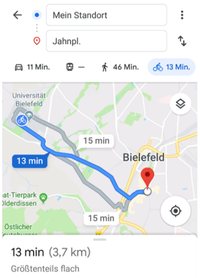
UNIVERSITÄT
BIELEFELD
Fakultät für
Wirtschaftswissenschaften

Non-parametric estimation of mixed discrete choice models

Dietmar BAUER, Manuel Batram, Sebastian Büscher

Modelling unobserved heterogeneity

- In many different applications one deals with the choice of one alternative out of many.
- One example is the choice of a transportation mode for a particular trip.
- The choice made depends on the characteristics of the alternatives for the trip (travel distance, travel time, costs involved) as well as the preferences of the decider (tradeoff between time and costs, for example; tastes).
- The preferences can be modelled using socio-demographic descriptions of the decider (sex, age, income, occupation, ...).
- Nevertheless often unexplained or unobserved taste heterogeneity remains.
- This unobserved heterogeneity has been modelled using the concept of 'mixing'.



Random utility models (RUM) for discrete choices

To fix notation:

- We observe the choice $y_{i,t}$ of several individuals $i = 1, \ldots I$ on several choice occasions $t = 1, ..., T_i$ out of a finite number J of well defined alternatives (which may differ from choice to choice).
- For each choice the set of available alternatives are characterised by their characteristics represented by $X_{j,i,t} \in \mathbb{R}^k, j = 1, ..., J$.
- The deciders are characterized via a set of sociodemographic variables $S_i \in \mathbb{R}^l$. These can be included into $X_{j,i,t}$ by interacting them with ASCs.

$$\text{Random utilities:} \quad U_{j,i,t} = \underbrace{X'_{j,i,t}\beta_i}_{V_{j,i,t}} + e_{j,i,t}$$

- $\beta_i \in \mathbb{R}^k$: individual specific preferences, β_i might depend on S_i.
- $V_{j,i,t}$... systematic utility of choice j in situation t for decider i.

Deciders then choose the alternative which delivers the highest utility.

Mixing distributions

The individual specific parameters β_i cannot be estimated consistently from a finite number of choices. More assumptions are needed.

Approaches:

1. **Latent class models:** We assume that there exist groups $I_s, s = 1, ..., S$ (group membership is not known to the modeller) such that $\beta_i = \beta_j$ if $i, j \in I_s$.
2. **Continuous mixing models:** The parameter β_i is chosen from an underlying distribution $Q(\beta)$.

The two concepts can be brought into one uniform framework, if the group membership is seen as a discrete random variable:

$$dQ_{LC}(\beta) = \sum_{s=1}^{S} \pi_s \delta_{\beta_s}(\beta)$$

Mixing model

Based e.g. on multinomial logit model:

$$\mathbb{P}(y_{i,t} = j | X_{:,i,t}, \beta_i) = p_j(X_{:,i,t}|\beta_i) = \frac{\exp(V_{j,i,t})}{\sum_{m=1}^{J} \exp(V_{m,i,t})}$$

- Mixed multinomial logit (MMNL)

$$\mathbb{P}(y_{i,t} = j | X_{:,i,t}) = \int_{\beta} p_j(X_{:,i,t}|\beta) dQ(\beta)$$

- Latent class model (with subclass logit mixing MNL; see Train, 2016):

$$\mathbb{P}(y_{i,t} = j | X_{:,i,t}) = \sum_{s=1}^{S} p_j(X_{:,i,t}|\beta_s)\pi_s$$

Basis can also be multinomial probit model (Bhat and Lavieri; 2018):

$$\mathbb{P}(y_{i,t} = j | X_{:,i,t}) = \int_{\beta} p_{j;MNP}(X_{:,i,t}|\beta) dQ(\beta)$$

Specification of the mixing distrbution

This raises the question, how to specify $Q(\beta)$:

- **type of distribution:** discrete (latent classes) or continuous.
- **class of continuous distributions:** support on real numbers, positive numbers, interval?
- **correlation structure** for different coordinates of the parameters.

And how to estimate $Q(\beta)$:

- latent class: specify support points β_s and frequencies π_s.
- continuous distribution: choice probabilities are not always known analytically $\Rightarrow$ MSL instead of ML?
- parametric or non-parametric form for distribution?

The literature provides partial answers. In particular the last point is only answered via trial and error comparing optimal likelihood values or cross validation.

Contribution

1. Discuss non-parametric maximum likelihood estimation for mixed models
2. Suggest adaptive grids as a method to decrease the problem of the curse of dimensionality
3. Combine the two techniques to obtain a proposed estimation algorithm.

Non-parametric MLE

2. Non-parametric MLE

- The statistics literature (e.g. surveys by Lindsay (1983) or Böhning (1995)) contains much information on the estimation of non-parametric mixed models, which is a more general situation than our setting.
- In general the mixing distribution can be estimated only provided identification holds, such that there exists a unique maximum of the limiting likelihood.
- In our setting there is a tradeoff between the variation in the characteristics $X_{i,j,t}$ and the identifiability of the models: The more flexible the distribution, the higher the requirements on variation in the characteristics.
- Fully flexible distributions $Q(\beta)$ require large support assumptions for the regressors (i.e. they need to be supported on $\mathbb{R}^k$).
- In order to identify latent class models, the characteristics $X_{j,i,t}$ need to be supported on sufficiently many support points (see e.g. Grün and Leisch, 2008).
- In between these extremes there are many variants of identification results, see e.g. Fox et al. (2012), Fox (2017) or the references in the recent papers Matzkin (2019), Chernozukov et al. (2019).

2. Non-parametric MLE

Results for the NP-MLE introduced by Kiefer and Wolfowitz (1956):

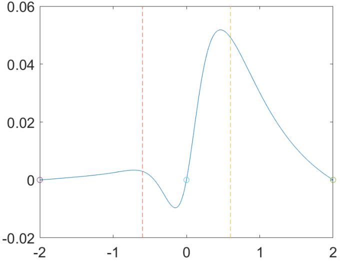
- Provided the regressors vary sufficiently, the mixing distribution can be estimated consistently, when optimization takes place over the space of all distribution functions.
- The maximum of the likelihood is achieved by a latent class model with a number of support points in the order of magnitude of the sample size.
- A distribution $\hat{Q}$ is optimal, if the directional derivative of the likelihood with respect to all point masses is non-positive:

$$D(\hat{Q}, \beta_n) = \lim_{\alpha \to 0} \frac{L((1-\alpha)\hat{Q} + \alpha\delta_{\beta_n}) - L(\hat{Q})}{\alpha} \leq 0$$

- If the directional derivative is positive at β_n, then adding a component with this support point improves the solution.
- Estimation algorithms for the location of the support points can be found using the expectation-maximization (EM) algorithm.

Using the directional derivative

- latent class MNP with one mixed parameter with support points $-0.6, 0.6$.
- evaluated at optimal fit with support points $-2, 0, 2$.
- zero at estimated support points, positive around true values.



Estimation Algorithm

These results suggest the following estimation scheme (see also Wang and Wang, 2013):

- Find an initial grid of support points β_s.
- For given support points β_s, find the optimal frequencies π_s from solving the convex, linearly constrained optimization problem:

$$L(\beta) = \sum_{i=1}^{n} \log\left(\sum_{s=1}^{S} \pi_s p(X_{:,i}|\beta_s)\right)$$

  Efficient algorithms for such optimization problems exist.
- Use the EM algorithm to update the support points β_s.
- Draw new support points β_n using the directional derivative as a guide. Continue at Step 2.

Extensions to situation with mixed deterministic and random coefficients have been derived (Bansai et al., 2018).

Adaptive grids

Curse of dimensionality

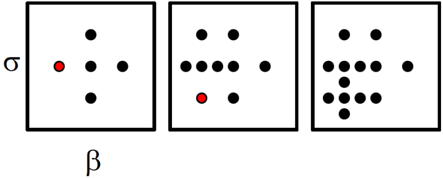
- The NP-MLE depends on the provision of an initial solution mostly consisting of the location of the support points β_s.
- The usage of a fixed grid (as advocated by Train (2016) e.g.) faces a curse of dimensionality problem for growing dimensions of β_s.
- The latent class model is only one possible approximation of the distribution $Q(\beta)$.
- Kernel density estimators show another approximation method:

$$Q(\beta) = \sum_{s=1}^{S} \pi_s \Phi(\beta; \mu_s, \Sigma_s)$$

  as a Gaussian mixture where Φ denotes a multivariate Gaussian distribution.
- This approximation can be more accurate for the same number of support points in some situations: e.g. mixing using a Gaussian distribution.
- The latent class model can be approximated via $\Sigma_s = I\frac{1}{k}$ for $k \to \infty$.

Adaptive grid

- Adaptive grids use a hierarchy in the grid architecture.
- They start with a coarse grid.
- Subsequently the grid is refined in directions that show potential.



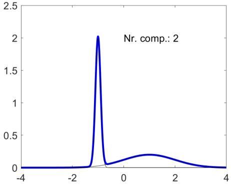
Plot adapted from Pflüger (2010).

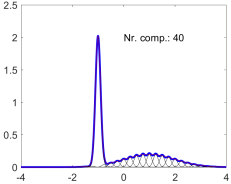
Adaptive grid in our case

- hierarchy: in terms of the variance in the various directions as well as the distance between support points on a fixed grid.
- Initialize with a coarse grid in terms of the location with a uniform variance matrix with independent components and relatively large variance.
- In every step with the current grid we refine in places where:
  - the frequency is substantial
  - the directional derivative of the new component is positive
- Hereby at each step for a support point to be refined we introduce new support points halfway to the next grid point on the same level in all directions and additionally with half the variance.
- From amongst all these potential new points we select only the ones with positive directional derivative.
- This procedure is iterated.

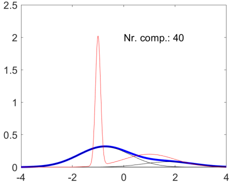
Example

True pdf: mix two normals:

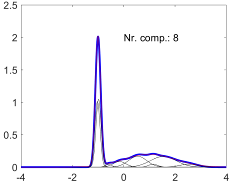




narrow grid



wide grid



adaptive grid

Proposed Algorithm

Combination of two approaches

Using the two main techniques we obtain the following estimation algorithm:

1. Initialize using an adaptive grid. Set $k = 1$.
2. Run 5 iterations of an EM algorithm to enhance the locations of the support points without changing the variances.
3. Generate new support points at randomly chosen promising locations according to $D(\beta, \hat{Q}^{(k)})$.
4. Set $k \to k + 1$ and continue at step 2 until convergence.

Small Scale Simulations

5. Small Scale Simulations

- cross sectional data set with sample size $n \in \{500, 1000, 2500, 5000\}$.
- three alternatives
- one regressor variable and two ASCs
- MNP model with identity as covariance matrix.

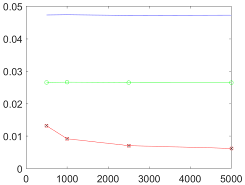
500 replications in two different scnearios:

Case 1 coefficient random with (a) two latent classes supported at ± 1; (b) univariate normally distributed; (c) log-normal

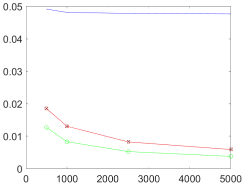
Case 2 the coefficient β is fixed to 1, two ASCs are random (a) multivariate normal with positive correlation; (b) mixture of two normals with different mean and variance.

Results Case 1

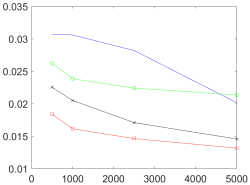
Mean absolute error in estimated choice probability for the choices:



(1a)
latent class



(1b)
normal distr.

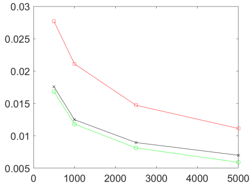


(1c)
log-normal

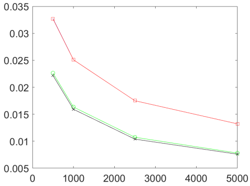
Adaptive grid in blue, proposed algorithm in green, EM based on fixed grid in black, best ML in red.

Results Case 2

Mean absolute error in estimated choice probability for the choices:



(2a)
normal



(2b)
mixed normal

Adaptive grid in blue, proposed algorithm in green, EM based on fixed grid in black, best ML in red.

Conclusions

6. Conclusions

- The proposed algorithm makes it possible to use NP-MLE techniques in the estimation of mixed discrete choice models.
- The adaptive grid techniques extend the range of dimensions that can be dealt with using the approach by accounting for both latent class models as well as Gaussian mixtures.
- Provided the regressor variables are sufficiently variable to ensure identification, the mixing distribution can be estimated consistently.
- Latent class models are included as a submodel.
- The distribution estimates provide estimates of key quantities of the mixing distribution: the expectation as well as the probability of one component to be positive.

Based on the proposed algorithm we are currently enlarging our experience in more complex and higher dimensional problems.

6. Conclusions

Thank you for your attention.
Questions?

Dietmar.Bauer@uni-bielefeld.de

References

7. References

- Prateek Bansal, Ricardo A Daziano, and Erick Guerra. Minorization-maximization (mm) algorithms for semiparametric logit models: Bottlenecks, extensions, and comparisons. *Transportation Research Part B: Methodological*, 115:17–40, 2018.
- Chandra Bhat and Patrícia Lavieri. A new mixed mnp model accommodating a variety of dependent non-normal coefficient distributions. *Theory and Decision*, 84(2):239–275, 2018.
- Dankmar Böhning. A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference*, 47(1-2):5–28, 1995.
- Victor Chernozhukov, Ivan Fernandez-Val, Whitney Newey. Nonseparable multinomial choice nmodels in cross-section and panel data. *Journal of Econometrics*, 211: 104-116, 2019.
- Jeremy Fox. A Note on Nonparametric Identification of Distributions of Random Coefficients in Multinomial Choice Models. Technical report, National Bureau of Economic Research, Cambridge, MA, jul 2017.
- Jeremy Fox, Kyoo il Kim, Stephen Ryan, and Patrick Bajari. The random coefficients logit model is identified. *Journal of Econometrics*, 166(2):204–212, 2012.
- Bettina Grün and Friedrich Leisch. Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification*, 25(2):225–247, 2008.
- Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.
- Bruce Lindsay. The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, pages 86–94, 1983.
- Rosa Matzkin. Constructive Identification in some nonseparable discrete choice models. *Journal of Econometrics*, 211, 83-103, 2019
- Dirk Pflueger. Spatially Adaptive Sparse Grids for High-Dimensional Problems. PhD Thesis, TU München, 2010.
- Kenneth Train. Mixed logit with a flexible mixing distribution. *Journal of Choice Modelling*, 19:40–53, 2016.
- Xuxu Wang and Yong Wang. Nonparametric multivariate density estimation using mixtures. *Statistics and Computing*, 25(2):349–364, 2013.