Universität Bielefeld

Using the maximum approximate composite marginal likelihood (MaCML) approach for the estimation of probit type random utility models

Manuel BATRAM and Dietmar BAUER

Random utility models (RUM) for discrete choices

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

Example: mode choice for a trip

- observing the transport modes used by a number of people on their daily routine
- different number of trips
- mode choice depends on characteristics of trip: distance travelled, mode specific costs, mode specific travel time
- and the characteristics of the decider: modes available, individual preferences, ...

Random utility models (RUM) for discrete choices

Such choices can be modelled using utility theory:

$$U_{j,i,t} = \underbrace{\alpha_{j,i} + X'_{j,t}\beta_i}_{V_{j,i,t}} + e_{j,i,t}$$

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

Deciders then choose the alternative which delivers the highest utility to them:

$$y_{i,t} = j \Leftrightarrow U_{j,i,t} \geq U_{m,i,t} \quad m = 1, ..., J.$$

Multinomial logit model (MNL)

Assumptions on noise term $e_{j,i,t}$ lead to different models:

Multinomial logit models (MNL):

- $e_{j,i,t}$ assumed to be iid over individuals, choice situation and alternatives, distributed according to Gumbel distribution.
- leads to the formula:

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

- easy to implement numerically
- individual preferences can be included using mixing of coefficient β_i:

$$\mathbb{P}(y_{i,t} = j | X_{:,t}) = \int_{\beta_i} \frac{\exp(\alpha_{j,i} + X'_{j,t}\beta_i)}{\sum_{m=1}^{J} \exp(\alpha_{m,i} + X'_{m,t}\beta_i)} f(\beta_i) d\beta_i$$

- These probabilities need to be simulated in general.

Multinomial probit models (MNP)

- $e_{:,i,t} \in \mathbb{R}^J$ assumed to be iid $(0, \Sigma)$ over individuals, choice situation, distributed according to a multivariate normal distribution.
- probabilities are given as:

$$\mathbb{P}(y_{i,t} = j | X_{:,t}) = \mathbb{P}(U_{m,i,t} - U_{j,i,t} \leq 0, m = 1, .., J | X_{:,t})$$

- Here $U_{m,i,t} - U_{j,i,t}, m \neq j$ is a multivariate normally distributed random variable. The choice probabilities are thus given by multivariate normal CDFs.
- Mixing with normal distributions can be included easily:

$$U_{m,i,t} - U_{j,i,t} = (X_{m,t} - X_{j,t})'\beta + \underbrace{e_{m,i,t} - e_{j,i,t} + (X_{m,t} - X_{j,t})'(\beta_i - \beta)}_{cond.norm.distr.\tilde{w}_{m,i,t}}$$

- Calculation of choice probabilities either using simulation or using approximations, both is time consuming.

Multinomial probit models (MNP) in the panel case

- In the panel case of repeated choices the choice situations are no longer independent, e.g. due to the terms

$$(X_{m,t} - X_{j,t})'(\beta_i - \beta).$$

- Thus the likelihood includes for each individual the probability

$$\mathbb{P}\left(\begin{bmatrix} y_{i,1} = j_1 \\ y_{i,2} = j_2 \\ \vdots \\ y_{i,T_i} = j_{T_i} \end{bmatrix} | X_{:,:}\right) = \mathbb{P}(U_{m,i,t} - U_{j_t,i,t} \leq 0, m = 1, .., J, t = 1, ..., T_i | X_{:,:})$$

- Calculation of the likelihood for MNP models in the panel case can involve the evaluation of high dimensional normal CDFs.
- The calculation of a high dimensional integral is time consuming, no matter how it is done.
- Approximations might not be accurate for large dimensions.
- Monte Carlo simulations may need large number of draws to result in accurate approximations.

Questions

1. Are there approximations that are numerically fast and 'reasonably accurate'?
2. What are the properties of these approximations relevant for estimation?
3. Can we replace MLE by a different concept that is easier to handle?
4. How do the approximation concepts compare to simulation based methods (MSL)?

Approximation Concepts

Main idea

- We want to evaluate the CDF of a multivariate normally distributed random variable:

$$\mathbb{P}(U_{m,i,t} - U_{j_t,i,t} = (X_{m,t} - X_{j,t})'\beta + \tilde{w}_{m,i,t} \leq 0, m = 1, .., J, t = 1, ..., T_i | X_{:,t})$$

- Separating systematic and random terms and rescaling we obtain w.r.o.g. $w_s \leq W_s, s = 1, ..., S$ for a normally distributed random vector $w = [w_s]_s$ with zero mean and variances equal to one: $\mathbb{P}(w_s \leq W_s, \forall s)$
- Using indicators $I_s = I(w_s \leq W_s)$ we obtain from conditioning

$$\mathbb{P}(I_s = 1, \forall s) = \mathbb{P}(I_1 = 1)\,\mathbb{P}(I_2 = 1 | I_1 = 1) \ldots \mathbb{P}(I_S = 1 | I_s = 1, s \leq S - 1)$$

- The sequence here is arbitrary, we could also use (or any other ordering)

$$\mathbb{P}(I_s = 1, \forall s) = \mathbb{P}(I_S = 1)\,\mathbb{P}(I_{S-1} = 1 | I_S = 1) \ldots \mathbb{P}(I_1 = 1 | I_s = 1, s \geq 2)$$

Approximation concepts use approximations to $\mathbb{P}(I_j = 1 | I_s = 1, s \in \mathcal{S})$ and provide ways to deal with ordering.

Solow-Joe (SJ)

Joe (1995) refines the method by Solow (1990) based on the idea that for an indicator variable the probability equals the expectation:

$$\mathbb{P}(I_j = 1 | I_s = 1, s \in \mathcal{S}) = \mathbb{E}(I_j | I_s = 1, s \in \mathcal{S}).$$

We obtain under the assumption of linearity:

$$I_j - \mathbb{E}\, I_j = \alpha_{j,\mathcal{S}}(I_{\mathcal{S}} - \mathbb{E}\, I_{\mathcal{S}}) + u_j$$

such that

$$\begin{aligned} \hat{\mathbb{P}}(I_j = 1 | I_s = 1, s \in \mathcal{S}) &= \mathbb{E}\, I_j + \alpha_{j,\mathcal{S}}(1 - \mathbb{E}\, I_{\mathcal{S}}) \Rightarrow \\ \hat{\mathbb{P}}(I_s = 1, \forall s) &= \hat{\mathbb{P}}(I_1 = 1)\hat{\mathbb{P}}(I_2 = 1 | I_1 = 1) \cdots \hat{\mathbb{P}}(I_S = 1 | I_s = 1, s \leq S - 1) \end{aligned}$$

Solow-Joe (SJ)

From OLS we know that

$$\alpha_{j,\mathcal{S}} = \Omega_{j,\mathcal{S}}\Omega_{\mathcal{S},\mathcal{S}}^{-1}$$

where $\Omega_{j,l} = \mathrm{Cov}(I_j, I_l) = \mathbb{E}(I_j - \mathbb{E}\, I_j)(I_l - \mathbb{E}\, I_l)$.

Thus we need to evaluate:

$$\begin{aligned} \mathbb{E}\, I_s &= \mathbb{P}(I_s = 1) = \mathbb{P}(w_s \leq W_s) = \Phi(W_s), \\ \mathbb{E}\, I_j I_l &= \mathbb{P}(I_j = 1, I_l = 1) = \Phi_2(W_j, W_l; \rho_{j,l}) \end{aligned}$$

This requires the evaluation of one and two dimensional normal CDFs.

For the ordering of the sequences there are three different versions:

SJ-1 One random ordering drawn separately for each calculation.

SJ-X Average over a fixed (say 10, 100 or 1000) number of random draws for each probability.

SJ-A Average over all possible orderings.

SJ: properties

- Joe (1995) recommends SJ-X over SJ-1 while SJ-A is infeasible for typical dimensions.
- Connors et al. (2014) recommend SJ-10 for MNP estimation.
- The approximation does not ensure that $\hat{\mathbb{P}}(y_{i,t} = j) \geq 0$. Therefore $\check{\mathbb{P}}(y_{i,t} = j) = \max(0, \hat{\mathbb{P}}(y_{i,t} = j))$ is used.
- Similarily $\sum_{j=1}^{J} \check{\mathbb{P}}(y_{i,t} = j) = 1$ does not necessarily hold. Consequently a *normalisation* may be used such that

$$\tilde{\mathbb{P}}(y_{i,t} = j) = \frac{\check{\mathbb{P}}(y_{i,t} = j)}{\sum_{s=1}^{J} \check{\mathbb{P}}(y_{i,t} = s)}$$

- It can be shown (Batram and Bauer, 2019) that the mapping $[W_s] \mapsto \tilde{\mathbb{P}}(y_{i,t} = j)$ for given ordering for SJ-1 and SJ-X is twice continuously differentiable and surjective for the simplex of choice probabilities.

Mendell-Elston (ME) approximation

- This approximation is based on the idea that the distribution of the normal random variable w_s conditioned on $I_s = 1, s \in \mathcal{S}$ is no longer normal but may be approximated as a normal distribution whose conditional mean and conditional variance can be calculated.
- A recursion formula is contained in the paper Mendell and Elston (1974).
- This formula only uses the evaluation of the pdf and cdf of univariate normal random variables.
- Given an ordering it is simple to code and provides expressions for derivatives.
- With respect to the ordering a number of systematic orderings exist that are said to enhance the accuracy. Also random orderings are possible.
- There also exist bivariate extensions. Some newer developments use similar ideas (Bhat, 2018; TVBS).

ME: properties

- Properties are similar to the SJ case: non-negativity is guaranteed, but the summation property needs to be ensured by normalisation.
- For fixed ordering it can be shown (Batram and Bauer, 2019) that the mapping $[W_s] \mapsto \tilde{\mathbb{P}}(y_{i,t} = j)$ is twice continuously differentiable and surjective for the simplex of choice probabilities.
- However, when using a systematic ordering, the mapping becomes discontinuous and potentially non-surjective.

Composite Marginal Likelihood (CML)

Definition of Composite Marginal Likelihood

- The calculation of the likelihood in the panel case is computationally hard, since all choices of one person need to be included in one high dimensional CDF calculation.
- Even for moderate size of the choice set this can be demanding.
- Identification of all parameters in typical panel models can be achieved from only two choices per person.
- Thus instead of maximizing the likelihood one can use a criterion function of the form

$$CML(\theta) = \prod_{i=1}^{I} \prod_{t=2}^{T_i} \prod_{s=1}^{t-1} \mathbb{P}(y_{i,s}, y_{i,t} | X_{:,:}; \theta)$$

- Alternatives involve only consecutive pairs of choices.
- This is a special case of the more general composite marginal likelihood methods (Varin et al., 2011).
- Analogously to quasi-likelihood methods this leads to consistent, asymptotically normal estimators that are not efficient (although the loss in efficiency might be small).

The MACML approach

- Bhat (2011) combines the two ideas to the *maximum approximate composite marginal likelihood* (MACML) estimator: The pairwise CML function is used as the criterion function where the normal CDF is approximated using SJ-1.
- Bhat and Sidhartan (2011) show by simulation in a panel data context that the method provides estimation that is numerically many times faster than MSL when achieving similar accuracy (or more accurate when using similar computation times).
- Asymptotic behaviour of the estimators are not provided in the article.
- Only a limited comparison of the different approximation concepts is provided.

Properties of MACML for estimation

Simple cross sectional example

To illustrate the features of using the approximations for estimation of MNP models consider the following example:

- Individuals choose from 4 alternatives.
- There are no characteristics apart from ASCs. The variance matrix of the errors is known for the estimation.
- The ASCs for the first three alternatives are known, only the fourth, θ say, is estimated.
- $\theta_0 = 0$: The true value of the fourth ASC equals zero.
- We investigate the limiting scaled log likelihood function:

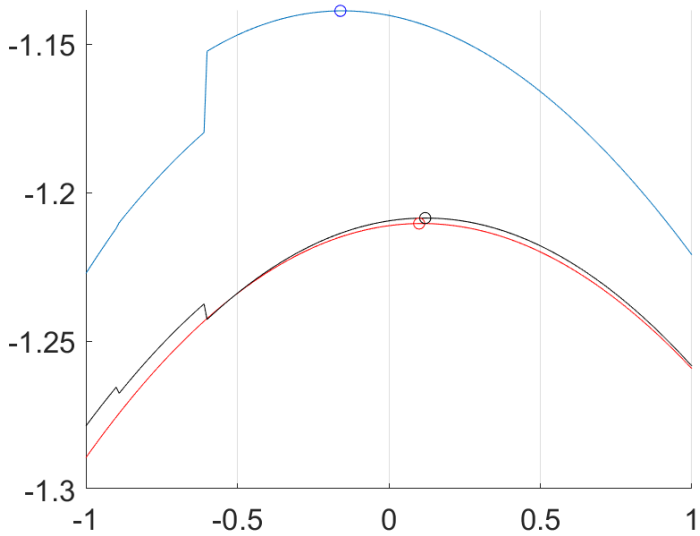
$$ll_\infty(\theta) = \sum_{j=1}^{4} \log \hat{\mathbb{P}}^{(m)}(j;\theta)\, \mathbb{P}_0(j)$$

where $\hat{\mathbb{P}}^{(m)}$ denotes the approximated choice probability (for m : *bvnu**, $SJ-1$, *ME*) and $\mathbb{P}_0(j)$ the true choice probabilities.

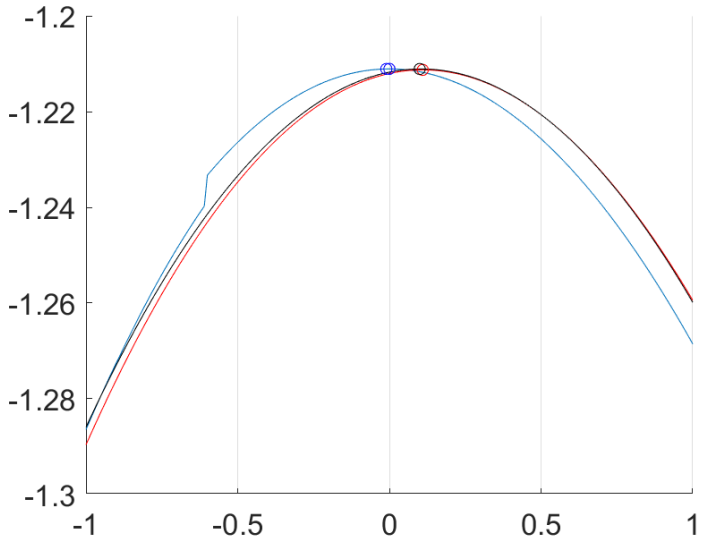
- This reveals the large sample bias.

* bvnu denotes the MVN CDF implementation of Alan Gentz in MATLAB.

Simple cross sectional example



No normalisation, systematic ordering



with normalization, no syst. ordering.

Colours: blue: bvnu, red: SJ-1, black: ME.

- Estimators of the parameters based on approximations are even asymtotically biased!
- Situation is much better for normalized approximations.

Simple cross sectional example

- In many cases we want to estimate choice probabilities and not the parameters θ per se.
- The limiting criterion function in the example equals

$$ll_\infty(\theta) = \sum_{j=1}^{4} \log \hat{\mathbb{P}}^{(m)}(j;\theta)\,\mathbb{P}_0(j)$$

- Over the simplex of probability distributions this is always optimized at $\hat{\mathbb{P}}^{(m)}(j;\hat{\theta}) = \mathbb{P}_0(j)$.
- Since in all cases we obtained surjective mappings onto the simplex asymptotically $\hat{\theta}$ corresponds to the true choice probabilities.
- Even if we obtain a bias, this only relates to the parameters. The choice probabilities are estimated consistently.

Second example: adding the impact of regressors

Consider the choice between the four alternatives according to the utility:

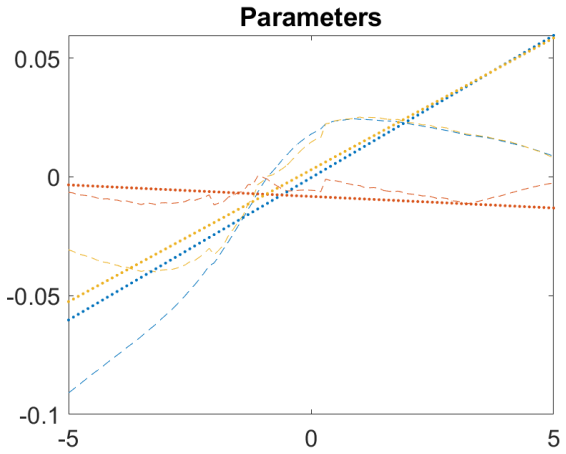
$$U_{j,i} = \alpha_j + \beta_j x_i + u_{j,i}$$

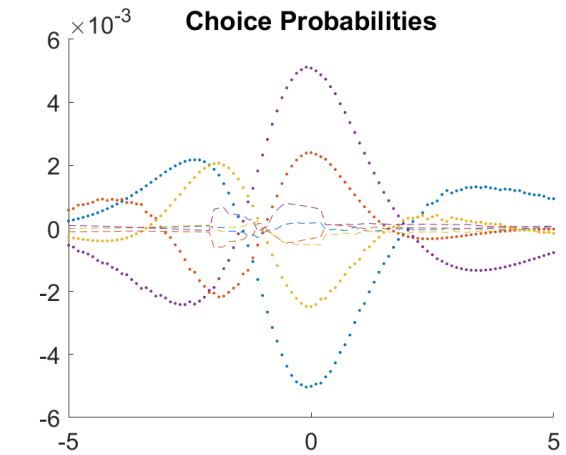
where $u_{:,i} \sim \mathcal{N}(0, \Sigma)$.

For the modeling and estimation two scenarios are possible:

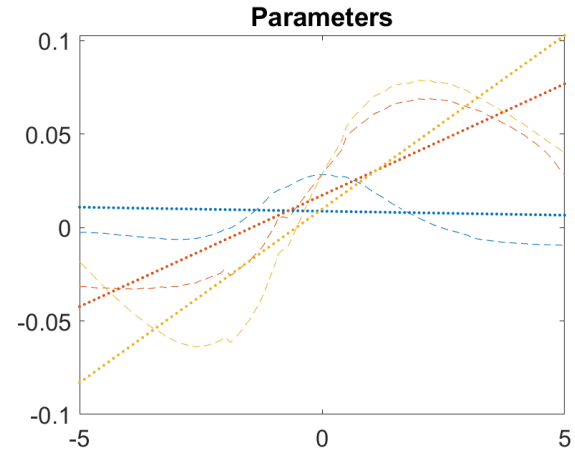
1. Estimation uses a linear model as above for the approximations: In this case the estimates of both the parameters and the choice probabilities will be biased.
2. Estimation uses a flexible (non-parametric) model for the dependence on x_i: In this case the parameter estimates will be biased but the probability estimates will be unbiased.

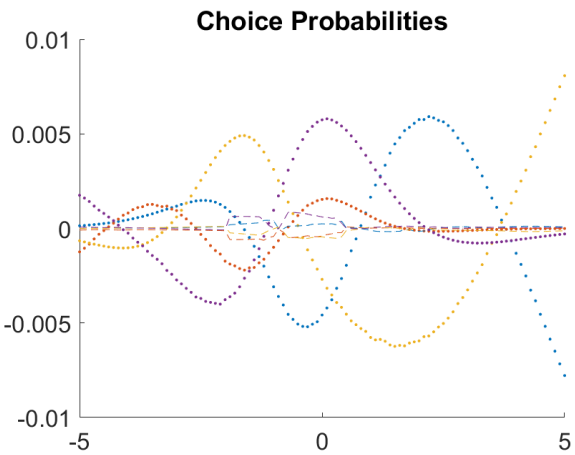
Second example: adding the impact of regressors





SJ-A, normalized





ME, normalized

Implications for more general situations

- The panel case is conceptually no different from the examples. All results carry over.
- This also holds true for the CML based estimators, as they basically match pairwise choice probabilities.
- Choice probabilities can always be estimated consistently using non-parametric methods (flexible dependence on underlying characteristics).
- Parameter estimates are biased even asymptotically.
- If one uses a fixed functional form (such as linear dependence on regressors), also choice probabilities show an asymptotic bias.

Different view of modeling

- All models implement for given regressors a mapping from parameters to choice probabilities:
$$\mathbb{P}(y_{i,t} = j | X_{:,t}) = g_j(X_{:,t}; \theta)$$
- Approximations to the MVN CDF provide maps that are approximations to the map induced by the MNP.
- The parameters for this approximated map can be estimated consistently and asymptotically normally using the MACML approach. Standard theory applies in this respect.
- The estimated model should be used with the same approximations, not the MNP form for predictions.
- The interpretation of a RUM is lost (or only holds in an approximate sense).

This is the identical situation of using the (mixed) MNL model for data generated from a MNP model, only that the differences might be smaller.

Does this matter? Finite Sample performance

- 40 cross sectional data sets of size 2500.
- 6 alternatives, depending on 6 regressor variables $X_i \in \mathbb{R}^6$ drawn from independent standard normal distributions.
- The preferences are given by a parameter vector β for each individual drawn from a normal distribution with

$$b = \mathbb{E}\,\beta_i = \begin{pmatrix} 1.5 \\ -1 \\ 2 \\ 1 \\ -2 \\ 0.5 \end{pmatrix}, \operatorname{Var}\beta_i = \begin{pmatrix} 1 & -0.5 & 0.25 & 0.75 & 0.3 & 0 \\ -0.5 & 1 & 0.25 & -0.5 & 0.2 & 0 \\ 0.25 & 0.25 & 1 & 0.33 & -0.15 & 0 \\ 0.75 & -0.5 & 0.33 & 1 & 0.2 & 0 \\ 0.3 & 0.2 & -0.15 & 0.2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

- The noise variance $\Sigma = 0.5 I_6$ is assumed known.
- MNP models are estimated with four different approximations: SJ-1, ME, TVBS, MSL (all in our own coding).
- Numerical search is started at the true values.
- For each trial we record the average percentage bias (APB) for the parameters and the mean computation times on a standard PC (in minutes).

Finite Sample Results

	no normalisation				normalisation			
	SJ-1	ME	TVBS	*MSL**	SJ-1	ME	TVBS	*MSL**
APB (b)	1.96	5.25	1.00	1.85	1.90	1.83	1.20	1.84
APB (all)	4.89	13.62	4.07	4.86	4.95	2.90	4.01	4.87
time (min)	1.50	2.19	13.86	8.37	9.17	15.73	87.06	51.04

- Even with normalisation a bias is visible.
- For estimating *b* TVBS is the most accurate method in this scenario.
- This is true both for the normalized and the unnormalized scenario.
- MSL is tuned to provide comparable accuracy but uses more time than some competitors.
- Normalisation helps the ME procedure most, while the others are not affected much.

* MSL uses 2000 Sobol draws.

Conclusions

5. Conclusions

- Using approximations to the Gaussian CDF reduces computation times for estimation.
- Normalisation of the probabilities are important for some approximations. Others appear to be more robust.
- Systematic reordering of components in the approximation can have dramatic side effects.
- In the panel case replacing the likelihood by the CML has the same goal.
- The approximation leads to bias terms which in simple examples are small but noticeable.
- Choice probabilities for given regressor values can be estimated consistently.
- Using flexible forms of dependence structures allows consistent estimation for all regressor variables. For fixed functional form biases result whose size is currently unclear.

References

6. References

- Batram, M., Bauer, D. 2019. On consistency of the MACML approach to discrete choice modelling. *Journal of Choice Modeling*, **30**, 1-16.
- Bhat, C.R., 2011. The maximum approximate composite marginal likelihood (macml) estimation of multinomial probit-based unordered response choice models. *Transport. Res. Part B* **45** (7), 923–939.
- Bhat, C.R., 2018. New Matrix-based Methods for the Analytic Evaluation of the Multivariate Cumulative Normal Distribution Function. *Transportation Research Part B* (forthcoming).
- Bhat, C.R., Sidhartan, R., 2011. A simulation evaluation of the maximum approximate composite marginal likelihood (macml) estimator for mixed multinomial probit models. *Transport. Res. Part B* **45** (7), 940–953.
- Connors, R., Hess, S., Daly, A., 2014. Analytic approximation for computing probit choice probabilities. *Transportmetrica* **10** (2), 119–139.
- Joe, H., 1995. Approximations to multivariate normal rectangle probabilities based on conditional expectations. *J. Am. Stat. Assoc.* **90** (431), 957–964.
- Mendell, N., Elston, R., 1974. Multifactorial qualitative traits: genetic analysis and prediction of recurrence risks. *Biometrics* **30**, 41–57.
- Solow, A.R., 1990. A method for approximating multivariate normal orthant probabilities. *J. Stat. Comput. Simulat.* **37**, 225–229.
- Varin, C., Reid, N., Firth, D., 2011. An overview of composite likelihood methods. *Stat. Sin.* **21**, 5–42.