Recovery of inter- and intra-personal heterogeneity using mixed logit models

Stephane Hess* Kenneth E. Train†

Abstract

Most applications of discrete choice models in transportation now utilise a random coefficient specification, such as mixed logit, to represent taste heterogeneity. However, little is known about the ability of these models to capture the heterogeneity in finite samples (as opposed to asymptotically). Also, due to the computational intensity of the standard estimation procedures, several alternative, less demanding methods have been proposed, and yet the relative accuracy of these methods has not been investigated. This is especially true in the context of work looking at joint inter-respondent and intra-respondent variation. This paper presents an overview of the various different estimators, gives insights into some of the theoretical properties, and analyses their *performance* in a large scale study on simulated data. In particular, we specify 31 different forms of heterogeneity, with multiple versions of each dataset, and with results from over $16,000$ mixed logit estimation runs. The findings suggest that variation in tastes over consumers is captured by all the methods, including the simpler versions, at least when sample size is sufficiently large. When tastes vary over choice situations for each consumer, as well as over consumers, the ability of the methods to capture and differentiate the two sources of heterogeneity becomes more tenuous. Only the most computationally intensive approach is able to capture adequately the two sources of variation, but at the cost of very high run times. Our results highlight the difficulty of retrieving taste heterogeneity with only cross-sectional data, providing further evidence of the benefits of repeated choice data. Our findings also suggest that the data requirements of random coefficients models may be more substantial than is commonly assumed, further reinforcing concerns about small sample issues.

Keywords: simulation-based estimation; approximation; random taste heterogeneity; mixed logit; intra-respondent; panel data

*Institute for Transport Studies, University of Leeds, s.hess@its.leeds.ac.uk, Tel: +44 (0)113 34 36611

†Department of Economics, University of California, Berkeley, train@econ.berkeley.edu, Tel: +1 415 291 1023

1 Introduction

In part as a result of improved estimation performance (cf. Bhat, 2001, 2003; Hess et al., 2006), researchers and practitioners in transportation and beyond are increasingly making use of the mixed logit model (cf. McFadden and Train, 2000; Train, 2009) for the representation of random taste heterogeneity across consumers. While mixed logit has clear theoretical advantages over specifications that assume taste homogeneity, relatively little attention has been paid to the question as to how well the estimated models are able to recover the *true* patterns of heterogeneity present in the data, especially with the sample sizes that are typically used in practice.

When tests have been performed (see e.g. Munizaga and Alvarez-Daziano, 2005; Cherchi and Guevara, 2009; Cherchi and Ortúzar, 2010; Rose et al., 2011), they have generally been limited to a few specifications and have focussed only on the variation in tastes over consumers in the context of data containing multiple choices by each consumer (i.e., panel data). Such an inter-respondent treatment of heterogeneity alongside an assumption of constant tastes across choices for the same respondent (cf. Revelt and Train, 1998) is now common-place, and has been shown to lead to very significant improvements in model fit as well as more reasonable estimates of taste variation (see for example the discussions in Hess and Rose, 2009). Recently, however, a number of authors (Bhat and Castelar, 2002; Bhat and Sardesai, 2006; Hess and Rose, 2009; Cherchi et al., 2009; Yáñez et al., 2011) have argued that tastes can vary across tasks for the same consumer and that this "intra-personal" heterogeneity occurs in addition to the variation over consumers (i.e., the "inter-personal" heterogeneity).

With the growing reliance on mixed logit models, the validity of the estimation results is of great importance, independently of whether analysts rely on cross-sectional or repeated choice data, and if the latter, whether intra-respondent heterogeneity is accommodated in addition to inter-respondent heterogeneity. The aim of the present paper is to test the ability of various mixed logit specifications to recover the true patterns of taste heterogeneity. We examine cases with inter-personal heterogeneity only, as well as cases with both intra- and inter-personal heterogeneity. We estimate the models on different sample sizes and with different types and levels of heterogeneity.

An important additional point arises here. In the case of cross-sectional data, only one estimator applies, with the likelihood given by a product of integrals of individual logit probabilities. In the case of repeated choice data with inter-respondent heterogeneity but intra-respondent homogeneity, the correct specification is the estimator in Revelt and Train (1998), with the likelihood still given by a product of integrals, where the integrands are however now sequences of choices, i.e. themselves a product of logit probabilities. This thus leads to a more complex specification, and an alternative may be to rely on a cross-sectional specification, which would entail estimating the model *as if* the multiple choices of each consumer were from different consumers. A variation on this cross-sectional approach has also been employed (see e.g. Paag et al., 2001), in which the same draws are used across choice situations for a given consumer. For either of these *approximations*, the finite sample accuracy has not been explored thus far, a point that we address in our empirical work. From a theoretical perspective, we also confirm that under certain conditions, the cross-sectional approximation on panel data is consistent, and similarly explore the as yet undefined properties of the Paag et al. (2001) estimator.

Generalising to a framework with both intra-personal and inter-personal heterogeneity, as discussed by Bhat and Castelar (2002), Bhat and Sardesai (2006), and Hess and Rose (2009), creates an even greater computational burden due to the presence of multiple layers of integration. Several procedures have been suggested in this context that can reduce estimation time considerably, and have been used for example by Yáñez et al. (2011). The question that we address is whether these computational savings can be realised without undue loss of accuracy. This issue is important not only with respect to inter- and intra-personal variation but also for more general forms of heterogeneity, such as the specification of Cherchi et al. (2009), which combines inter-personal heterogeneity with two layers of intra-personal heterogeneity: across choices made on different days of the week, and across choices made in different weeks, and which uses an approximation in estimation. Our empirical work shows the shortcomings of these *approximations* in the recovery of the true patterns of heterogeneity, and we highlight the importance of further work to look into the asymptotic properties of estimators for such models.

For the purely cross-sectional experiments, we note substantial data requirements, stressing the advantages of repeated choice data. Across our remaining experiments, the findings highlight the importance of relying on the *correct* estimator. While the approximations to the simple inter-respondent specification are indeed consistent, the sample size requirements are more substantial than is commonly assumed, even with the simple utility functions used here. In addition, we note that these approximations do not in fact lead to any computational gains. In the case of joint inter-respondent and intra-respondent heterogeneity, the approximations do have clear computational advantages, but are unable to reproduce the correct patterns of heterogeneity, even at large sample sizes or when making use of a very high number of draws. Additionally, even for the *correct* estimator, the sample size requirements are substantial when aiming to retrieve both inter-respondent and intra-respondent heterogeneity.

The remainder of this paper is organised as follows. The following section gives an overview of the various specifications. This is followed in Section 3 by a discussion of the empirical framework used in the analysis. Results of the analysis are summarised in Section 4, while the conclusions are presented in Section 5.

2 Model specification and estimation

In this section, we look in detail at the specification of mixed logit models on cross-sectional data, on panel data with purely inter-personal heterogeneity, and on panel data with both intra- and inter-personal heterogeneity. In each case, we show the maximum simulated likelihood estimator method, while for the two types of panel data, we also discuss other estimation procedures that have been proposed.

2.1 Cross-sectional data

Estimation of a mixed logit model on cross-sectional data is relatively simple computationally. However, the estimates are often considerably less precise than with panel data. Correlations over multiple choices faced by a given consumer assist in identifying taste heterogeneity[1], and cross-sectional data do not provide information on these correlations since only one choice is observed for each consumer. The question that we address is whether taste heterogeneity can be accurately estimated on cross-sectional data, and what sample size is needed to attain an acceptable level of precision.

Notation is the following. We observe a sample of N consumers, indexed as $n = 1, \ldots, N$, where each consumer is observed to face only one choice situation. Let β_n be a vector of the true, but unobserved taste coefficients for consumer n. We assume that $\beta_n \, \forall n$ is *iid* over consumers with density $g(\beta \mid \Omega)$, where Ω is a vector of parameters of this distribution, such as the mean and variance. Let j_n be the alternative chosen by consumer n, such that $P_n(j_n \mid \beta)$ gives the probability of the observed choice for consumer n, conditional on β. The mixed logit probability of consumer n's chosen alternative is

$$P_n(j_n \mid \Omega) = \int_\beta P_n(j_n \mid \beta)\, g(\beta \mid \Omega)\, d\beta. \quad (1)$$

[1]As an example, if one consumer is observed always to choose the cheapest alternative and another consumer always chooses the most expensive alternative, then an inference that the price coefficient differs for these two consumers can be reasonably made.

The log-likelihood function is then given by:

$$\mathrm{LL}(\Omega) = \sum_{n=1}^{N} \ln \left(\int_{\beta} P_n (j_n \mid \beta)\, g(\beta \mid \Omega)\, \mathrm{d}\beta \right), \tag{2}$$

Since the integrals do not take a closed form, they are approximated by simulation. The simulated log-likelihood is:

$$\mathrm{SLL}(\Omega) = \sum_{n=1}^{N} \ln \left(\frac{1}{R} \sum_{r=1}^{R} P_n (j_n \mid \beta_{r,n}) \right). \tag{3}$$

where $\beta_{r,n}$ gives the r^{th} draw (out of R) from $g(\beta \mid \Omega)$ for consumer n. Different draws are used for the N consumers, for a total of NR draws.

2.2 Panel data with inter-personal variation only

Panel data allow us to utilise correlations over choice situations for a given consumer. As above, we observe a sample of N consumers, identified as n with $n = 1, \ldots, N$, but now consumer n faces T_n choice situations. In this section we allow tastes to vary over consumers but we assume that the tastes of each consumer are constant over choice situations. Consistent with this assumption, let β_n be a vector of the true, but unobserved taste coefficients for consumer n. We assume that β_n is *iid* over consumers with density $g(\beta \mid \Omega)$. Let $P_{n,t}(i \mid \beta)$ denote the logit probability that consumer n chooses alternative i in choice situation t, conditional on β. Now let $j_{n,t}$ be the alternative chosen by consumer n in choice situation t, such that $P_{n,t}(j_{n,t} \mid \beta)$ gives the logit probability of the observed choice for consumer n in choice situation t, conditional on β. The mixed logit probability of consumer n's observed *sequence* of choices (i.e., the choices in all the situations that the consumer faced) is

$$P_n(\Omega) = \int_{\beta} \prod_{t=1}^{T_n} P_{n,t}(j_n \mid \beta)\, g(\beta \mid \Omega)\, d\beta. \tag{4}$$

Note that, since the same tastes apply to all choices by a given consumer, the integration over the density of β applies to all the consumer's choices combined, rather than each one separately.

The log-likelihood function for the observed choices is then:

$$\mathrm{LL}(\Omega) = \sum_{n=1}^{N} \ln \left(\int_{\beta} \left[\prod_{t=1}^{T_n} \left(P_{n,t}(j_{n,t} \mid \beta) \right) \right] g(\beta \mid \Omega)\, \mathrm{d}\beta \right). \tag{5}$$

The simulated LL (SLL) is:

$$\mathrm{SLL}\left(\Omega\right)=\sum_{n=1}^{N}\ln\left(\frac{1}{R}\sum_{r=1}^{R}\left[\prod_{t=1}^{T_n}\left(P_{n,t}\left(j_{n,t}\mid\beta_{r,n}\right)\right)\right]\right). \tag{6}$$

Note that in this formulation, the product over choice situations is calculated for each draw; the product is averaged over draws; and *then* the log of the average is taken. The SLL is the sum over consumers of the log of the average (across draws) of products. The calculation of the contribution to the SLL function for consumer n involves the computation of RT_n logit probabilities.

Instead of utilising the panel nature of the data, the model could be estimated *as if* each choice were from a different consumer. That is, the panel data could be *treated as if* they were cross-sectional. The objective function is similar to Equation 2 except that the multiple choice situations by each consumer are represented as being for different consumers:

$$\mathrm{LL}\left(\Omega\right)=\sum_{n=1}^{N}\sum_{t=1}^{T_n}\ln\left(\int_{\beta}P_{n,t}\left(j_{n,t}\mid\beta\right)g\left(\beta\mid\Omega\right)\mathrm{d}\beta\right), \tag{7}$$

where the integration across the distribution of taste coefficients is applied to each choice, rather than to each consumer's sequence of choices. This function is simulated as:

$$\mathrm{SLL}\left(\Omega\right)=\sum_{n=1}^{N}\sum_{t=1}^{T_n}\ln\left(\frac{1}{R}\sum_{r=1}^{R}P_{n,t}\left(j_{n,t}\mid\beta_{r,t,n}\right)\right). \tag{8}$$

where $\beta_{r,t,n}$ is the r^{th} draw from $g(\beta\mid\Omega)$ for choice situation t for consumer n. Different draws are used for the T_n choice situations for consumer n, as well as for the N consumers. Consumer n's contribution to the SLL function utilises RT_n draws of β rather than R draws as in Equation 6, but involves the computation of the same number of logit probabilities as before, namely, RT_n. The difference is that the averaging across draws is performed before taking the product across choice situations.

If the parameters are identified by cross-sectional data (i.e, if the parameters could be estimated with only one choice situation per consumer), then the estimator based on this approach is consistent[2]. This consistency follows from the general theorem for consistency of extremum estimators (e.g. Ruud, 2000, Lemma 15.2, p. 322). Consider a statistic $s_n(\theta)$ that depends on parameters θ and

[2]An estimator is consistent if it converges on the true value when sample size rises without bound.

varies in the population. The general consistency theorem states: If $E_n(s_n(\theta))$ is uniquely maximised at the true value of θ, then, under standard regularity conditions, the estimator $argmax \sum_n s_n(\theta)$ on a random sample from the population is consistent. The assumption that $E_n(s_n(\theta))$ is *uniquely* maximised is the condition for identification, since otherwise different values of θ would attain the same maximum. Now consider our situation. Let $L_{nt}(\theta)$ be a person's log-likelihood value for one choice situation t. Suppose that $E_n(L_{nt}(\theta))$ is uniquely maximised at the true parameters for any t, such that the parameters are identified with only one choice per person and the maximum likelihood estimator based on one choice for a sample of consumers is consistent. Now consider the statistic that sums the log-likelihood of each choice over T choices for each person: $\sum_{t=1}^{T} L_{nt}(\theta)$. The expectation of this sum, $E_n(\sum_t L_{nt}(\theta)) = \sum_t E_n(L_{nt}(\theta))$, is also uniquely maximised at θ since each element in the sum is uniquely maximised at θ. The estimator defined by $argmax \sum_n \sum_t L_{nt}(\theta)$ on a sample of people is therefore consistent. Efficiency[3] is reduced, relative to the panel specification, because the correlation over observations by a given consumer is not utilised in the estimation criterion. The procedure is an example of a quasi-likelihood estimator[4] described by, e.g. Varin (2007), that in general are consistent with a loss of efficiency. The sandwich estimator is appropriate for its covariance matrix, since the sandwich formula does not simplify, as with full-information maximum likelihood on a correctly specified model, to the inverse of the hessian.

Another alternative (see e.g. Paag et al., 2001) is to utilise the cross-sectional formulation but, instead of taking different draws for each choice by a given consumer, to use the same draws in all the choice situations for the same consumer. The SLL under this approach is:

$$\mathrm{SLL}\left(\Omega\right) = \sum_{n=1}^{N} \sum_{t=1}^{T_n} \ln \left(\frac{1}{R} \sum_{r=1}^{R} P_{n,t} \left(j_{n,t} \mid \beta_{r,n} \right) \right). \tag{9}$$

The only difference in comparison with Equation 8 lies in dropping the additional subscript t from the draws of β, where the same set of R draws is now reused in the simulation of all T_n choices for consumer n, thus leading to a requirement for NR draws, identical to the maximum likelihood approach for panel data, and different from the $R \sum_{n=1}^{N} T_n$ draws for the cross-sectional estimation in Equation 8. This approach attempts to accommodate the panel nature of the data by reusing the same draws across choices for a given consumer. It is possible that the approach

[3] An estimator is efficient within a class (eg among consistent estimators) if its asymptotic sampling variance is lower than any other estimator within the class. One consistent estimator is more efficient than another if the former has lower asymptotic sampling variance.

[4] also called composite likelihood or pseudo-likelihood estimators

accommodates some correlation across replications for a given individual through using the same draws in simulation, and this may increase efficiency relative to the purely cross-sectional approach, at least for a small number of draws. However, asymptotically, as the number of draws rises as needed with sample size, this estimator is equivalent to the previous one, which treats each choice separately. Any efficiency that is perhaps gained with a small number of draws disappears as the number of draws rises. As a result, the estimator is consistent with a loss of asymptotic efficiency under the same conditions as the previous estimator. Limited evidence reported by Choudhury et al. (2009) seems to support this theoretical evaluation.

At this stage, it seems useful to briefly compare the three specifications discussed in this section. The specification in Equation 6 accommodates the within respondent homogeneity by performing the averaging across draws at the level of a sequence of choices for a given respondent. The probability of the entire sequence of choices is calculated for a single draw, thus recognising that tastes stay constant over choice situations. The averaging across draws recognises the uncertainty from the analyst's perspective in the specific values for β for the given respondent. In the cross-sectional specification in Equation 8, we do not utilise the fact that we have multiple observations for each respondent , estimating as if choices were made by different people. The same applies in Equation 9, with the only difference that the same draws are reused across observations for the same respondent. However, each probability is still simulated on its own.

2.3 Panel data with both inter- and intra-personal variation

We now generalise the specification on panel data to include intra-personal taste heterogeneity in addition to inter-personal heterogeneity.[5] Let $\beta_{n,t} = \alpha_n + \gamma_{n,t}$ where α_n is distributed across consumers but not over choice situations for a given consumer, and $\gamma_{n,t}$ is distributed over choice situations as well as consumers. That is, α_n captures inter-personal variation in tastes while $\gamma_{n,t}$ captures intra-personal variation. Their densities are denoted as $f(\alpha)$ and $h(\gamma)$, respectively,[6] where their dependence on underlying parameters, contained collectively in Ω, is suppressed for convenience.

[5] We focus on the simple case of *unstructured* additional heterogeneity across tasks for the same consumer (cf. Bhat and Castelar, 2002; Bhat and Sardesai, 2006; Hess and Rose, 2009). An example of a more *structured* approach is given by Cherchi et al. (2009).

[6] The mean of β_n is captured in α_n such that the mean of $\gamma_{n,t}$ is zero.

The LL function is given by:

$$LL(\Omega) = \sum_{n=1}^{N} \ln \left[\int_{\alpha} \left(\prod_{t=1}^{T_n} \left(\int_{\gamma} P_{n,t}(j_{n,t} \mid \alpha, \gamma) \, h(\gamma) \, \mathrm{d}\gamma \right) \right) f(\alpha) \, \mathrm{d}\alpha \right]. \quad (10)$$

The two levels of integration create two levels of simulation, which can be specified as:

$$SLL = \sum_{n=1}^{N} \ln \left[\frac{1}{R} \sum_{r=1}^{R} \left(\prod_{t=1}^{T_n} \frac{1}{K} \sum_{k=1}^{K} \left(P_{n,t}(j_{n,t} \mid \alpha_{r,n}, \gamma_{k,t,n}) \right) \right) \right]. \quad (11)$$

This simulation uses R draws of α for consumer n, along with $K\,T_n$ draws of γ. Note that, in this specification, the same draws of γ are used for all draws of α. That is, $\gamma_{k,t,n}$ does not have an additional subscript for r.[7] The total number of evaluations of a logit probability for consumer n is equal to $R\,K\,T_n$, compared to $R\,T_n$ when there is only inter-personal variation.

The computational cost of implementing this method with large K is very high, and thus far, it has not been implemented in any of the major packages. BIOGEME (Bierlaire, 2003) allows the user to estimate models combining inter-personal and intra-personal heterogeneity by using one draw of γ for each draw of α, where the intra-respondent nature of γ is recognised by using different draws for different choice situations. This approach is used for example by Yáñez et al. (2011). The SLL takes the form:

$$SLL = \sum_{n=1}^{N} \ln \left[\frac{1}{R} \sum_{r=1}^{R} \left(\prod_{t=1}^{T_n} P_{n,t}(j_{n,t} \mid \alpha_{r,n}, \gamma_{r,t,n}) \right) \right]. \quad (12)$$

The draws of γ are now subscipted by r (since a different draw of γ is used for each draw of α) and t (since different draws are taken across the t tasks), but are not subscripted by k (since only 1 draw is taken for each value of α.) This specification reduces the number of computations for consumer n from $R\,K\,T_n$ back to $R\,T_n$. The method differs from Equation 11 in two ways: by reducing K to 1 and by using different draws of γ for each draw of α. By using just one draw from γ for each draw from α, we now longer recognise the fact that for a given individual, i.e. a fixed value of α, we have variation in the values of γ. As a consequence, it is our expectation that this model will fail in retrieving the intra-respondent component of heterogeneity.

[7] It would be possible, in principle, to use different draws of γ for each draw of α, with $\gamma_{r,k,t,n}$ replacing $\gamma_{k,t,n}$. However, doing so creates an even greater computational burden by increasing the number of draws for respondent n from $R + K\,T_n$ to $R\,K\,T_n$.

An alternative approach is to simulate each choice probability separately but to use the same draws of α in all the choices by a given consumer. This approach is analogous to Equation 9 above but adapted for intra-personal variation in tastes. The objective function is

$$SLL = \sum_{n=1}^{N} \sum_{t=1}^{T_n} \ln \left[\frac{1}{R} \sum_{r=1}^{R} \left(P_{n,t} \left(j_{n,t} \mid \alpha_{r,n}, \gamma_{r,t,n} \right) \right) \right]. \tag{13}$$

This approach carries out all simulation at the level of individual choices, but the same draws of α are reused across choices for the same consumer. For γ, new draws are used in each choice situation. The use of the same draws of α across choices is intended to provide some identification of the intra-personal variation relative to inter-personal. However, as the number of draws rises, Equation 13 becomes simply a cross-sectional estimator in which the two forms of heterogeneity are not distinguished.

Given the high computational cost of estimating models based on Equation 11 with large K, the use of equations 12 and 13 could potentially lead to significant savings. However, the accuracy of these alternatives is unknown. This issue is set to become even more important as new model structures are developed that allow for increasingly complex patterns of heterogeneity. For example, the model developed by Cherchi et al. (2009) allows such a complex patterns of heterogeneity that the authors utilise an approach to estimation in which all averaging is performed at the level of individual consumers, drawing parallels with Equation 13.

In the present paper, we do not attempt to derive the asymptotic properties of the three estimators described in this section, but this remains a crucial area for future work. We do however postulate that both approximations will struggle in differentiating between the two types of heterogeneity. Equation 12 will face difficulties in retrieving intra-respondent heterogeneity, while Equation 13 will have issues with inter-respondent heterogeneity.

As in Section 2.2, we now briefly compare the three specifications discussed in this Section. The simulated log-likelihood in Equation 11 recognises that part of the taste heterogeneity is across (rather than within) respondents by calculating the probability for the entire sequence of choices for a single value of α before averaging across draws. However, to accommodate the additional heterogeneity within the sequence of choices, the probability for a single choice conditional on a given draw from α is itself obtained by simulating across the distribution for γ. In other words, while the averaging across draws from α takes place at the level of an individual respondent, the averaging across draws from γ takes place at the level of a single observation. On the other hand, in Equation 12, we do not simulate across the distribution of γ conditional on a given value for α, but simulate jointly across the distributions for α and γ, albeit with associating different draws from γ with the draws from α for different observations. Finally, in Equation 13, we again simulate jointly across the distributions for α and γ, but do it separately for each choice task, while reusing the same draws from α across the observations for a given individual.

3 Empirical framework

To test the ability of different methods to capture the true heterogeneity in a dataset, we constructed a variety of true data generation processes, simulated datasets under these situations, and applied the estimation methods to the datasets. Each choice situation consists of two alternatives with two attributes for each alternative, namely travel time (in minutes) and travel cost (in £). The underlying data comes from an experimental design with 50 rows, with 5 blocks of 10 choices each. On the basis of this design, we simulated datasets with up to $5,000$ choice situations. For the cross-sectional datasets, we assigned a single choice situation to each consumer. For the panel datasets, we assigned an entire block of 10 choice situations to each consumer.

3.1 Case studies

We define four different "case studies" that incorporate different *types* of heterogeneity in the sensitivities to travel time and travel cost. Each of these case studies includes several "versions" that differ in the *degree* of its type of heterogeneity. Sample size is specified to range from 100 to 5,000 choice situations (using 12 different sample sizes). For each sample size of each versions of each case study, ten different datasets were generated. Estimation was conducted on each dataset by several relevant methods. The average over the ten datasets of the comparison between estimates and true parameters provides information on the bias, if any, in the estimator. The root means squared error between the estimates and the true parameters provides information on the efficiency of each estimator.

We will now describe the four case studies, with an overview given in Table 1. We use the notation β_{T} for the time coefficient an β_{C} for the cost coefficient, with the specification of these coefficients differing over case studies. In all case studies, the data generation and estimation include a constant for the cheaper of the two alternatives, with a *true* value of 1.

3.1.1 Case Study 1

This case study specifies a single random coefficient, namely the time coefficient β_{T}, with the travel cost coefficient β_{C} being fixed to a value of -1. The time coefficient is specified to have a mean of 0.2, which is labelled μ. Three different versions are specified that differ in the standard deviation of the travel cost coefficient (i.e., the degree of heterogeneity). The standard deviation, labelled σ, is set to 0.05, 0.1, and 0.2 in the three versions, respectively, which implies that the coefficient of variation ($cv = \sigma/\mu$) is 0.25, 0.5, and 1, respectively. For this case study, two different types of data were produced. Cross-sectional datasets were generated with only one choice situation per consumer, while panel datasets were generated with ten choice situations per consumer.

3.1.2 Case Study 2

The second case study specifies heterogeneity in both the time and cost coefficients, with three different levels of heterogeneity for each coefficient (*cv* of 0.25, 0.5, and 1), giving rise to nine possible combinations. In addition, two levels of correlation between the time and cost coefficients are considered, namely, no correlation and a correlation of -0.5, giving rise to a total of 18 versions of this case study. Cross-sectional and panel datasets are created in the same way as for case study 1.

3.1.3 Case Study 3

The third case study incorporates heterogeneity only in the time coefficient, but specifies two layers of heterogeneity: inter- and intra-personal variation. The time coefficient consists of (i) a component that varies over consumers only, with a standard deviation of 0.05, 0.1 or 0.2 (*cv* of 0.25, 0.5, or 1) and (ii) a component that varies of over choice situations for each consumer, with a standard deviation of 0.025, 0.05, or 0.1 (*cv* of 0.125, 0.25, or 0.75). Combining each of the levels gives nine versions of this case study. Since the two layers of variation cannot be expected to be identified in cross-sectional data, only panel datasets are created for this case study.

3.1.4 Group 4 case study

The final case study incorporates inter- and intra-personal heterogeneity for both the time and cost coefficients. Given the high computational cost of estimating models under this specification, only a single version was specified. For this one version, the coefficient of inter-personal variation is set at 0.5 for the time coefficient and 0.625 for the cost coefficient, and the coefficient of intra-personal variation is set at 0.25 for the time coefficient and 0.375 for the cost coefficient. Also, we specify a correlation of -0.5 between the two coefficients at the level of inter-personal variation, with no correlation between the intra-personal components. As with case study 3, only panel datasets are generated.

3.2 Estimation

Halton draws (Halton, 1960) were used for the simulation that is required in estimation. After extensive pre-testing, we utilised $R = 200$ draws [8]. All models were coded and estimated in Ox 4.1 (Doornik, 2001).

Different estimation procedures were used in the different case studies. The lettering in the following chart (ie. A, B, etc.) is used to refer to the methods when we report results in the next section. We gently suggest that the reader retain this chart when reading the remainder of the paper.

- Case studies 1 and 2:
  - On the panel datasets, we applied the three approaches given in Section 2.2, namely:
    - (A) equation 6, which is the standard maximum simulated likelihood approach for panel data with inter-personal heterogeneity,
    - (B) equation 8, which treats the data as if they were cross-sectional, and
    - (C) equation 9, which is like B except that the same draws are used for all the choices of a consumer.
  - On the cross-sectional datasets, we applied:
    - (D) equation 3, which is the standard simulated likelihood function for cross-sectional data.
- In case studies 3 and 4, we applied the three procedures in Section 2.3, namely:
  - (E) equation 11, which uses K draws of the intra-personal component (γ) along with R draws of the inter-personal component (α), where $K = R$, and with each draw from γ being used with each draw from α,

[8] The number of draws was increased up to $10,000$ draws, but no changes in results were observed beyond about 100 draws. To keep estimation times manageable in the face of the large number of models ($16,080$ in total), we settled on $R = 200$.

– (F) equation 12, which utilises one draw of the intra-personal component for each of the R draws of the inter-personal component, and
– (G) equation 13, which is similar to C in that it treats the data as if they were cross-sectional but uses the same draws for the inter-personal component across all the choices of a consumer.

The estimation was performed on each of ten datasets, for each version of each case study, and for each sample size – giving a total of $16,080$ models that were estimated. The next section discusses the results of these estimations.

4 Results

Given the number of estimations, it is clearly impossible to present detailed results for each. We have taken several steps to reduce the informational burden and yet meaningfully represent the findings. In particular, we focus on the coefficient of variation cv, which measures the degree of heterogeneity, and, when applicable, the correlation between the time and cost coefficients. In any one estimation, the estimates are compared to the true value in the simulated dataset[9]. The mean error (ME, where the error is the estimate minus the true value) over the ten datasets is calculated, as well as the root mean squared error (RMSE). The ME provides an indication of bias, and the RMSE provides a measure of the standard deviation of the estimates around the true value, which, in the absence of bias, is a measure of efficiency. We also report the mean adjusted ρ^2 and the mean estimation time across the ten runs. Finally, early experience showed problems with convergence for some of the methods especially on small samples, and so we also report the number of out the ten runs that converged.

Even with the above approach to reporting results, we still have 30 versions across the first three case studies, with 12 different sample sizes and several estimation methods. As an additional way of producing a more concise overview of our findings, we present (i) results for *all* sample sizes for only *one* version of each case study, and (ii) results at *one* sample size (the largest) for *all* versions of each case study.

[9]Since the datasets are samples taken from underlying distributions, the mean and variance of the random coefficients in the dataset differ from that of the underlying distribution from which the sample is drawn, especially with small sample sizes. For comparison purposes, we use the former as the "true" value to avoid this additional sampling noise.

4.1 Case study 1

The findings for case study 1 are presented in two parts: Table 2 presents the detailed findings for version 2, which uses a true coefficient of variation of 0.5, at all sample sizes; while Table 3 reports the results of all versions of this case study at a sample size of $5,000$ observations.

Recall that both cross-sectional and panel datasets were generated for case study 1. In the cross-sectional datasets, the number of observations is the number of sampled consumers, N. In the panel datasets, each consumer faces 10 choice situations. The number of observations (as defined for the purposed of the tables) is the number of choice situations, which in our panel datasets is $10N$. Therefore, when the tables lists the number of observations as, e.g., 100, the number of sampled consumers in the panel datasets is 10 and in the cross-sectional datasets is 100.

Consider Table 2 first. Each part of the table provides results for a different statistic, e.g, ME of the estimated cv in the top-left part, and RMSE in the top-right. In each part, there are columns that correspond to the estimation methods enumerated in Section 3.2 above. In particular, the first column, labelled A, gives results for the maximum simulated likelihood estimator on the panel datasets. The second and third columns, labelled B and C, give results for the two simplifications on panel datasets. The last column, labelled D, gives the results of maximum simulated likelihood on the cross-sectional datasets. The dotted line before the last column is intended to reinforce the distinction that the estimates for the last column are obtained on different datasets (the cross-sectional datasets) than the estimates for the other columns (the panel datasets).

We observe essentially no bias in estimating the coefficient of variation using any of the three estimation procedures on the panel data (A, B, and C) when using the largest sample (the last row). A closer inspection of the results across the 12 different sample sizes however indicates that methods B and C require larger samples than method A in order to achieve relatively low levels of bias. The greater efficiency of A is evidenced in the RMSE results, which show much lower variation across the ten runs when using A as opposed to B or C. The only exception to this relation is with a sample size of 100, which would equate to only 10 consumers.

The model fit is uniformly better with A than B or C. Convergence problems are observed for B and C at the two smallest sample sizes, which do not occur with A, and A obtains slightly shorter estimation times than B or C. These results regarding convergence rates and estimation times contradict the occasionally held view that the more complex derivatives of the likelihood for A relative to B and C can hamper estimation.

Our findings suggest that method B, which is consistent (when identified), provides fairly unbiased estimates except in small samples. However, its loss in efficiency is non-trivial, even with large samples: the RMSE is 5 times greater with B than A at the largest sample size. Interestingly, method C produces exactly the same results as B, confirming the conjecture (discussed above) that the two methods should be similar when estimated with a sufficiently large number of draws, and showing that any attempt with C to regain some of the efficiency lost when moving from A to B is not particularly effective.

It is also of interest to look at the results for the cross-sectional datasets, using method D, which is the maximum likelihood estimator for these data. We observe that the ME is fairly large even with large samples, and that the RMSE is larger for all sample sizes than with maximum likelihood on panel data (method A). These results re-enforce the discussion above that taste heterogeneity is more difficult to identify on cross-sectional data than panel data.

These findings are confirmed by the results for all versions of the case study, given in Table 3. All three estimation methods on the panel data produce little bias[10], with method A obtaining the lowest RMSE. There is again essentially no difference between B and C, and the estimation on cross-sectional data (method D) again results in a fairly large ME and RMSE. There is also evidence that the difficulties with B and C (as well as D) are more accentuated when working with higher levels of heterogeneity.

4.2 Case study 2

We now turn our attention to the results of the second case study, which incorporates heterogeneity for both coefficients. Table 4 presents results for version 5 of this case study, in which the coefficients are not correlated and each has a coefficient of variation of 0.5. The corresponding results with correlated coefficients, which is version 14, are shown in Table 5.

Looking first at the case with uncorrelated coefficients, we see more fluctuation in ME with increasing sample size than was the case when working with a single random coefficients. Nevertheless, method A clearly exhibits less bias than methods B and C, or than D on cross-sectional data. Interestingly, D (on the cross-sectional data) obtains lower ME than A (on panel data) for the cost heterogeneity but higher ME for the time heterogeneity; and D exhibits greater variation across runs for both cost and time heterogeneity. Method A maintains its estimation-time and model-fit advantages. In addition, we now observe more

[10] The slightly larger ME for method A on version 3 is negligible when looking at the RMSE findings.

problems with convergence for the remaining methods than was the case with a single random coefficient.

These findings are also supported by the results for all nine uncorrelated versions reported in Table 6. We observe non-negligible bias for B and C in some settings, especially for β_{T}, with similar problems for D. There are also clear advantages for method A in terms of stability across runs, as well as model fit and estimation time. The differences across methods are once again especially noticeable in those cases where we have high *true* levels of heterogeneity. In both Table 4 and Table 6, we observe a small estimation time advantage for C over B, which can possibly be linked to the use of a smaller set of draws ($N\,R$ rather than $N\,T\,R$), requiring less initial setup time.

Turning our attention to the case with correlation between the cost and time coefficients, the results (Table 5) for version 14 show, as above, less bias and greater efficiency for A than B and C. However, a number of additional observations can be made. All three models show problems with retrieving the true level of correlation, although this is less severe for A. Also, while C performs well for the heterogeneity in the time coefficient with the full sample size, this value seems to be an outlier (when compared to other sample sizes), and performance especially for the cost coefficient is in fact inferior to B. The problems with convergence also increase in severity for B and C, and we again observe the above-mentioned differences in model fit and estimation times. Method D on cross-sectional data seems to perform better than method B on the panel data, but problems with recovering the true patterns of heterogeneity remain.

Table 7 gives results for all versions with correlated coefficients and the largest sample size. The results evidence superior performance by method A, especially with high levels of heterogeneity. Also, while there are cases where C performs better than B for β_{T}, this is usually accompanied by greater error for β_{C}, and there is thus no conclusive evidence that C produces less bias or attains a greater efficiency than B.

4.3 Case study 3

Table 8 presents the detailed results for version 5 of the third case study, where we now have both inter-personal ($cv = 0.5$) and intra-personal ($cv = 0.25$) heterogeneity in the time coefficient. Methods E and G evidence little bias in either type of heterogeneity. In contrast, method F exhibits considerable bias in the intra-personal heterogeneity while remaining fairly unbiased for inter-personal heterogeneity. The RMSE's are lowest for method E. For method F, they are only slightly higher for the inter-personal heterogeneity, but much higher for intra-personal heterogeneity. For method G, the opposite is the case.

We observe only small problems with model convergence at small sample sizes. In terms of fit, methods E and F obtain higher values than method G. In terms of estimation times, the main observation however relates to the massively higher estimation times for method E, which takes over 200 times as long as the others, which is a direct result of the large number of logit calculations.

As discussed in section 2.3, method E requires RKT logit calculations for consumer n, while methods F and G require RT_n logit calculations. The questions arises: does the superior performance of E arise simply because of a larger number of logit calculations (i.e., effectively more draws in simulation), or because of the difference in the way the logit probabilities are combined in the formula for E relative to the formulas for F and G? To investigate this issue, we re-applied methods F and G with $R = 40,000$, such that they utilise the same number of logit calculations as E (which, as stated above, uses $R = 200$ and $K = 200$). The results are given in Table 9. As the table indicates, method F performs essentially the same with $40,000$ draws as with 200 draws[11]. In particular, it continues to estimate essentially no intra-personal variance, even with 40,000 draws. This result suggests that the problem arises from the formula for F rather than the number of draws: by using only one draw of intra-personal heterogeneity for each draw of inter-personal heterogeneity, the estimator seems not able to distinguish intra-personal heterogeneity. Method G performs worse with $40,000$ draws than 200. Recall that the formula for G treats the choices as if they were cross-sectional, but uses the same draws of inter-personal variation in all the choices for a given consumer. We conjectured that this use of common draws creates correlation over choices that serves to identify the two types of heterogeneity, but that the differentiation diminishes as the number of draws rises, since the exact (unsimulated) probabilities are independent over choices in this formula. The results in Table 9 are consistent with this conjecture and show, the same as for method F, that the degraded performance of G relative to E is due to G's formulation rather than the number of draws.

We return now to our original implementation of F and G with $R = 200$. The findings for all versions of case study 3 are given in Table 10. The results indicate that method E performs well in all versions. Method G performs well in version 4 but not the other versions. And method F seems, as noted above, capable of estimating inter-personal, but not intra-personal heterogeneity, although some problems for inter-personal heterogeneity are also observed in versions 8 and 9. In summary, method F performs well for inter-personal heterogeneity, while method G performs well for intra-personal heterogeneity. Only method E performs well

[11] The performance improves in the 4^{th} decimal place for inter-respondent heterogeneity, and gets worse in the 4^{th} decimal for intra-respondent heterogeneity.

for both types of heterogeneity.

4.4 Case study 4

The results for the single case study in group 4 are shown in Table 11 and Table 12, based on data generated with inter- and intra-personal heterogeneity in both the time and cost coefficients, with a correlation of -0.5 between the coefficients for the inter-personal component. Method E performs well for both types of heterogeneity for both coefficients, as well as in the correlations. This finding is useful, since it shows that fairly complex patterns of heterogeneity can be captured by this method, albeit at a high computational cost. Method F, as before, performs well for inter-personal variation but not for inter-personal variation. And Method G performs relatively poorly in all regards. Apparently, its good performance in version 5 of case study 3 was an aberration, since the accuracy in that instance is not evidenced in other versions of that case study nor in the current case study. In addition, as shown Table 12, in method G encountered convergence problems in a large number of datasets.

The overall conclusion seems to be that method E is clearly superior to the others if computer time is not a binding constraint on the researcher. If shorter run times are necessary, then it might be tempting to use method F, as it seems preferable to G when considering bias, efficiency and convergence combined. However, the shorter run times with method F come at a substantial lost of accuracy relative to method E, so such a shortcut is not advisable.

5 Conclusions

This paper has examined the issue of estimating the *true* patterns of heterogeneity across consumers as well as across choice situations for a given consumer. This topic is of great interest given the growing reliance on mixed logit models in transportation and other fields. Despite some results in terms of theoretical properties of the different models, little is known about their ability to retrieve the *true* patterns of heterogeneity in the data. Additionally, with increasing interest in structures allowing for complex patterns of heterogeneity, such as joint inter-respondent and intra-respondent variation, there is a temptation to use computationally attractive approximations, such as in Cherchi et al. (2009) and Yáñez et al. (2011), or indeed in Paag et al. (2001) for the case of inter-respondent heterogeneity only. Worryingly, the qualities of these approximations are unknown, and the results from the present paper should serve to stop their use before they become more widely used. Alongside discussing the theoretical properties of a number of estimators, the main aim of this paper was to compare the performance of different specifications in a large scale simulated data study.

With the wealth of results presented across case studies, a summary of the observations seems appropriate. Firstly, retrieving the *true* patterns of heterogeneity is found to be considerably less precise on cross-sectional data than on panel data, even with large sample sizes, which confirms the value of having multiple choices per consumer. When such repeated choices are available, a distinction needs to be made between cases with inter-personal variation only, and cases with additional intra-respondent heterogeneity. With inter-personal variation only, maximum simulated likelihood (method A) performs well, as expected. However, we found fairly large RMSE and bias in some cases with small samples sizes, which points to the value of larger samples – larger perhaps than are typically used for mixed logit estimation. It seems that treating the panel data as if they were cross-sectional (method B) results in fairly accurate estimates provided that the sample size is sufficiently high (and that the parameters are identified by cross-sectional data, as in our study). The performance depended on the presence of one or two random coefficients as well as the actual degree of heterogeneity. There is of course a loss of efficiency, which we found to be as large as a factor of four or five. Treating the panel data as if they were cross-sectional but simulating with common draws for each consumer (method C) performs the same as method B, as we had conjectured: the added complication of using of common draws has no meaningful effect, given that the data are treated as cross-sectional. With intra- and inter-personal variation, maximum likelihood with extensive simulation at each level (method E) performs well, as expected, but run times are very high. While appealing from a computational perspective, the two simplifications that have been proposed (methods F and G) do not reach anywhere near the level of accuracy, especially for the intra-personal heterogeneity.

The three main findings from the paper can be summarised succinctly as follows. Firstly, the data requirements of mixed logit models are substantial, and the small sample issues are arguably more important than is commonly assumed. Secondly, the availability of multiple observations per respondents greatly facilitates the study of taste heterogeneity. Thirdly, to guarantee recovery of the true patterns of heterogeneity, analysts should make use of the correct specification of the (simulated) log-likelihood function and avoid any shortcuts, however computationally attractive they may me.

Acknowledgements

This work described in this paper was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the project "Estimation of travel demand models from panel data", grant EP/G033609/1. The first author also acknowledges the financial support by the Leverhulme Trust, in the form of a *Leverhulme Early Career Fellowship*. The authors would like to thank Andrew Daly, Stefan Mabit and Mogens Fosgerau for comments on an earlier version of this paper.

References

Bhat, C. R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial Logit model. Transportation Research Part B 35 (7), 677–693.

Bhat, C. R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. Transportation Research Part B 37 (9), 837–855.

Bhat, C. R., Castelar, S., 2002. A unified mixed logit framework for modeling revealed and stated preferences: formulation and application to congestion pricing analysis in the San Francisco Bay area. Transportation Research Part B 36 (7), 593–616.

Bhat, C. R., Sardesai, R., 2006. The impact of stop-making and travel time reliability on commute mode choice. Transportation Research Part B 40 (9), 709–730.

Bierlaire, M., 2003. BIOGEME: a free package for the estimation of discrete choice models. Proceedings of the 3^{rd} Swiss Transport Research Conference, Monte Verità, Ascona.

Cherchi, E., Cirillo, C., Ortúzar, J. de D., 2009. A mixed logit mode choice model for panel data: accounting for different correlation over time periods. paper presented at the International Choice Modelling Conference, Harrogate.

Cherchi, E., Guevara, C. A., 2009. A monte carlo experiment to analyse the curse of dimensionality in estimating random coefficients models with a full variance-covariance matrix. paper presented at the 12^{th} International Conference on Travel Behaviour Research, Jaipur, India.

Cherchi, E., Ortúzar, J. de D., 2010. Can mixed logit reveal the actual data generating process? some implications for environmental assessment. Transportation Research Part D 15 (7), 428–442.

Choudhury, C., Sivakumar, A., Rohr, C., Burge, P., Daly, A., 2009. Dealing with repeated choices in stated preference data: an empirical analysis. paper presented at the European Transport Conference, Noordwijkerhout, The Netherlands.

Doornik, J. A., 2001. Ox: An Object-Oriented Matrix Language. Timberlake Consultants Press, London.

Halton, J., 1960. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. Numerische Mathematik 2 (1), 84–90.

Hess, S., Rose, J. M., 2009. Allowing for intra-respondent variations in coefficients estimated on repeated choice data. Transportation Research Part B 43 (6), 708–719.

Hess, S., Train, K., Polak, J. W., 2006. On the use of a Modified Latin Hypercube Sampling (MLHS) method in the estimation of a Mixed Logit model for vehicle choice. Transportation Research Part B 40 (2), 147–163.

McFadden, D., Train, K., 2000. Mixed MNL Models for discrete response. Journal of Applied Econometrics 15 (5), 447–470.

Munizaga, M. A., Alvarez-Daziano, R., 2005. Testing mixed logit and probit models by simulation. Transportation Research Record 1921, 53–62.

Paag, H., Daly, A., Rohr, C., 2001. Predicting the use of the copenhagen harbour tunnel. In: Hensher, D. A. (Ed.), Travel Behaviour Research: the Leading Edge. Pergamon, Oxford, pp. 627–646.

Revelt, D., Train, K., 1998. Mixed Logit with repeated choices: households' choices of appliance efficiency level. Review of Economics and Statistics 80 (4), 647–657.

Rose, J. M., Hess, S., Collins, A. C., 2011. Impact of violations of model assumptions: the role of lexicography, inconsistent choices and changing preferences in model estimation. paper presented at the 90^{th} Annual Meeting of the Transportation Research Board, Washington, D.C.

Ruud, P. A., 2000. An Introduction to Classical Econometric Theory. Oxford University Press, New York.

Train, K., 2009. Discrete Choice Methods with Simulation, second edition Edition. Cambridge University Press, Cambridge, MA.

Varin, C., 2007. On composite marginal likelihoods. AStA Advances in Statistical Analysis 92 (1), 1–28.

Yáñez, M. F., Cherchi, E., Heydecker, B., Ortúzar, J. de D., 2011. On the Treatment of Repeated Observations in Panel Data: Efficiency of Mixed Logit Parameter Estimates. Networks and Spatial Economics, forthcoming.

Table 1: Settings used for coefficients in data generation

Case Study 1 (10 replications of simulated data for each version)

Version	β_{T} μ	σ	cv	β_{C}
1	-0.2	0.05	0.25	-1
2	-0.2	0.1	0.5	-1
3	-0.2	0.2	1	-1

Case Study 2 (10 replications of simulated data for each version)

	β_{T}			β_{C}			$corr.\,(\beta_{\mathrm{T}}, \beta_{\mathrm{C}})$	
Version	μ	σ	cv	μ	σ	cv	$1-9$	$10-18$
1&10	-0.2	0.05	0.25	-1	0.25	0.25	0	-0.5
2&11	-0.2	0.1	0.5	-1	0.25	0.25	0	-0.5
3&12	-0.2	0.2	1	-1	0.25	0.25	0	-0.5
4&13	-0.2	0.05	0.25	-1	0.5	0.5	0	-0.5
5&14	-0.2	0.1	0.5	-1	0.5	0.5	0	-0.5
6&15	-0.2	0.2	1	-1	0.5	0.5	0	-0.5
7&16	-0.2	0.05	0.25	-1	1	1	0	-0.5
8&17	-0.2	0.1	0.5	-1	1	1	0	-0.5
9&18	-0.2	0.2	1	-1	1	1	0	-0.5

Case Study 3 (10 replications of simulated data for each version)

	β_{T}					
Version	μ	σ	γ	cv (inter)	cv (intra)	β_{C}
1	-0.2	0.05	0.025	0.25	0.125	-1
2	-0.2	0.1	0.025	0.5	0.125	-1
3	-0.2	0.2	0.025	1	0.125	-1
4	-0.2	0.05	0.05	0.25	0.25	-1
5	-0.2	0.1	0.05	0.5	0.25	-1
6	-0.2	0.2	0.05	1	0.25	-1
7	-0.2	0.05	0.15	0.25	0.75	-1
8	-0.2	0.1	0.15	0.5	0.75	-1
9	-0.2	0.2	0.15	1	0.75	-1

Case Study 4 (10 replications of simulated data)

β_{T}				
μ	σ	γ	cv (inter)	cv (intra)
-0.2	0.1	0.05	0.5	0.25

$corr.\,(\beta_{\mathrm{T}}, \beta_{\mathrm{C}})$	
inter	intra
-0.5	0

β_{C}				
μ	σ	γ	cv (inter)	cv (intra)
-1	0.625	0.375	0.625	0.375

Table 2: Detailed estimation results for version 2 of case study 1

	ME(cv)				RMSE(cv)			
Obs.	A	B	C	D	A	B	C	D
100	0.44	-0.34	-0.27	-0.07	0.85	0.44	0.45	0.37
200	0.17	-0.15	-0.15	-0.10	0.28	0.41	0.42	0.32
300	0.10	-0.14	-0.15	-0.17	0.18	0.35	0.35	0.31
400	0.08	-0.09	-0.08	-0.16	0.18	0.43	0.44	0.29
500	0.05	-0.07	-0.07	-0.21	0.15	0.48	0.47	0.36
750	-0.01	0.08	0.08	-0.03	0.09	0.22	0.23	0.24
1,000	0.01	0.16	0.16	0.05	0.07	0.28	0.28	0.17
1,500	0.00	0.09	0.09	-0.10	0.05	0.16	0.16	0.21
2,000	0.01	0.12	0.12	-0.12	0.05	0.15	0.15	0.18
3,000	0.02	0.10	0.09	-0.06	0.03	0.15	0.15	0.13
4,000	0.01	0.06	0.06	-0.06	0.02	0.14	0.13	0.11
5,000	0.00	0.01	0.01	-0.05	0.02	0.10	0.10	0.09

	Runs converged				Mean adj. ρ^2			
Obs.	A	B	C	D	A	B	C	D
100	10	8	8	7	0.32	0.26	0.27	0.30
200	10	9	9	8	0.31	0.23	0.23	0.27
300	10	10	10	10	0.33	0.27	0.27	0.30
400	10	10	10	10	0.34	0.28	0.28	0.30
500	10	10	10	10	0.35	0.28	0.28	0.30
750	10	10	10	10	0.36	0.30	0.30	0.31
1,000	10	10	10	10	0.35	0.30	0.30	0.31
1,500	10	10	10	10	0.36	0.30	0.30	0.32
2,000	10	10	10	10	0.36	0.30	0.30	0.31
3,000	10	10	10	10	0.36	0.30	0.30	0.31
4,000	10	10	10	10	0.36	0.31	0.31	0.31
5,000	10	10	10	10	0.37	0.31	0.31	0.31

	Mean est. time (s)			
Obs.	A	B	C	D
100	1	1	1	1
200	1	2	2	2
300	2	3	2	3
400	3	4	4	4
500	3	5	5	4
750	5	7	8	7
1,000	6	10	9	9
1,500	10	14	14	15
2,000	13	19	19	20
3,000	20	30	29	31
4,000	27	39	38	40
5,000	34	48	47	50

Table 3: Summary estimation results for all versions of case study 1, with $5,000$ observations each

	ME(cv)				RMSE(cv)			
Version	A	B	C	D	A	B	C	D
1	0.00	-0.03	-0.03	-0.03	0.02	0.11	0.11	0.08
2	0.00	0.01	0.01	-0.05	0.02	0.10	0.10	0.09
3	-0.02	0.01	0.01	-0.09	0.04	0.21	0.21	0.14

	Mean adj. ρ^2				Mean est. time (s)			
Version	A	B	C	D	A	B	C	D
1	0.36	0.35	0.35	0.35	32	46	43	45
2	0.37	0.31	0.31	0.31	34	48	47	50
3	0.42	0.25	0.25	0.25	37	51	49	52

	Runs converged			
Version	A	B	C	D
1	10	10	10	10
2	10	10	10	10
3	10	10	10	10

Table 4: Detailed estimation results for version 5 of case study 2

	ME(cv_T)				ME(cv_C)			
Obs.	A	B	C	D	A	B	C	D
100	0.10	-0.51	-0.44	-0.15	0.62	0.34	0.08	0.11
200	0.16	0.03	0.28	0.00	0.15	-0.02	0.04	0.17
300	0.01	-0.42	-0.43	0.08	0.11	0.09	0.00	0.15
400	-0.02	-0.16	-0.12	-0.14	0.06	-0.02	0.02	0.08
500	0.00	-0.32	-0.17	-0.33	0.07	-0.14	-0.04	0.05
750	-0.04	-0.21	-0.25	-0.21	0.02	-0.10	-0.11	0.01
1,000	-0.03	0.00	-0.03	0.00	0.03	-0.14	-0.14	0.06
1,500	-0.02	0.09	0.08	-0.22	0.07	-0.08	-0.07	-0.02
2,000	-0.01	0.09	0.11	-0.28	0.04	-0.04	0.02	-0.05
3,000	0.02	0.14	0.15	-0.16	0.02	0.05	0.07	-0.06
4,000	0.01	0.04	0.05	-0.09	0.02	0.04	0.05	0.00
5,000	0.00	-0.05	-0.04	-0.08	0.02	-0.11	0.00	0.01

	RMSE(cv_T)				RMSE(cv_C)			
Obs.	A	B	C	D	A	B	C	D
100	0.57	0.51	0.45	0.26	0.91	0.36	0.18	0.20
200	0.31	0.90	1.33	0.35	0.26	0.33	0.38	0.32
300	0.18	0.42	0.43	0.57	0.20	0.39	0.33	0.27
400	0.19	0.85	0.83	0.20	0.15	0.33	0.34	0.24
500	0.15	0.47	0.44	0.34	0.14	0.37	0.32	0.18
750	0.11	0.33	0.30	0.25	0.09	0.28	0.23	0.14
1,000	0.08	0.21	0.24	0.19	0.06	0.27	0.27	0.12
1,500	0.06	0.19	0.17	0.28	0.08	0.27	0.22	0.17
2,000	0.05	0.14	0.18	0.30	0.06	0.25	0.14	0.18
3,000	0.05	0.19	0.20	0.23	0.04	0.11	0.10	0.17
4,000	0.04	0.16	0.15	0.14	0.04	0.08	0.07	0.09
5,000	0.04	0.17	0.15	0.11	0.04	0.23	0.06	0.07

	Runs converged				Mean adj. ρ^2				Mean est. time (s)			
Obs.	A	B	C	D	A	B	C	D	A	B	C	D
100	10	2	2	3	0.32	0.20	0.14	0.29	1	2	2	2
200	10	9	9	8	0.30	0.13	0.13	0.25	2	3	3	5
300	10	8	7	6	0.29	0.14	0.14	0.26	3	5	4	6
400	10	9	10	8	0.31	0.18	0.18	0.26	4	7	7	7
500	10	9	10	9	0.31	0.18	0.18	0.27	6	7	9	8
750	10	10	9	9	0.31	0.20	0.20	0.29	8	13	12	14
1,000	10	9	9	10	0.31	0.19	0.19	0.28	10	20	17	20
1,500	10	10	9	10	0.32	0.19	0.19	0.29	16	28	27	24
2,000	10	10	10	10	0.33	0.20	0.20	0.29	22	37	38	34
3,000	10	10	10	10	0.33	0.20	0.20	0.28	33	65	57	54
4,000	10	10	10	10	0.33	0.21	0.21	0.29	41	92	67	69
5,000	10	10	10	10	0.33	0.22	0.22	0.29	49	92	82	92

Table 5: Detailed estimation results version 14 of case study 2

	ME(cv_{T})				ME(cv_{C})				ME($corr.$)			
Obs.	A	B	C	D	A	B	C	D	A	B	C	D
100	0.06	0.18	-0.23	-0.40	0.55	0.18	-0.53	-0.22	1.45	1.25	0.84	0.95
200	0.16	0.60	1.64	0.02	0.12	0.17	0.11	0.15	1.59	1.20	1.22	1.04
300	-0.04	8.52	-0.12	0.03	0.08	0.03	0.25	-0.17	0.24	0.75	0.82	0.89
400	0.06	3.21	17.36	-0.12	0.05	0.16	0.10	-0.13	0.93	0.43	0.77	0.75
500	0.04	-0.16	0.09	-0.21	0.05	0.04	0.03	-0.05	0.03	0.44	-0.30	0.35
750	-0.05	-0.08	-0.12	-0.15	-0.01	-0.17	-0.19	-0.19	-0.24	0.31	-0.20	1.05
1,000	-0.05	0.06	0.12	0.02	-0.01	-0.11	-0.17	-0.07	-0.15	0.70	-0.18	1.27
1,500	-0.04	0.05	0.07	-0.19	0.02	-0.11	-0.13	-0.23	0.07	0.26	-0.27	0.43
2,000	-0.02	0.06	0.10	-0.24	0.00	-0.10	-0.18	-0.08	-0.23	0.05	-0.03	0.50
3,000	0.00	0.09	0.15	-0.14	-0.03	-0.12	-0.23	-0.13	-0.18	-0.03	0.08	-0.14
4,000	-0.02	-0.05	0.07	-0.11	-0.02	-0.08	-0.32	-0.09	-0.14	0.42	0.27	0.11
5,000	-0.04	-0.10	-0.03	-0.10	0.01	-0.17	-0.30	-0.16	-0.12	0.23	0.22	-0.02

	RMSE(cv_{T})				RMSE(cv_{C})				RMSE($corr.$)			
Obs.	A	B	C	D	A	B	C	D	A	B	C	D
100	0.45	0.54	0.23	0.40	0.76	0.71	0.53	0.22	1.55	1.38	0.84	0.95
200	0.45	2.07	3.99	0.21	0.33	0.48	0.39	0.44	1.60	1.46	1.40	1.29
300	0.21	24.59	0.34	0.26	0.27	0.34	0.60	0.22	0.87	1.09	1.15	1.08
400	0.29	10.54	50.11	0.15	0.25	0.46	0.45	0.20	1.15	0.95	1.06	0.97
500	0.26	0.46	0.79	0.31	0.20	0.47	0.47	0.26	0.58	0.95	0.34	0.77
750	0.18	0.34	0.32	0.22	0.12	0.25	0.26	0.24	0.38	0.92	0.40	1.12
1,000	0.13	0.22	0.37	0.12	0.08	0.22	0.24	0.16	0.59	1.16	0.65	1.36
1,500	0.10	0.16	0.15	0.25	0.06	0.23	0.24	0.29	0.68	0.93	0.48	0.72
2,000	0.08	0.16	0.16	0.26	0.07	0.21	0.26	0.12	0.41	0.73	0.50	0.96
3,000	0.07	0.17	0.21	0.18	0.06	0.24	0.31	0.14	0.34	0.62	0.44	0.49
4,000	0.06	0.14	0.15	0.16	0.04	0.17	0.34	0.14	0.30	0.98	0.38	0.70
5,000	0.07	0.15	0.12	0.13	0.03	0.25	0.32	0.20	0.27	0.74	0.34	0.47

	Runs converged				Mean adj. ρ^2				Mean est. time (s)			
Obs.	A	B	C	D	A	B	C	D	A	B	C	D
100	10	2	1	1	0.38	0.11	0.07	0.31	2	3	2	4
200	10	8	6	5	0.32	0.08	0.08	0.18	3	5	5	5
300	10	8	7	3	0.31	0.11	0.11	0.17	4	8	8	8
400	10	9	8	6	0.33	0.15	0.14	0.18	6	11	9	10
500	10	8	9	9	0.34	0.16	0.15	0.19	8	11	13	12
750	10	8	9	8	0.33	0.18	0.18	0.22	11	19	20	21
1,000	10	8	8	7	0.34	0.18	0.18	0.20	15	27	28	29
1,500	10	9	10	10	0.35	0.17	0.17	0.21	23	49	41	44
2,000	10	10	10	10	0.34	0.18	0.18	0.21	32	56	60	50
3,000	10	10	10	10	0.35	0.18	0.18	0.21	51	89	98	75
4,000	10	10	9	10	0.35	0.19	0.19	0.21	66	112	118	108
5,000	10	10	10	10	0.35	0.20	0.20	0.21	80	135	141	137

Table 6: Summary estimation results for all versions of case study 2 with uncorrelated coefficients, with $5,000$ observations

	ME(cv_{T})				ME(cv_{C})			
Version	A	B	C	D	A	B	C	D
1	0.00	-0.03	-0.02	-0.05	0.02	-0.09	0.00	0.01
2	0.00	-0.02	-0.02	-0.06	0.02	-0.03	-0.01	0.02
3	-0.01	-0.02	0.00	-0.11	0.03	-0.17	0.02	-0.04
4	0.00	-0.02	-0.03	-0.06	0.01	-0.02	-0.01	-0.04
5	0.00	-0.05	-0.04	-0.08	0.02	-0.11	0.00	0.01
6	-0.02	-0.11	-0.07	-0.11	0.03	-0.11	0.03	0.02
7	-0.04	-0.16	-0.17	-0.03	0.01	-0.05	-0.05	-0.05
8	0.00	-0.25	-0.26	-0.08	0.01	-0.10	-0.10	-0.03
9	-0.06	-0.27	-0.21	-0.25	0.06	-0.08	-0.05	0.02

	RMSE(cv_{T})				RMSE(cv_{C})			
Version	A	B	C	D	A	B	C	D
1	0.04	0.10	0.08	0.12	0.04	0.15	0.09	0.08
2	0.03	0.08	0.09	0.11	0.03	0.12	0.12	0.07
3	0.06	0.14	0.13	0.17	0.05	0.20	0.11	0.15
4	0.04	0.12	0.13	0.15	0.03	0.07	0.06	0.16
5	0.04	0.17	0.15	0.11	0.04	0.23	0.06	0.07
6	0.06	0.19	0.16	0.19	0.05	0.22	0.09	0.10
7	0.07	0.18	0.18	0.15	0.04	0.08	0.08	0.07
8	0.04	0.29	0.29	0.18	0.05	0.15	0.14	0.08
9	0.10	0.41	0.40	0.29	0.11	0.15	0.13	0.10

	Runs converged				Mean adj. ρ^2				Mean est. time (s)			
Version	A	B	C	D	A	B	C	D	A	B	C	D
1	10	10	10	10	0.33	0.31	0.31	0.32	50	85	86	95
2	10	10	10	10	0.35	0.28	0.28	0.29	51	92	82	91
3	10	10	10	10	0.41	0.22	0.22	0.23	58	74	86	87
4	10	10	10	10	0.31	0.24	0.24	0.26	52	98	79	86
5	10	10	10	10	0.33	0.22	0.22	0.24	49	92	82	92
6	10	10	10	10	0.39	0.18	0.18	0.20	56	101	90	105
7	10	10	10	10	0.35	0.13	0.13	0.15	55	94	81	99
8	10	10	10	10	0.36	0.12	0.12	0.14	54	97	76	104
9	10	10	10	7	0.40	0.10	0.10	0.12	58	114	87	107

Table 7: Summary estimation results for all versions of case study 2 with correlated coefficients, with $5,000$ observations

	ME(cv_{T})				ME(cv_{C})				ME($corr.$)			
Version	A	B	C	D	A	B	C	D	A	B	C	D
10	-0.01	0.01	0.03	0.00	0.00	-0.10	-0.16	-0.05	0.16	0.27	0.48	0.15
11	-0.01	-0.02	0.03	-0.08	0.01	-0.08	-0.13	-0.08	-0.09	1.24	0.40	1.16
12	-0.03	-0.04	0.02	-0.12	0.01	-0.09	-0.10	-0.09	-0.12	0.26	0.08	0.18
13	-0.04	-0.02	0.05	-0.03	0.00	-0.08	-0.24	-0.14	0.02	0.32	0.42	0.31
14	-0.04	-0.10	-0.03	-0.10	0.01	-0.17	-0.30	-0.16	-0.12	0.23	0.22	-0.02
15	-0.04	-0.16	-0.07	-0.17	0.02	-0.09	-0.28	-0.14	-0.03	0.10	0.13	-0.18
16	-0.07	-0.05	-0.02	0.08	0.00	-0.33	-0.54	-0.36	-0.16	0.22	0.25	-0.11
17	-0.03	-0.18	-0.15	-0.02	-0.02	-0.40	-0.53	-0.37	-0.10	0.07	0.05	-0.08
18	-0.01	-0.21	0.02	-0.37	-0.01	-0.26	-0.48	-0.12	-0.01	-0.22	-0.01	-0.15

	RMSE(cv_{T})				RMSE(cv_{C})				RMSE($corr.$)			
Version	A	B	C	D	A	B	C	D	A	B	C	D
10	0.05	0.08	0.09	0.12	0.03	0.13	0.18	0.09	0.59	0.74	0.62	0.70
11	0.04	0.09	0.11	0.12	0.05	0.11	0.14	0.11	0.30	1.32	0.60	1.21
12	0.07	0.14	0.16	0.20	0.06	0.14	0.13	0.12	0.29	0.89	0.47	0.74
13	0.06	0.12	0.12	0.10	0.04	0.15	0.26	0.21	0.66	0.89	0.67	0.79
14	0.07	0.15	0.12	0.13	0.03	0.25	0.32	0.20	0.27	0.74	0.34	0.47
15	0.10	0.22	0.15	0.25	0.05	0.13	0.30	0.20	0.17	0.89	0.28	0.29
16	0.10	0.15	0.16	0.21	0.07	0.48	0.56	0.46	0.60	0.75	0.47	0.35
17	0.06	0.22	0.20	0.18	0.05	0.49	0.55	0.49	0.22	0.59	0.13	0.33
18	0.07	0.32	0.25	0.39	0.08	0.34	0.52	0.25	0.10	0.64	0.18	0.58

	Runs converged				Mean adj. ρ^2				Mean est. time (s)			
Version	A	B	C	D	A	B	C	D	A	B	C	D
10	10	10	10	10	0.33	0.30	0.30	0.31	69	127	135	130
11	10	10	10	10	0.35	0.26	0.26	0.27	67	121	131	135
12	10	10	10	10	0.41	0.20	0.20	0.21	78	131	135	136
13	10	10	10	10	0.32	0.23	0.23	0.24	68	122	125	132
14	10	10	10	10	0.35	0.20	0.20	0.21	80	135	141	137
15	10	10	9	9	0.41	0.15	0.16	0.16	82	126	158	163
16	10	10	10	9	0.36	0.12	0.12	0.14	76	140	132	179
17	10	9	9	6	0.39	0.11	0.11	0.12	107	135	124	184
18	10	10	10	9	0.44	0.09	0.09	0.10	87	128	178	146

Table 8: Detailed estimation results for version 5 of case study 3

Obs.	ME($cv_{\text{T-inter}}$) E	F	G	ME($cv_{\text{T-intra}}$) E	F	G
100	0.13	0.02	-0.27	0.06	-0.12	3.81
200	0.11	0.02	-0.38	0.17	-0.17	0.06
300	0.09	0.02	-0.18	0.09	-0.19	0.03
400	0.09	0.03	-0.21	0.05	-0.22	-0.06
500	0.06	0.01	-0.26	-0.01	-0.21	-0.03
750	0.07	0.01	-0.16	0.05	-0.21	0.22
1,000	0.05	-0.01	-0.27	0.08	-0.22	0.32
1,500	-0.01	-0.03	-0.10	-0.04	-0.22	0.11
2,000	0.00	-0.02	-0.01	-0.05	-0.22	0.01
3,000	0.01	-0.01	0.07	-0.05	-0.23	-0.05
4,000	0.01	-0.01	0.03	-0.04	-0.23	-0.05
5,000	0.00	-0.02	-0.04	-0.03	-0.24	0.00

Obs.	RMSE($cv_{\text{T-inter}}$) E	F	G	RMSE($cv_{\text{T-intra}}$) E	F	G
100	0.70	0.49	0.61	0.30	0.14	11.79
200	0.29	0.19	0.42	0.37	0.18	0.19
300	0.25	0.14	0.57	0.36	0.19	0.19
400	0.23	0.16	0.37	0.26	0.22	0.18
500	0.21	0.15	0.33	0.26	0.21	0.15
750	0.19	0.11	0.24	0.24	0.21	0.34
1,000	0.15	0.08	0.35	0.25	0.23	0.40
1,500	0.07	0.06	0.13	0.13	0.22	0.14
2,000	0.06	0.05	0.06	0.13	0.22	0.15
3,000	0.03	0.03	0.12	0.12	0.23	0.13
4,000	0.03	0.02	0.13	0.13	0.23	0.14
5,000	0.02	0.03	0.14	0.11	0.24	0.10

Obs.	Runs converged E	F	G	Mean adj. ρ^2 E	F	G	Mean est. time (s) E	F	G
100	7	10	9	0.30	0.29	0.25	219	1	1
200	10	10	9	0.29	0.29	0.22	389	2	3
300	10	10	10	0.30	0.30	0.26	532	3	5
400	10	10	10	0.32	0.32	0.27	716	4	7
500	10	10	10	0.33	0.33	0.27	887	5	8
750	10	10	10	0.34	0.34	0.29	1,390	8	15
1,000	10	10	10	0.33	0.33	0.29	1,857	10	17
1,500	10	10	10	0.34	0.34	0.29	2,857	16	31
2,000	10	10	10	0.34	0.34	0.29	3,554	21	33
3,000	10	10	10	0.34	0.34	0.29	5,309	32	61
4,000	10	10	10	0.35	0.35	0.30	6,928	41	79
5,000	10	10	10	0.35	0.35	0.30	8,991	52	90

Table 9: Estimation results for methods F and G for version 5 of case study 3: runs on full sample with $R = 200$ and $R = 40,000$

	ME($cv_{\text{T-inter}}$)		ME($cv_{\text{T-intra}}$)	
Draws	F	G	F	G
$R = 200$	-0.02	-0.04	-0.24	0.00
$R = 40,000$	-0.02	-0.30	-0.24	0.20

	RMSE($cv_{\text{T-inter}}$)		RMSE($cv_{\text{T-intra}}$)	
Draws	F	G	F	G
$R = 200$	0.03	0.14	0.24	0.10
$R = 40,000$	0.03	0.33	0.24	0.26

	Runs converged		Mean adj. ρ^2		Mean est. time (s)	
Draws	F	G	F	G	F	G
$R = 200$	10	10	0.35	0.30	52	90
$R = 40,000$	10	10	0.35	0.30	12,222	12,245

Table 10: Summary estimation results for all versions of case study 3, with $5,000$ observations

	ME($cv_{\text{T-inter}}$)			ME($cv_{\text{T-intra}}$)		
Version	E	F	G	E	F	G
1	0.01	0.00	-0.05	0.02	-0.12	0.02
2	0.00	-0.01	-0.09	0.01	-0.12	0.11
3	0.02	-0.02	-0.22	0.05	-0.11	0.44
4	0.01	0.00	-0.02	-0.03	-0.24	-0.01
5	0.00	-0.02	-0.04	-0.03	-0.24	0.00
6	0.03	-0.04	-0.11	0.04	-0.24	0.13
7	0.01	-0.04	0.34	-0.08	-0.74	-0.42
8	0.00	-0.11	-0.04	-0.08	-0.74	-0.13
9	-0.01	-0.23	-0.21	-0.05	-0.75	-0.02

	RMSE($cv_{\text{T-inter}}$)			RMSE($cv_{\text{T-intra}}$)		
Version	E	F	G	E	F	G
1	0.03	0.02	0.14	0.09	0.12	0.09
2	0.02	0.02	0.14	0.11	0.12	0.15
3	0.07	0.04	0.38	0.15	0.11	0.52
4	0.02	0.02	0.11	0.13	0.24	0.08
5	0.02	0.03	0.14	0.11	0.24	0.10
6	0.08	0.05	0.35	0.13	0.24	0.38
7	0.02	0.04	0.39	0.12	0.74	0.48
8	0.03	0.11	0.27	0.13	0.74	0.29
9	0.07	0.23	0.47	0.10	0.75	0.43

	Runs converged			Mean adj. ρ^2			Mean est. time (s)		
Version	E	F	G	E	F	G	E	F	G
1	10	10	10	0.35	0.35	0.34	8,503	51	78
2	10	10	10	0.36	0.36	0.31	8,862	52	80
3	10	10	10	0.42	0.42	0.25	9,112	60	93
4	10	10	10	0.34	0.34	0.33	8,633	52	81
5	10	10	10	0.35	0.35	0.30	8,991	52	90
6	10	10	10	0.40	0.40	0.24	8,966	63	92
7	10	10	10	0.27	0.27	0.27	8,490	49	101
8	10	10	10	0.28	0.28	0.26	9,406	48	104
9	10	10	10	0.32	0.31	0.22	9,014	53	96

Table 11: Estimation results for case study 4: part I

	ME($cv_{\text{T-inter}}$)			ME($cv_{\text{T-intra}}$)			ME($cv_{\text{C-inter}}$)			ME($cv_{\text{C-intra}}$)		
Obs.	E	F	G	E	F	G	E	F	G	E	F	G
100	1.23	3.53	-0.12	0.28	0.19	-0.02	-0.18	0.23	-0.51	-0.16	-0.15	-0.23
200	0.18	0.03	0.09	0.03	-0.16	0.09	0.06	0.25	-0.26	-0.06	0.04	-0.09
300	-0.05	0.03	-0.34	-0.02	-0.18	-0.06	0.13	0.13	0.15	0.13	-0.05	0.07
400	0.06	0.09	-0.27	-0.11	-0.20	-0.16	0.05	0.06	0.14	0.07	-0.07	0.00
500	0.07	0.07	-0.34	-0.06	-0.22	-0.02	-0.02	-0.02	0.15	0.00	-0.12	0.04
750	0.04	0.03	-0.30	-0.04	-0.21	-0.08	-0.10	-0.09	-0.16	0.01	-0.14	-0.01
1,000	0.02	-0.03	-0.17	-0.08	-0.23	0.00	-0.04	-0.02	-0.07	-0.13	-0.16	0.05
1,500	-0.02	-0.03	-0.15	-0.09	-0.21	-0.03	0.00	0.00	-0.24	-0.04	-0.17	0.16
2,000	-0.05	-0.04	-0.14	-0.12	-0.22	0.10	0.01	0.01	0.00	-0.01	-0.16	0.01
3,000	0.02	0.05	-0.03	-0.12	-0.23	0.05	-0.06	-0.06	0.13	-0.01	-0.20	-0.18
4,000	0.00	0.02	-0.20	-0.09	-0.23	0.23	-0.05	-0.06	0.17	0.01	-0.20	-0.27
5,000	-0.02	-0.01	-0.23	-0.05	-0.24	0.21	-0.03	-0.04	0.07	0.02	-0.18	-0.13

	RMSE($cv_{\text{T-inter}}$)			RMSE($cv_{\text{T-intra}}$)			RMSE($cv_{\text{C-inter}}$)			RMSE($cv_{\text{C-intra}}$)		
Obs.	E	F	G	E	F	G	E	F	G	E	F	G
100	2.12	11.18	0.44	0.59	0.81	0.19	0.43	0.41	0.51	0.19	0.21	0.24
200	0.50	0.29	0.89	0.28	0.17	0.38	0.34	0.41	0.44	0.28	0.22	0.19
300	0.26	0.18	0.36	0.16	0.18	0.10	0.31	0.28	0.70	0.28	0.14	0.08
400	0.40	0.34	0.28	0.23	0.21	0.17	0.21	0.23	0.42	0.21	0.19	0.28
500	0.26	0.30	0.39	0.24	0.22	0.24	0.20	0.19	0.47	0.24	0.20	0.24
750	0.19	0.15	0.32	0.26	0.21	0.17	0.15	0.14	0.21	0.16	0.19	0.14
1,000	0.19	0.18	0.32	0.23	0.23	0.19	0.10	0.09	0.22	0.20	0.19	0.10
1,500	0.08	0.09	0.17	0.17	0.22	0.20	0.07	0.07	0.26	0.15	0.20	0.21
2,000	0.09	0.09	0.17	0.18	0.22	0.30	0.08	0.09	0.22	0.09	0.18	0.22
3,000	0.07	0.09	0.17	0.17	0.23	0.20	0.08	0.09	0.17	0.09	0.21	0.19
4,000	0.06	0.06	0.24	0.15	0.23	0.31	0.06	0.07	0.20	0.07	0.21	0.28
5,000	0.07	0.06	0.27	0.13	0.24	0.32	0.04	0.05	0.24	0.07	0.19	0.29

	ME (*corr.*-inter)			RMSE (*corr.*-inter)			ME (*corr.*-intra)			RMSE (*corr.*-intra)		
Obs.	E	F	G	E	F	G	E	F	G	E	F	G
100	0.91	1.62	1.34	1.30	1.67	1.37	0.10	0.10	0.10	0.10	0.10	0.10
200	1.65	1.58	1.21	1.65	1.59	1.35	0.10	0.10	0.10	0.10	0.10	0.10
300	-0.13	0.08	-0.31	0.62	0.62	0.31	0.09	0.09	0.09	0.09	0.09	0.09
400	0.58	0.67	1.28	1.00	0.96	1.30	0.05	0.05	0.05	0.05	0.05	0.05
500	-0.19	-0.37	-0.01	0.39	0.39	0.52	0.04	0.04	0.04	0.04	0.04	0.04
750	-0.08	-0.15	0.18	0.59	0.59	0.59	0.01	0.01	0.01	0.01	0.01	0.01
1,000	0.45	0.40	0.74	0.91	0.95	1.05	0.01	0.01	0.01	0.01	0.01	0.01
1,500	0.08	0.07	0.08	0.67	0.74	0.78	-0.01	-0.01	-0.01	0.01	0.01	0.01
2,000	-0.19	-0.23	-0.14	0.32	0.33	0.35	-0.02	-0.02	-0.02	0.02	0.02	0.02
3,000	-0.17	-0.12	-0.22	0.31	0.25	0.34	-0.01	-0.01	-0.01	0.01	0.01	0.01
4,000	-0.11	-0.12	0.02	0.25	0.26	0.20	0.00	0.00	0.00	0.00	0.00	0.00
5,000	0.09	0.03	-0.02	0.51	0.52	0.12	0.00	0.00	0.00	0.00	0.00	0.00

Table 12: Estimation results for case study 4: part II

	Runs converged			Mean adj. ρ^2		
Obs.	E	F	G	E	F	G
100	2	9	3	0.41	0.36	0.09
200	7	10	4	0.30	0.31	0.05
300	9	10	2	0.28	0.28	0.08
400	9	10	5	0.30	0.30	0.11
500	10	10	8	0.30	0.31	0.11
750	10	10	6	0.29	0.29	0.13
1,000	9	10	4	0.29	0.29	0.13
1,500	10	10	4	0.29	0.29	0.12
2,000	10	10	5	0.29	0.29	0.13
3,000	10	10	4	0.29	0.29	0.13
4,000	10	10	6	0.29	0.29	0.14
5,000	10	10	8	0.30	0.30	0.15

	Mean est. time (s)		
Obs.	E	F	G
100	535	3	3
200	1,032	5	8
300	1,543	8	15
400	1,958	11	16
500	2,591	12	19
750	3,797	20	32
1,000	4,547	27	76
1,500	7,703	42	66
2,000	11,256	57	190
3,000	15,207	89	204
4,000	20,163	111	326
5,000	24,964	146	301