrow$ Total expenditure rather an important variable to describe household's preferences

2nd Example: Returns to Education

- Card (2001) assumes that log earnings Y are described through education W and experience E as

$$Y = B_0 + B_1 W + B_2 E$$

where

$$B_1 = A_1 - const.W$$

and random parameter A_1.

- Accounts for heterogeneity in economic benefits and marginal costs of schooling
- Linear specification of heterogeneity not justified by economic theory

Varying Random Coefficient Model

We consider linear model

$$Y = B_0 + B_1' X$$

where for $1 \leq l \leq d-1$:

$$B_{1,l} = \underbrace{g_l(W)}_{\textit{varying coefficients}} + \underbrace{A_{1,l}}_{\textit{random coefficients}}$$

The varying random coefficients thus consist of

1. varying coefficients that are explained by covariates $W \subset X$
2. random coefficients A that are independent of X

Aim of this Talk

The aim of this talk:

1. *Identification of the distribution of marginal effects*
   - ▷ Relies on the specific structure how covariates enter parameter of heterogeneity B
   - ▷ Does not require any external variation through instruments
2. *Provide a two step estimation procedure*
   - ▷ First step relies on varying coefficient through the conditional mean restriction
   - ▷ Second step relies on sieve minimum distance approach of characteristic functions

Literature

Random Coefficient Models with $B \perp X$

Beran, Feuerverger & Hall (1993), Hoderlein, Klemelä & Mammen (2010), Gautier & Kitamura (2011) , Dunker, Hoderlein & Kaido (2014), Fox & Lazzati (2014), ...

Instrumental variables in Random Coefficient models

Florens, Heckman, Meghir, and Vytlacil (2008), Kasy (2013), Masten and Torgovitsky (2014), Masten (2015), Hoderlein, Holzmann & Meister (2015)

Sieve Estimation in Discrete Choice Models with Random Coefficients

Fox, Kim & Yang (2015)

Varying Coefficients Literature

Hastie & Tibshirani (1993), Härdle, Hall & Ichimura (1993),..., Fan, Yao, and Cai (2003), Cui & Härdle (2011), Park, Mammen & Lee (2015), Ma & Song (2015)

Table of contents

The Model

We consider linear model

$$Y = B_0 + B_1' X$$

with parameters of heterogeneity $B_1 = (B_{1,1}, \ldots, B_{1,d-1})$ modeled with with $W \subset X$ as

$$B_{1,l} = g_l(W) + A_{1,l} \quad \text{for } 1 \leq l \leq d-1$$

Assumption 1

The covariates X are independent of A.

- Independence not sufficient for identification of model $Y = B_0 + B_1' X$ where $B_1 = \psi(W, A)$ and ψ is not add. separable

Mean effect: Varying Coefficients Model

- Model can be rewritten as

$$Y = A_0 + \sum_{l=1}^{d-1} \big(A_{1,l} + g_l(W)\big) X_l$$

- Condition $A \perp X$ implies varying coefficients model

$$\boldsymbol{E}[Y|X] = g_0(X) + \sum_{l=1}^{d-1} g_l(W) X_l$$

where g_0 linear with $g_0(x) = \boldsymbol{E}A_0 + x'\boldsymbol{E}A_1$ (see e.g. Cai, Fan, Li (2000), ...)

Potential Outcome and marginal Effects

- B depends on X: Interpretation of B as marginal effects fails.
- Introduce the potential outcome notation

$$Y^x := (1, x')B^x = A_0 + \sum_{l=1}^{d-1} \big(g_l(w) + A_{1,l}\big)x_l$$

- Effect of a marginal change in x_k on Y_x:

$$\partial B_k^x := g_k(w) + A_{1,k} + \sum_{l=1}^{d-1} g_l'(w)x_l$$

Notations

- Let $x = (w', v')'$. Introduce regression function

$$g(x) = \boldsymbol{E}[Y|X = x] \quad \left(= g_0(w) + \sum_{l=1}^{d-1} g_l(w) x_l \right)$$

- Vector of marginal effects

$$\partial B^x = \partial \mathbf{g}(x) + A$$

with $\partial \mathbf{g}(x) = (0, \partial g_1(x), \ldots \partial g_{d-1}(x))$ where $\partial g_k(x) = dg(x)/dx_k$

- Conditional characteristic function

$$\varphi(x, t) = \boldsymbol{E}[exp(itY)|X = x] \quad \left(= [\mathcal{F} f_{Y|X=x}](t) \right)$$

Identification of random coefficients A

Theorem

Assume large support assumption of X. Then,

$$f_A(a) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} |t|^{d-1} \exp\big(- it(a_1 + a'_{-1}x + g(x))\big)\varphi(x,t)dt\,dx.$$

- Argument: Fourier inversion yields

$$\begin{aligned}(2\pi)^d f_A(a) &= \int_{\mathbb{R}^d} \exp(-ia'u)[\mathcal{F}f_A](u)du \\ &= \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} |t|^{d-1} \exp\big(- it(a_1 + a'_{-1}x)\big)[\mathcal{F}f_A](t,tx)dt\,dx \\ &= \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} |t|^{d-1} \exp\big(- it(a_1 + a'_{-1}x)\big)[\mathcal{F}f_{Y|X=x}(\cdot + g(x))](t)dt\,dx\end{aligned}$$

Identification of marginal effects ∂B^x

Corollary

The p.d.f. of $\partial B^{x_0} = \partial \mathbf{g}(x_0) + A$ *is identified trough*

$$f_{\partial B^{x_0}}(b) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} |t|^{d-1} \exp[-it(b_1 + b'_{-1}x + \partial h(x_0, x))] \varphi(x, t) dt dx$$

- Here we denoted $\partial h(x_0, x) = (1, x') \, \partial \mathbf{g}(x_0) - g(x)$
- Follows directly from $f_{\partial B^x}(b) = f_A(b - \partial \mathbf{g}(x))$

Relation to Radon transform

- In the linear RC case, i.e, $g_I = 0$, normalize the model to $U = B'S$ where $U = Y/\|X\|$ and $S = X/\|X\|$.
- In this case, it is well known that f_B is identified through the

$$f_{U|S=s} = Rf_B$$

  where R denotes the radon transform.
- Statistical inverse problems involving Radon transform appear in several problems in tomography (Korostelev & Tsybakov (1993), Cavalier (2000, 2001)).
- Radon needlet thresholding based on svd of Radon transform (Kerkyacharian, Le Pennec & Picard (2012))
- In the context of RC: Beran, Feuerverger & Hall (1996) and Feuerverger & Vardi (2000), Hoderlein, Klemelä & Mammen (2010), ...

Estimation Strategy

- Due to nonlinearity of varying coefficients, theory for Radon transform based estimators appears difficult.
- We have seen that $f_{\partial B^{x_0}}$ is identified through

$$\left[\mathcal{F} f_{\partial B^{x_0}}(\cdot + \partial \mathbf{g}(x_0))\right](t, tx) = \exp\left(-itg(x)\right)\varphi(x,t)$$

  for all $t \in \mathbb{R}$ and $x \in \mathbb{R}^{d-1}$.

- Approach is based on sieve minimum distance approach, which has become increasingly popular in the econometric literature (Ai & Chen (2003), Chen & Pouzo (2012), Horowitz (2011), B, Simoni & Mammen (2015),...)

Minimum Distance Estimation

- Approach relies on weighted L^2 distance

$$\int \int_0^\infty \left| \left[\mathcal{F} f_{\partial B^{x_0}} (\cdot + \partial \mathbf{g}(x_0)) \right](t, tx) - \exp\left(- itg(x) \right) \varphi(x, t) \right|^2 \nu(t) dt dx = 0,$$

  given some weighting function ν with $\int_{\mathbb{R}} \nu = 1$.

- Parametric minimum distance estimators based on weighted L^2 was considered by Hoderlein, Holzmann & Meister (2015).

- The estimator relies on two steps:
  1. Estimate ccf. φ and rf. g and its derivative
  2. Estimate $f_{\partial B^{x_0}}$ via an empirical version of the L^2 criterion

1. Step Estimators

- Let $p^K(x) = \big(p_1(x), \ldots, p_K(x)\big)'$ and $p_d^K(x) = p^K(w) \otimes x$ of dimension $K = K(n)$
- Estimate ccf. φ and rf. g by

$$\widehat{\varphi}_n(x,t) = p^K(x)\big(P'P\big)^- \sum_{j=1}^{n} \exp(itY_j)p^K(X_j)$$

and

$$\widehat{g}_n(x) = p_d^K(x)'(P_d'P_d)^- \sum_{j=1}^{n} Y_j p_d^K(X_j).$$

where $P = \big(p^K(X_1), \ldots, p^K(X_n)\big)'$, $P_d = \big(p_d^K(X_1), \ldots, p_d^K(X_n)\big)'$.

- Estimate $\partial g_k(x)$ by $\partial\,\widehat{g}_k(x) = d\widehat{g}_n(x)/dx_k$

2. Step Estimator of pdf. of B^{x_0}

Consider the sieve minimum distance estimator $\widehat{f}_{\partial B^{x_0}}$ of $f_{\partial B^{x_0}}$ given by

$$\arg\min_{f \in \mathcal{B}_n} \int \int_0^\infty \left| \widehat{\varphi}_n(x,t) \exp\left(- i\, t\, \widehat{g}_n(x) \right) - \left[\mathcal{F} f(\cdot + \partial \widehat{\mathbf{g}}_n(x_0)) \right](t, tx) \right|^2 d\nu(t) dx$$

- $\mathcal{B}_n$ is a sieve space of dimension $L = L(n) < \infty$ with basis functions $\{q_l\}_{l \geq 1}$.
- Sieve estimation is also convenient to impose additional constraints on the unknown functions (see also Chetverikov & Wilhelm (2015))

Explicit Estimator

- Without additional constraints obtain explicit solution to minimum distance critertion
- Thus, $\widehat{f}_{\partial B^x}(\cdot) = q^L(\cdot + \partial\widehat{\mathbf{g}}_n(x))'\widehat{\beta}_n$ where

$$\widehat{\beta}_n = \Big(\int_{R^{d-1}} \int_0^\infty [\mathcal{F}q^L](t, tx)[\mathcal{F}q^L](t, tx)' d\nu(t) dx \Big)^- \times \int_{R^{d-1}} \int_0^\infty [\mathcal{F}q^L](t, tx)\widehat{\varphi}_n(x, t) \exp\big(- i\, t\, \widehat{g}_n(x)\big) d\nu(t)\, dx.$$

Assumptions on Basis functions

Assumption 2

(i) $(Y_1, X_1), \ldots, (Y_n, X_n)$ *i.i.d. sample*

(ii) $\sup_x \|p^K(x)\|^2 \log n = o(n\lambda_K)$ *and* $\sup_x \|q^L(x)\|^2 \log n = o(n)$.

Discussion:

- $\sup_x \|p^K(x)\|^2 = O(K)$ for many basis functions like B-splines, wavelets, or trigonometric
- For (*ii*) see Belloni et al. (2015) and Chen & Christensen (2015)
- λ_K denotes the minimal eigenvalue of the associated matrix $\boldsymbol{E}[p^K(X)p^K(X)']$

Assumptions on Basis functions (cont'd)

Assumption 2

(iii) *There exists a strictly positive and nonincreasing sequence* $(\lambda_k)_{k\geq 1}$ *such that*

0.1 $\lambda_{min}\left(\lambda_K^{-1}\boldsymbol{E}[p^K(X)p^K(X)']\right) \gtrsim 1$

0.2 $\lambda_{min}\left(\boldsymbol{E}[XX'|W=w]\right) \gtrsim 1.$

Discussion:

- If $f_X \geq const.$ and p^K is a vector of orthonormal basis functions $\Rightarrow$ 0.1 holds with $\lambda_K \geq const.$
- Eigenvalue restrictions as in 0.2 are commonly imposed in VC literature, see (see Xia & Härdle (2006), Lee, Mammen & Park (2012), Ma & Song (2015))

Sieve Link Condition

Assumption 3

There exists a strictly positive and nonincreasing sequence $(\tau_l)_{l\geq 1}$ *such that*

(i) $\lambda_{min}\Big(\tau_L^{-1}\int_{\boldsymbol{R}^{d-1}}\int_0^\infty [\mathcal{F}q^L](t,tx)[\mathcal{F}q^L](t,tx)'d\nu(t)dx\Big) \gtrsim 1$

(ii) $\|\mathcal{F}(\Pi_L f_{\partial B^x} - f_{\partial B^x})\|_\nu^2 = O\big(\tau_L \int |\Pi_L f_{\partial B^x}(b) - f_{\partial B^x}(b)|^2 db\big)$.

- We denote $\|\phi\|_\nu = \big(\int_0^\infty \int_{\boldsymbol{R}^{d-1}} |\phi(x,t)|^2 dx d\nu(t)\big)^{1/2}$.
- Similar to Chen & Reiss (2011) and Chen & Pouzo (2013) in nonparametric instrumental regression.
- In classical statistical inverse problems involving the Radon transform, the degree of ill-posedness is $(d-1)/2$ (see, e.g., Korostelev and Tsybakov (1993))

Discussion of Link Condition

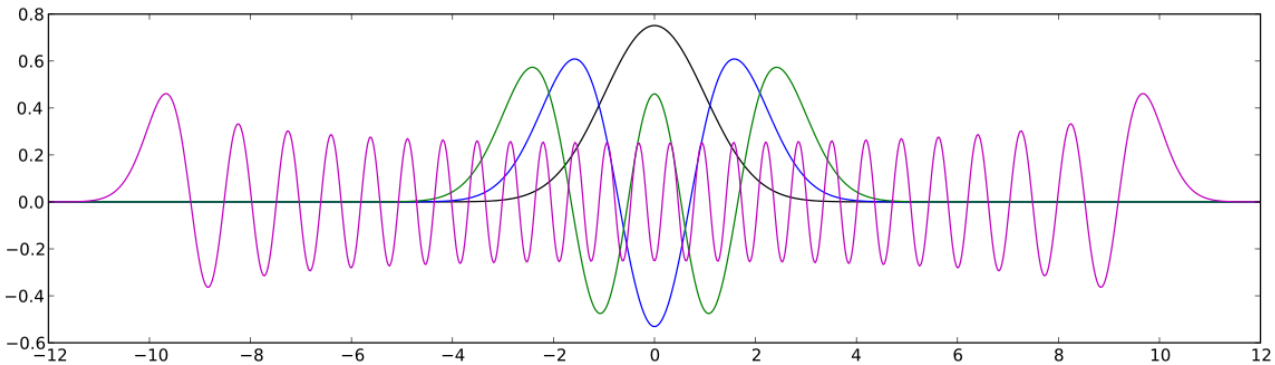
Let $\widetilde{\nu}(t) = t^{d-1}\nu(t)$ where ν is p.d.f. and $\{q_l\}_{l\geq 1}$ Hermite functions.

Proposition

(i) *If for all* $n \geq 1$ *and any* $a \neq 0$*:*

$$\int_0^\infty (a'q^L(t))^2 \, 1\{\widetilde{\nu}(t) < \tau_L\} dt < \int_0^\infty (a'q^L(t))^2 dt.$$

then $\lambda_{min}\left(\tau_L^{-1} \int_{R^{d-1}} \int_0^\infty q^L(t,tx) q^L(t,tx)' d\nu(t) dx\right) \gtrsim 1$



Discussion of Link Condition

Let $\widetilde{\nu}(t) = t^{d-1}\nu(t)$ where ν is p.d.f. and $\{q_l\}_{l\geq 1}$ Hermite functions.

Proposition

(i) *If for all* $n \geq 1$ *and any* $a \neq 0$*:*

$$\int_0^\infty (a' q^L(t))^2 \, 1\{\widetilde{\nu}(t) < \tau_L\} dt < \int_0^\infty (a' q^L(t))^2 dt.$$

then $\lambda_{min}\left(\tau_L^{-1} \int_{R^{d-1}} \int_0^\infty q^L(t, tx) q^L(t, tx)' d\nu(t) dx\right) \gtrsim 1$

(ii) *If*

$$\sum_{k>L} \int_0^\infty \left| q_k(t) \widetilde{\nu}(t) \right|^2 dt = O(\tau_L)$$

then $\|\mathcal{F}(\Pi_L f_{\partial B^x} - f_{\partial B^x})\|_\nu^2 = O\left(\tau_L \int |\Pi_L f_{\partial B^x}(b) - f_{\partial B^x}(b)|^2 db\right)$.

Discussion of Link Condition

- Condition (i) is a stronger condition if $\nu(t)$ is only a light tail distribution.
- If $\nu(t)$ is a heavy tail measure such as the log normal, then $\widetilde{\nu}(t) < \tau_L$ if and only if

$$(2\pi)^{-1/2} \exp\big(-(\log t)^2/2\big) < \tau_L t^{d-1}$$

or

$$t > \exp\Big(-2\log\big(\sqrt{2\pi}\,\tau_L t^{d-1}\big)\Big)$$

which is

$$t > \Big(\frac{1}{2\pi\,\tau_L^2}\Big)^{1/(2d-1)}.$$

Main Result: Rate of convergence

Theorem

Under the above assumptions we have

$$\int_{R^d} \left| \widehat{f}_{\partial B^x}(b) - f_{\partial B^x}(b) \right|^2 db$$
$$= O_p\left(\tau_L^{-1} \left(\frac{K^{d_w/(d_w-1)}}{n\lambda_K} + K^{(1-2r)/(d_w-1)} \right) + L^{-2s/d} \right)$$

- Result given sieve approximation error
  - (i) $\int_{R^d} |\Pi_L f_{\partial B^x}(b) - f_{\partial B^x}(b)|^2 db = O(L^{-2s/d})$
  - (ii) $\int_{R^{d-1}} |\gamma' \partial p^K(x) - \partial g(x)|^2 dw = O(K^{(1-2r)/(d_w-1)})$
- Obtain larger variance term due the first step estimation of the derivative of g, i.e., $K^{1/(d_w-1)}$

Rate of convergence

Corollary

Let $\lambda_K > 0$, $\tau_L \sim L^{-2a/d}$, $K \sim n^{(d_w-1)/(2r+d_w-1)}$, $L \sim n^{d/(2s+2a+d)}$. *If*

$$r/(d-1) \geq (s+a)/d$$

then $\int_{R^d} \left|\widehat{f}_{\partial B^\times}(b) - f_{\partial B^\times}(b)\right|^2 db$ *attains the rate*

1. $O_p\left(n^{(d/(d_w-1)-2s)/(2s+2a+d)}\right)$ *if* $2r\, d_w > (d-1)(2r-1)$
2. $O_p\left(n^{-2s/(2s+2a+d)}\right)$ *otherwise.*

Dicussion of Rate of convergence

- If $2r\,d_w > (d-1)(2r-1)$: Optimal rate for estimating the $d/(d_w-1)$–derivative in inverse problems with degree of ill-posedness a.
- If If $2r\,d_w \leq (d-1)(2r-1)$: Optimal rate for estimation in inverse problems with degree of ill-posedness a.
- For estimating the distribution of potential outcome no derivative of g is required
  $\Rightarrow$ Automatically obtain optimal rate $O_p\left(n^{-2s/(2s+2a+d)}\right)$.

Asymptotic Variance Formula

- Study the asymptotic behavior of

$$\sqrt{n}\big(\widehat{f}_{\partial B^{\times}}(b) - f_{\partial B^{\times}}(b)\big)$$

  for some b in the support of B

- Require a normalization factor V which increases with sample size at rate $\tau_L^{-1} L$

Normalization Factor

The normalization factor is given by

$$V(b) = \mathbb{V}\mathrm{ar}\left(\tau_L^{-1/2} q^L(b)' \left(\int R(t)\rho(t)d\nu(t)p^K(X) + S\,\eta\, p_d^K(X) \right) \right)$$

where

- $\eta = Y - g(X)$ and $\rho(t) = \exp(itY) - \varphi(X, t)$, for matrices S and $R(t)$

Asymptotic Pointwise Normality

Theorem

Given undersmoothing conditions on the dimension parameter L obtain

$$\sqrt{n/V(b-\partial g(x))}\left(\widehat{f}_{\partial B^{\times}}(b)-f_{\partial B^{\times}}(b)\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Monte Carlo Simulations

- Realizations of dependent variable Y are generated by

$$Y = B_0 + WB_1,$$

where $W \sim \mathcal{N}(0, 2)$, $B_0 \sim \mathcal{N}(0, 1)$,

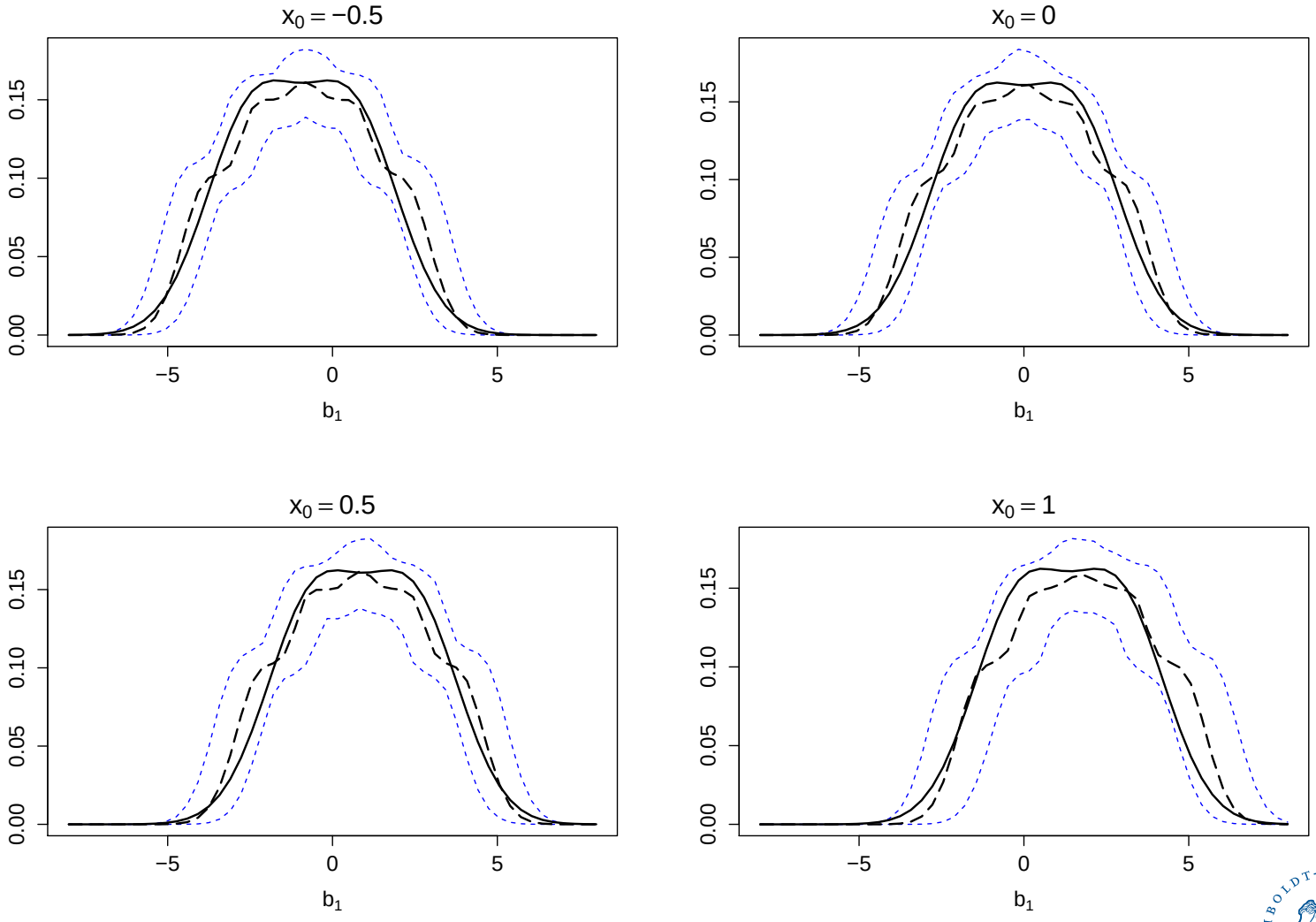
- Generate

$$B_1 = \sin(W) + A_1$$

and A_1 is a mixture of normals, i.e., $f_{A_1} = c_1\varphi_{-3/2,2} + c_2\varphi_{3/2,2}$

- Weighting function ν is lognormal density and use Hermite basis functions

Estimated Density of ∂B^{x_0}



Conclusion

1. This paper investigates the impact of both, observed and unobserved characteristics on individual heterogeneity
2. Combine ideas from RC and VC literature to study such complex heterogeneity
3. Method and model assumptions are convenient for implementation

Thank you!