UNIVERSITÄT BIELEFELD
Fakultät für Wirtschaftswissenschaften

Determinants for the efficiency loss due to using a composite marginal likelihood for estimating a probit model in the panel setting

Dietmar BAUER, Sebastian Büscher, Lennart Oelschläger

1. Introduction

- In this talk we deal with probit modelling for the choice of one alternative out of finitely many, say J.
- One example is the choice of a transportation mode for a particular trip.
- The choice $y_{i,t} \in \{1, ..., J\}$
  - made by the i-th decider (out of n)
  - in the t-th decision (out of T_i)
  - depends on the characteristics $X_{j,i,t} \in \mathbb{R}^k$ of the alternatives for the trip (travel distance, travel time, costs involved, ...).
- Often this is modelled using *random utility models*, where the choice is assumed to be made based on maximizing the underlying random utility:

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

- $V_{j,i,t}$ ... systematic utility of choice j for decider i in her t-th choice situation.

Modeling Preferences

- In the above formulation the individual preferences are encoded in the parameter β_i which depends on the decider i.
- Sometimes the preferences can be modelled explicitly using socio-demographic descriptions $S_i \in \mathbb{R}^d$ of the decider (for example sex, age, income, occupation, attitudinal surveys, ...):

$$\beta_i = f(S_i; \beta)$$

- Nevertheless often unexplained or unobserved taste heterogeneity remains.
- This unobserved heterogeneity has been modelled using the concept of 'mixing' which assumes a distribution of the random effects β_i in the population

Mixed Multinomial Probit Model (MMNP)

The MMNP model is defined via specifying the distributions of β_i, $e_{j,i,t}$.

MMNP model:

1. **Normally distributed errors:** The vector $e_{:,i,t} \in \mathbb{R}^J$ is jointly normally distributed with mean zero and variance Σ, iid over individuals and choice occasions.
2. **Random Effects:** $\beta_i \sim \mathcal{N}(\beta, \Omega)$: individual specific effects are (I) iid over individuals, (II) independent of the characteristics $X_{j,i,t}$ and S_i and (III) normally distributed.

This does not lead to a closed form solution for the choice probabilities, but contributions of one individual to the likelihood of the form (recoding S_i to be represented in $X_{j,i,t}$):

$$\begin{aligned}\mathbb{P}(y_{i,t} = j_t | X_{:,i,t}, \forall t; \beta, \Omega, \Sigma) &= \mathbb{P}(e_{l,i,t} - e_{j_t,i,t} \leq V_{j_t,i,t} - V_{l,i,t}, \forall l \neq j_t, \forall t; \beta, \Omega, \Sigma) \\ &= \mathbb{P}(\tilde{e}_i \leq \tilde{V}_i; \beta, \Omega, \Sigma)\end{aligned}$$

Thus we need to evaluate a $(J-1) \times T$ dimensional Gaussian CDF!

Composite Marginal Likelihood

Compromise between numerical speed ...

- The evaluation of a large Gaussian CDF is numerically cumbersome as it involves numerical integration (except for special cases).
- Therefore approximations are used (simulation, quadrature, ...). But this still takes time.
- An alternative is the **composite marginal likelihood** approach.
- This replaces the likelihood by an alternative criterion function that is easier to calculate.
- Examples are
  - the *independence likelihood* (ignoring correlations over time, only using marginals),
  - the *pairwise likelihood* (using pairs of decisions) and
  - the *full likelihood* (MLE).
- All approaches define for each individual a function $f(y_i, X_i; \theta)$ measuring the fit for this decider with characteristics X_i and choices y_i.

2. Composite Marginal Likelihood

... and statistical accuracy

- In all cases under identifiability (imposing enough restrictions on the parameterisation) we obtain consistent and asymptotically normal estimators.
- The asymptotic variance equals (sandwich form)

$$(\mathbb{E}\,\partial^2 f(y_i, X_i; \theta_0))^{-1}(\mathbb{E}(\partial f(y_i, X_i; \theta_0) f(y_i, X_i; \theta_0)'))(\mathbb{E}\,\partial^2 f(y_i, X_i; \theta_0))^{-1}.$$

- This is smallest for the full likelihood where the information equality implies

$$-\mathbb{E}\,\partial^2 f(y_i, X_i; \theta_0) = \mathbb{E}(\partial f(y_i, X_i; \theta_0) f(y_i, X_i; \theta_0)').$$

Therefore there is a trade-off between numerical speed and statistical accuracy when we pick an appropriate CML.

What is known about the relative merits?

- Different versions of the CML have been proposed.
- Two components in these CMLs are the marginal probability of single choices (independence CML) as well as pairs of choices (pairwise CML).
- When using pairs often not all pairs are included but only adjacent pairs $(t, t+1)$ further reducing the numerical load.
- Also both, marginal and pairwise CML can be combined, see Cox and Reid (2004).
- Bhat and Sidhartan (2011) and Katsikatsou et al (2012) provide simulation evidence that the efficiency loss is only modest when using the pairwise CML.
- Cessie and Houwelingen (1994), Kuk and Nott (2000) and Joe and Lee (2008) all deal with weighting in CML dealing mostly with unbalanced cases where individuals face different number of choice situations or situations with errors correlated over time -> talk by Sebastian Büscher tomorrow.

Contribution of this paper

We contribute to answers for the following questions:

1. Which method suffers how much? What is the relative ranking?
2. Which parameters are most concerned?
3. How bad can it be? What is the worst case relation of variances that we find?
4. Which situations are bad for CML estimation?

We include the following CMLs for comparison:

- full likelihood (MLE): all choices are modelled jointly.
- full pairwise likelihood (FP): using all pairs of choices ($T(T-1)/2$ pairs)
- adjacent pairwise likelihood (AP): using only adjacent pairs $((t, t+1), t = 1, ..., T)$
- adjacent pairwise likelihood plus independence likelihood (APU): $T-1$ pairs plus T marginal probabilities

Approach

3. Approach

We use a simple model that allows to separate a number of effects:

- MMNP with two alternative varying regressors iid uniformly $[-1, 1]$ distributed. [Leaving out the ASCs identifies the level.]
- $\beta = [\tilde{\beta}_1, 1]$: The second coefficient is fixed to 1 [This fixes the scale.].
- The first $\tilde{\beta}_1 \sim \mathcal{N}(\beta_1, \omega^2)$ is drawn randomly from a normal distribution with expectation β_1 and variance ω^2 (random effect).
- The error $e_{:,i,t} \in \mathbb{R}^J$ is normally distributed with expectation zero and variance $\sigma^2 I_J$.
- Each individual chooses from $J = 2$ alternatives in T consecutive choice occasions (balanced case).
- The choice probabilities for the choices y_i of the i-th decider depend on X_i and the true parameters $\theta_\circ = [\beta_1, \omega, \sigma]$ and are given as $p(y_i|X_i; \theta_\circ)$.

Asymptotic variance

$$(\mathbb{E}\,\partial^2 f(y_i, X_i; \theta_0))^{-1}(\mathbb{E}(\partial f(y_i, X_i; \theta_0) f(y_i, X_i; \theta_0)'))(\mathbb{E}\,\partial^2 f(y_i, X_i; \theta_0))^{-1}.$$

- The asymptotic variance is given in the sandwich form involving $\mathbb{E}\, h(y_i, X_i; \theta_0)$ for $h = \partial^2 f$ and $h = (\partial f)(\partial f)'$.
- Noting that y_i is chosen conditional on X_i we can write this as

$$\mathbb{E}_X\left(\mathbb{E}_y(h(y, X; \theta_0)|X)\right) = \mathbb{E}_X\left(\sum_y p(y|X; \theta_\circ) h(y, X; \theta_0)\right)$$

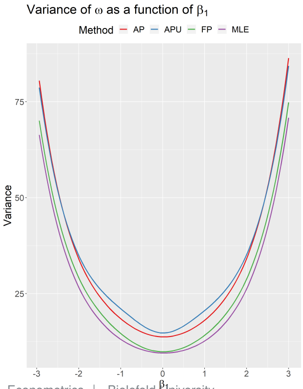
- Analytic expressions for $p, \partial f, \partial^2 f$ are used.
- Expectations over X are calculated by assuming a discrete uniform distribution over M randomly drawn points.
- We experimented with different number of draws M as well as different distributions (normally distributed, continuous uniformly distributed).
- For $M \geq 100$ the results do not change qualitatively.

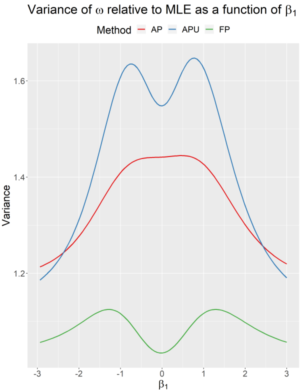
Results

4. Results

$\beta_1 \in (-3, 3), \omega = 0.02, \sigma = 1, T = 3, J = 2.$

- As β_1 gets larger in absolute value the variance increases. This happens for all parameters.
- For estimation of ω the relative performance gets worst around ±1.
- APU is significantly worse than AP for small values of β_1, for values larger than 2 the addition of the marginals pays off.

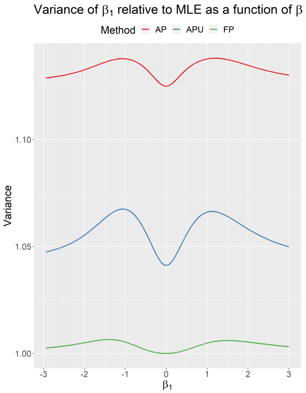

Variance of ω as a function of β_1

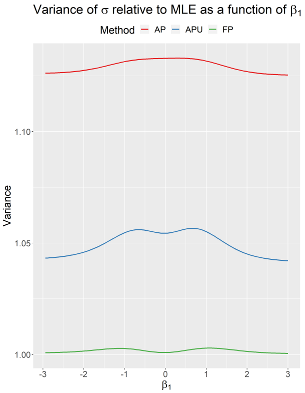

Variance of ω relative to MLE as a function of β_1

4. Results

Estimation of β_1, σ

- For $T = 3$ the loss in FP is only marginal.
- APU beats AP considerably: While the loss for AP amounts to roughly 13%, APU only uses about 5% in accuracy.


Variance of β_1 relative to MLE as a function of β

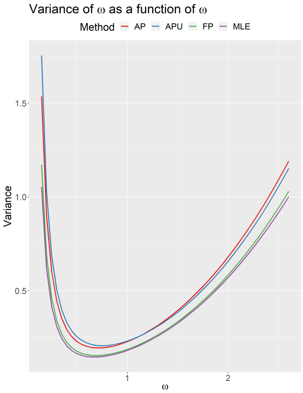

Variance of σ relative to MLE as a function of β_1

4. Results

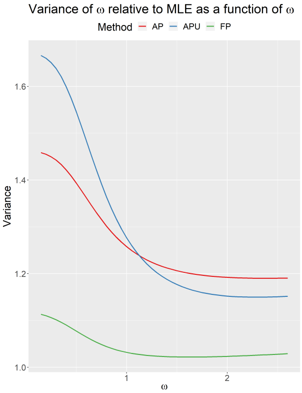
Varying ω

$$\beta_1 = 1, \omega \in (0.15, 2.6), \sigma = 1$$

- When estimating ω the situation of small ω is particularly critical.
- APU comes close to 70% efficiency loss, AP 45% while FP looses only 15%.



Variance of ω as a function of ω



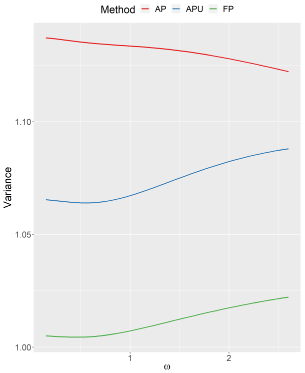
Variance of ω relative to MLE as a function of ω

4. Results

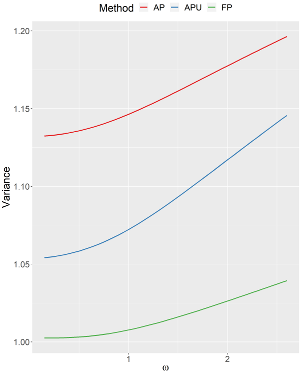
Estimation of β_1, σ

- For small ω the loss in FP is only marginal, for larger ω modest but increasing.
- APU again beats AP considerably: For both the accuracy loss is less pronounced than for the estimation of ω.

Variance of β_1 relative to MLE as a function of ω



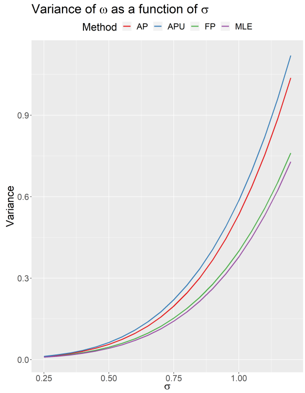
Variance of σ relative to MLE as a function of ω

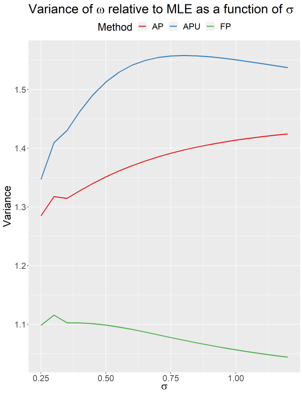


Varying *sigma*

$\beta_1 = 0.5, \omega = 0.25, \sigma \in (0.25, 1.2)$

- estimating ω gets harder for larger σ (more noise)
- for ω AP beats APU , where the efficiency loss can be substantial.

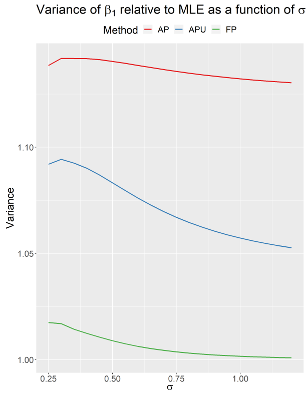

Variance of ω as a function of σ


Variance of ω relative to MLE as a function of σ

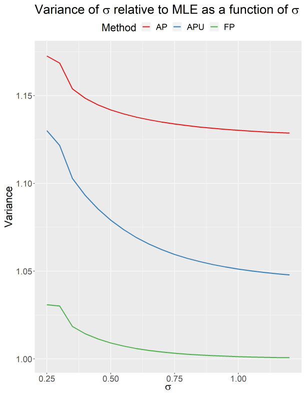
4. Results

Estimation of β_1, σ

- the loss in FP is only marginal, for AP and APU larger σ leads to decreasing modest efficiency losses.
- APU again beats AP considerably.



Variance of β_1 relative to MLE as a function of σ

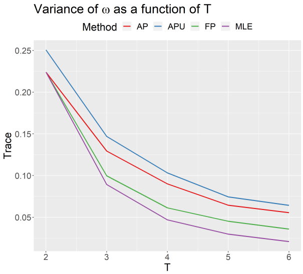


Variance of σ relative to MLE as a function of σ

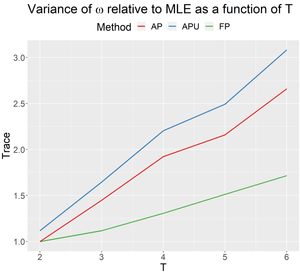
The influence of the number of choice occasions T_i

$$\beta_1 = 0.5, \omega = 0.1, \sigma = 0.5$$

- As expected the variance decreases as the number of choice situations increases.
- At the same time for the estimation of ω relative to MLE the other variants get worse the larger T_i.



Variance of ω as a function of T

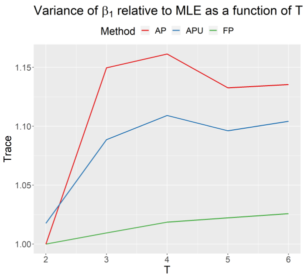


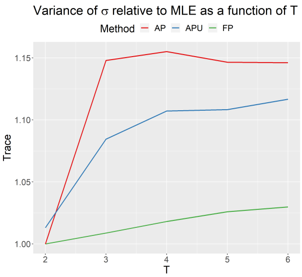
Variance of ω relative to MLE as a function of T

4. Results

Estimation of β_1, σ

- As expected the variance decreases as the number of choice situations increases (not shown).
- Estimation of β_1, σ relative to MLE the other variants do not get worse for larger T_i.


Variance of β_1 relative to MLE as a function of T


Variance of σ relative to MLE as a function of T

Conclusions

5. Conclusions

- Our results show that (not surprisingly) the biggest influence on the relative efficiency lies in the number of choice situations T: Overall the relative performance of all CML procedures compared to MLE degrades with increasing T.
- For the estimation of β_1, σ the effect is not as bad as for the estimation of ω^2.
- AP and APU are considerably worse than FP (also not surprising).
- Comparing AP and APU we see that including the marginals reduces the variance for estimating β_1 and σ, while it increases the variance for ω (except for situations of large values of β_1 and/or ω).
- In most cases the efficiency loss for FP is modest (less than 10% for $T = 3$), for estimating β_1, σ also for larger T up to $T = 6$.
- But situations with many choice occasions can lead to significant efficiency losses for estimating ω also for FP .

5. Conclusions

Thank you for your attention.

Questions?

Dietmar.Bauer@uni-bielefeld.de

References

6. References

- Bhat, C. R., Sidharthan, R. (2011). A simulation evaluation of the maximum approximate composite marginal likelihood (MACML) estimator for mixed multinomial probit models. *Transportation Research Part B: Methodological*, **45(7)**, 940-953
- Bhat, C. R., Varin, C., Ferdous, N. (2010). A comparison of the maximum simulated likelihood and composite marginal likelihood estimation approaches in the context of the multivariate ordered-response model. In 'Maximum simulated likelihood methods and applications.' Emerald Group Publishing Limited.
- Cox, D. R., Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, **91(3)**, 729-737.
- Joe, H., Lee, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis*, **100(4)**, 670-685.
- Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics and Data Analysis*, **56(12)**, 4243-4258.
- Kuk, A. Y., Nott, D. J. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statistics & Probability Letters*, **47(4)**, 329-335.
- le Cessie, S., Van Houwelingen, J. C. (1994). Logistic regression for correlated binary data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **43(1)**, 95-108.

6. 0.

Bhats favourite.

Bhat et al. (2010): ’As in the CMOP case, the estimated efficiency of the CML approach is as good as the MSL approach in the low correlation case (the relative efficiency ranges between 90%-103%, with a mean of 97%). For the high correlation case, the relative efficiency of parameters in the CML approach ranges between 82%-96% (mean of 91%) of the efficiency of the MSL approach’