An Early Stopping Bayesian Data Assimilation Approach for improved Mixed Multinomial Logit transferability

An Early Stopping Bayesian Data Assimilation Approach for improved Mixed Multinomial Logit transferability

Abstract

Mixed Multinomial Logit (MMNL) models can provide valuable insights into inter and intra-individual heterogeneity in transportation choice modelling. However, the high computational and data requirements for MMNL models has limited the application of MMNL models in practice. These requirements are particularly problematic when investigating the behaviour of specific population sub-groups or market segments, where a modeller may want to estimate separate models for a number of similar contexts, each with low data availability. The same challenges arise when adapting one model to a new location or time period.

To overcome these barriers, we establish a new Early Stopping Bayesian Data Assimilation (ESBDA) approach which updates a previously estimated MMNL on a new data sample or subsample through iterative *Bayesian inference*. This approach therefore enables an existing model from one context to be transferred to a new context with lower data availability.

The ESBDA approach is benchmarked against two reference estimators: (i) a standard Bayesian estimator (MMNL); and (ii) a Bayesian Data Assimilation (BDA) estimator without *early stopping*. The results show that the proposed ESBDA approach can effectively overcome over-fitting and non-convergence. ESBDA models outperform the models estimated by the reference estimators in terms of behavioural consistency of parameter estimates and the out-of-sample predictive performance of the model. Even when using few collected data, ESBDA can still produce suitable and stable MMNL model with parameter estimates consistent with established behavioural theory.

Keywords: Multinomial Logit, Bayesian Data Assimilation, Model Transferability, Early Stopping

1. Introduction

Discrete Choice Models (DCMs) are crucial modelling tools in transport, economics, health, and other disciplines where individual choice behaviour is a key research interest. The most prominent DCM used in practice is the Multinomial Logit (MNL) model, where the parameters in the utility functions have fixed values across the population. However, the fixed parameters in MNL models do not account for the significant inter and intra-individual heterogeneity in individual choice behaviour. Variants of the standard MNL have emerged to accommodate heterogeneity through assuming a distribution in the modelled parameters over the population. These distributions can be discrete, as in the Latent Class Model (LCM) (Bhat, 1997; Greene and Hensher, 2003);

or continuous, as in the Mixed Multinomial Logit (MMNL) (Cardell and Reddy, 1977; Ben-Akiva and Bolduc, 1996; McFadden and Train, 2000). However, both MMNL and LCMs have higher computational and data requirements than MNL models, which has limited their application in practice.

The computational and data requirements of MMNL models become restrictive when investigating market segmentation, where separate models for a different population subgroups are estimated, each with low data availability. Meanwhile, as the parameter combinations of DCMs estimated for different populations can vary greatly, it is difficult to apply or transfer models to a new modelling context (e.g., modelling for a new location or for the future). This means areas or population segments with poor data availability cannot benefit substantially from existing models. There is therefore a need to address the problem of data shortage in modelling heterogeneous choice behaviour.

This paper addresses this need through the introduction of the Early Stopping Bayesian Data Assimilation (ESBDA) estimator. Following the idea of model transferibility (Ben-Akiva and Bolduc, 1987), the ESBDA estimator is designed to adapt a previously established model to a new population subgroup, location, or time period. The adaptation is achieved through Bayesian Data Assimilation (BDA). The proposed estimator is equipped with *early stopping* procedures to help prevent over-fitting or non-convergence, which are recurrent conundrums of small sample modelling of DCMs. The major contribution of ESBDA is the enabling a modeller to exploit an existing model estimated on a different population to estimate a model on a new population. This is of particular benefit when the modeller does not have access to sufficient data to estimate a reliable model directly on the new population.

2. The Early Stopping Bayesian Data Assimilation (ESBDA) Estimator

2.1. Modelling Approach

This section presents the modelling approach used, focusing on Bayesian Data Assimilation (BDA) and *early stopping* procedures, where ESBDA gets its major advantages over the conventional *hierarchical Bayesian* estimator of MMNL. For a complete review of modelling MMNL, we direct the reader to Chapter 6 of Train (2003).

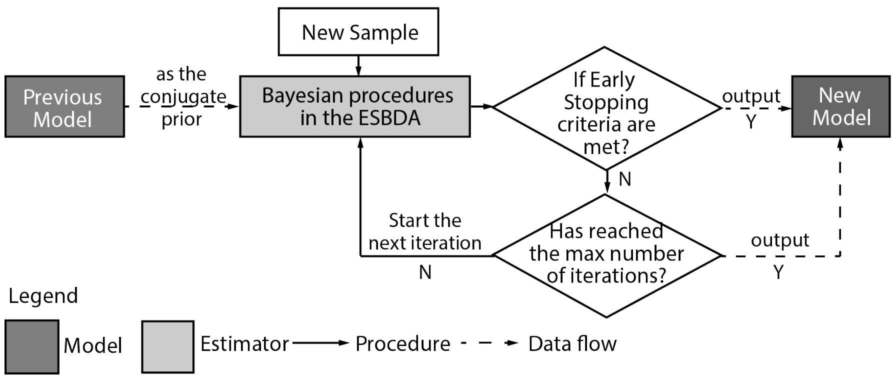

Figure 1: Flowchart of iterative Bayesian procedures and early stopping procedures of ESBDA.

2.1.1. Bayesian Data Assimilation (BDA)

BDA is a phrase first used in the time-series modelling literature to describe a time-related model refinement/calibration technique which uses new information as it becomes available (Jazwinski, 1970; Reich and Cotter, 2015). This paper extends the definition of BDA to 'the technique of *data assimilation* through *Bayesian inference* for general transfer/update of a previously established model'.

In the context of updating a model for a new population, the conjugate prior is the previously estimated parameter combination of a similar model. Through assimilating the sample data of the modelling object, the *prior* evolves into a new parameter combination (*posterior*) fitting to the target context. The assimilation is processed through iterative Bayesian inference, where each parameter is updated in response to the condition of the rest of the parameter combination. We will introduce the procedure in Section 2.2.1.

2.1.2. Early Stopping

Early stopping is a commonly used technique in Machine Learning (ML), which stops the training before convergence in order to regularise the model and prevent over-fitting. It tracks the real-time validation and/or training set error(s) and terminates the modelling when an *early stopping* criterion is met. This technique is not used in the conventional MMNL estimation as this model type rarely has a high dimensional parameter space and therefore has a relatively low risk of *over-fitting*.

However, *early stopping* could also be of benefit for the MMNL models when modelling with very small sample sizes. *Early stopping* can prevent the resultant model from over-fitting to the insufficient sample data which may not be representative of the population to be modelled. Meanwhile, the insertion of *early stopping* procedures essentially configures the core of Maximum Simulated Likelihood (MSL) estimator into the Bayesian estimator. As such, ESBDA becomes essentially a hybrid of the two most prominent MMNL estimators — the MSL and the hierarchical Bayes (HB) procedure. The HB procedure typically terminates the simulation when the MMNL model converges. However, the model may never converge when the sample size is small, which makes when to stop the modelling become a tough decision. In this case, this decision could be left to the *early stopping* procedures. Despite these potential benefits, to the best of the authors' knowledge, this paper is the first effort to configure the *early stopping* procedure into a MMNL estimator.

Here we present the mainstream classes of *early stopping* criteria that are applicable to our estimator. Let E denote the modelling error and $E_{opt}(T)$ denote the lowest error obtained in epochs until T:

$$E_{opt}(T) = \min_{T' \leq T} E(T') \tag{1}$$

The relative generalisation loss at epoch T (in percent) is:

$$L(T) = 100 \cdot (\frac{E(T)}{E_{opt}(T)} - 1) \tag{2}$$

(i) The first class of stopping criteria, L_α, stops modelling as soon as the generalisation loss drops to a certain threshold, i.e., stop when $L(T) > \alpha$.

(ii) The second class, Q_α, stops training when the decreasing speed of E,$P_k(t)$, drops below a certain threshold, i.e., $\frac{L(T)}{P_k(T)} > \alpha$, where k refers to consecutive k epochs, e.g., a training strip. And the decreasing speed (in per thousand) of E is:

$$P_k(T) = 1000 \cdot \left(\frac{\sum_{T'=T-k+1}^{T} E(T')}{k \cdot \min_{T'=T-k+1}^{T} E(T')} - 1\right) \tag{3}$$

(iii) The third class triggers stopping when the L increased in s successive strips:
S_s: stop after epoch T iff $E(T) > E(T-k)$ and S_{s-1} stops after epoch $T-k$; or
S_1: stop after the end of first strip t with $E(T) > E(T-k)$.

Early stopping criteria (except for Q_α) are typically applied to the validation set only because the training set error is automatically tracked in the form of likelihood during the MSL modelling progress. However, as the HB itself does not track the likelihood in the training set, we apply *early stopping* to the both datasets. The one in the training set serves as a lightweight MSL to supervise the modelling error during the HB procedure.

2.2. Estimator Formulation

2.2.1. Algorithm

Our framework is built on a fundamental Bayesian estimator which employs the HB procedure. The HB method was initially established by Rossi et al. (1996) and Allenby (1997) and the estimator was then coded by Train (2006). We have recoded extensively to realise modelling error tracking and plotting, to adapt the codes to the new model, etc. For simplicity, the diagram of the algorithm (Fig.2) illustrates only the modifications that make a sound difference to the estimation results. The key extensions of the new estimator from the standard HB procedure are on the two ends of the original algorithm: (i) the adoption of a *conjugate prior* parameter combination in the beginning and (ii) the *early stopping* procedures to terminate modelling.

The estimator assimilates new data and approximate the posterior estimates by Markov Chain Monte Carlo (MCMC) sampling through iterative *Bayesian inference* procedures. We illustrate the procedures using the multivariate normal, as it is relatively easy to follow numerically[1]. For $\beta_n \sim N(b, W)$, we have the utility function:

$$U_{nj} = \alpha' z_{nj} + (\bar{\beta}'_n + \sigma_n \zeta_{nj}) x_{nj} + \varepsilon_{nj} \tag{4}$$

where α' and $\bar{\beta}'_n + \sigma_n \zeta_{nj}$ are vectors of fixed and random coefficients respectively. z_{nj} is the vector of fixed-weighted explainable variables and x_{nj} is that of random-weighted variables. ε_{nj} is the remaining unobserved utility which is independent and identically distributed (iid) Extreme-Value 1 type (EV1). Then the conditional posteriors in each layer of Bayesian inference are:

1. $K(\beta_n|\alpha, b, W) \propto L(y_n|\alpha, \beta_n)\phi(\beta_n|b, W)$[2]. It is not in a closed form, so Metropolis–Hastings Algorithm (M-H) is used to obtain a simulated β_n on the pooled data.

2. $K(b|W, \beta_n \forall n)$ is $N(\sum_n \beta_n/N, W/N))$. Note α does not enter this layer directly. Its affect on posterior b is passed through the draws of β_n from the first layer.

[1] The limited space only allow us to present out methodology as succinct as possible. For an in-depth study of the HB procedure, we direct the reader to Chapter 9 and 12 of Train (2003).

[2] We use β_n to denote the real-time simulated α' and $\bar{\beta}'_n + \sigma_n \zeta_{nj}$.

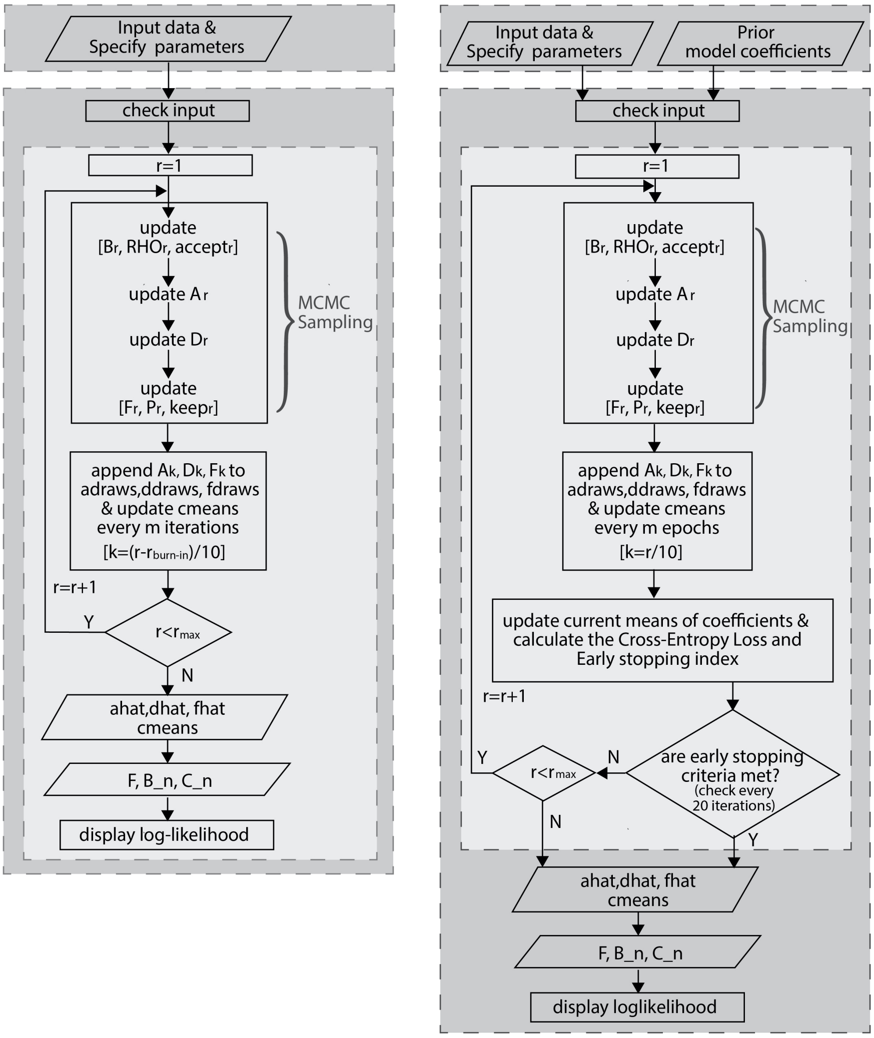

Figure 2: Estimation procedures of the original Train's Hierarchical Bayes procedures (left) and the proposed ESBDA estimator.

3. $K(W|b, \beta_n \forall n)$ is $IW(K+N, (KI+N\bar{S})/(K+N))$ where $\bar{S} = \sum_n (\beta_n - b)(\beta_n - b)^{'}/N$. Similarly, α does not involve directly.

4. $K(\alpha|\beta_n) \propto \Pi_n L(y_n|\alpha, \beta_n)$. M-H may be used again when the *prior* on α is essentially flat.

The method can be conveniently adapted to variants of normal distribution simply through distribution transformation. Denote the weights of random utility terms in person n's utility function as c_n, and $c_n = T(\bar{\beta}_n' + \sigma_n \zeta_{nj})$, where T refers to a distribution transformation which depends only on the latent distribution parameters and which is weakly monotonic (to maintain $\partial c_n^k / \partial c_n^{\bar{\beta}_n' + \sigma_n \zeta_{nj}} \geq 0$ for elements in β_n or c_n). The distributed random parameter is drawn in the same manner in modelling but it enters the utility function in its transformed form:

$$U_{nj} = \alpha' z_{nj} + T(\bar{\beta}_n' + \sigma_n \zeta_{nj})' x_{nj} + \varepsilon_{nj} \tag{5}$$

Whilst the derivation of the resulting posterior in each layer may change in other flexible distributions, the procedures are broadly similar.

2.2.2. *Hyper-parameters*

The class(es) of early stopping criteria and the threshold value(s) are usually selected in an interactive fashion (Precheit, 1998). For model performance in estimation and for hyper-parameters optimisation, our estimator employs Cross-Entropy Loss (CEL) (Eq.6), which is a normalised measure independent on the sample size. And the *early stopping* criteria that we incorporate are L_α and S_s.

$$\begin{aligned} G_{CEL} &= -\frac{1}{N} G_{\text{log-likelihood}}, \\ &= -\frac{1}{N} \sum_{n=1}^{N} \ln P(i_n|x_n) \end{aligned} \tag{6}$$

To guarantee model termination, stopping criteria are complemented by a rule that terminates modelling after a set number of epochs. Other hyper-parameters, e.g., the total number of epochs, the number of draws for simulating the distributed parameters, are also set on an ad-hoc basis. To relieve serial correlation of M-H, draws of *posterior* distribution of α, β_n are retained at regular intervals instead of consecutively (every T_1 epochs). The same rule is applied to CEL tracking and plotting (every T_2 epochs).

2.2.3. *Target Scenarios*

The estimator is developed principally to transfer/update a previously estimated model to adapt (i) another location, (ii) demographic/ locational population segments, or (iii) a different time period (given the prediction of demographic change). It can also be used for normal MMNL estimation.

For model segmentation, a hierarchical modelling structure can be established to investigate heterogeneous choice behaviour hierarchically — from a general level to specific detailed segments — through layers of ESBDAs (see Fig. 3). In each level of segmentation, the coefficients, i.e., the *posterior*, estimated for the upper level model are input to the ESBDA as the *prior* to estimate this level *posterior* coefficients, which will then serve as the *prior* of the next level.

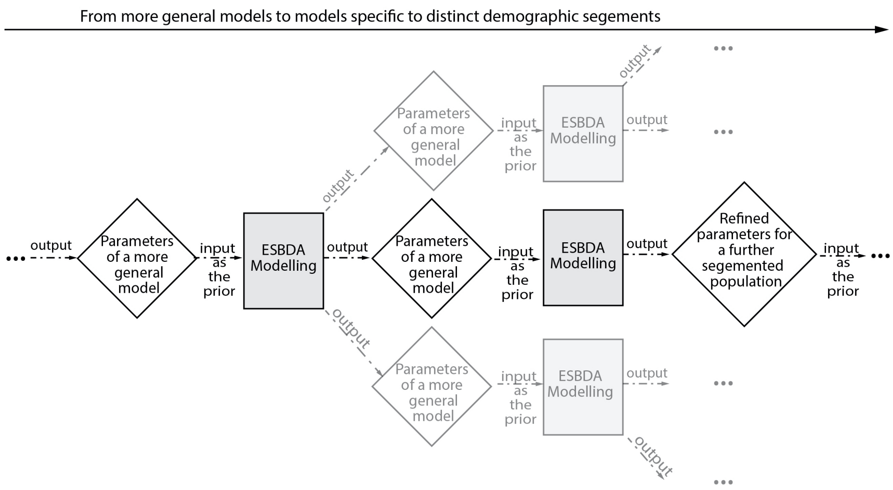

Figure 3: Hierarchical modelling structure for a systematic DCM model segmentation

3. Testing procedure

Simulation experiments are carried out to benchmark the ESBDA estimator against two reference estimators. The alternative estimators are run to estimate the same MMNL model. The comparison is made on the basis of the estimation results.

3.1. Benchmark Estimators

As Table 1 shows, the basic benchmark estimator is the HB procedure (Train, 2006). The starting values of the model coefficients are selected at random. The estimation is therefore based solely on the limited data of the modelling target. We further employ an intermediate estimator, i.e., BDA, as another reference estimator. It takes full advantage of the *Bayesian inference* to develop the new model informatively from a informative *prior* model. Unlike the proposed estimator, this estimator is not equipped with *early stopping* procedures.

Table 1: The ESBDA estimator and the benchmark estimators

Estimator	*prior*-based ESBDA	*prior*-based BDA	Non-*prior* Bayesian Estimator
Simulation mechanism	Bayesian modelling	Bayesian modelling	Bayesian modelling
prior	Previously estimated parameters	Previously estimated parameters	No-*prior*
Early stopping procedures	Yes	No	No

3.2. Measures of Estimator Performance

The estimators are compared across three performance dimensions. The first is the statistics of their modelling estimates, e.g., statistical significance, sign errors. The second indicator is the modelling progress: how steady CEL progresses throughout a complete modelling; whether the estimation result truly converge; and where does the modelling early stop. The last is whether the parameter combination is interpretable. The ratios of model coefficients against the weight of money can provide valuable insights into people's willingness to pay of different factors (time in particular) in choice behaviour. A MMNL estimator should give the model a parameter combination which is highly interpretable and therefore informative to exploring people's valuation of different factors in choice behaviour. On the contrary, the modelling methodology may need to be reconsidered if there is a suspicious coefficient ratio. As such, we employ an additional estimator evaluation dimension, which is the modeller's judgement of the resultant time-cost-ratio, i.e., how far the estimated ratio deviates from empirical values of Value of Time (VOT).

3.3. *Case Study: Modelling Travel Mode Choice in London*

The model used to test the three estimators is a MMNL model that we develop for modelling the travel mode choice in London. The dataset, available online, is adapted from a closely tailored London travel dataset[3](Hillel, 2019) which recreates the travel mode choice-set that are faced by the respondentsat the time of travel.

The time/cost ratio is of particular interest in transport modelling. To investigate the ratio, the values of time and of cost cannot be both random at the same time. Therefore, we assign a normal distribution to the cost value and maintain all other coefficients/constants as fixed value parameters. People's perception and valuation of time varies when travelling in different modes. So we set alternative-specific parameters for the utility functions of the four modes, i.e., driving (U_{n_driving}), public transit (U_{n_public}), cycling (U_{n_cycling}), and walking (U_{n_walking}). The utility functions are as follows (Eq.7–10)[4]. The explanatory variables and the coefficients of the model are presented in Table 2.

$$U_{n\_\text{driving}} = (\bar{\beta}_{n\_\text{cost}} + \sigma_{n\_\text{cost}}\zeta_{n\_\text{cost}})c_{n\_d} + \alpha_{\text{driving-time}} t_{n\_d} + \alpha_{\text{var}}\nu_{n\_d} + \varepsilon_{n\_\text{driving}} \tag{7}$$

$$\begin{aligned} U_{n\_\text{public}} =& (\bar{\beta}_{n\_\text{cost}} + \sigma_{n\_\text{cost}}\zeta_{n\_\text{cost}})c_{n\_p} + \alpha_{\text{access-time}} t_{n\_a} + \alpha_{\text{bus-time}} t_{n\_b} + \alpha_{\text{rail-time}} t_{n\_r} + \\ & \alpha_{\text{change-walking-time}} t_{n\_\text{change1}} + \alpha_{\text{change-waiting-time}} t_{n\_\text{change2}} + C_{\text{public}} + \varepsilon_{n\_\text{public}} \end{aligned} \tag{8}$$

$$U_{n\_\text{cycling}} = \alpha_{\text{cycling-time}} t_{n\_c} + C_{\text{cycling}} + \varepsilon_{n\_\text{cycling}} \tag{9}$$

$$U_{n\_\text{walking}} = \alpha_{\text{walking-time}} t_{n\_w} + C_{\text{walking}} + \varepsilon_{n\_\text{walking}} \tag{10}$$

Table 2: Variables and Coefficients of the London Travel Mode Choice Model, and the distributions of the coefficients

Variable / Constant	Symbol	Coefficient	Distribution
Travel Cost	c_{n_d}(driving); c_{n_p}(public)	β_{n_cost}	normal
Driving time	t_{n_d}	$\alpha_{\text{driving-time}}$	fixed
Access time	t_{n_a}	$\alpha_{\text{access-time}}$	fixed
In-vehicle time on bus	t_{n_b}	$\alpha_{\text{bus-time}}$	fixed
In-vehicle time on rail	t_{n_r}	$\alpha_{\text{rail-time}}$	fixed
Interchange walking time	t_{n_change1}	$\alpha_{\text{change-walking-time}}$	fixed
Interchange waiting time	t_{n_change2}	$\alpha_{\text{change-waiting-time}}$	fixed
Cycling time	t_{n_c}	$\alpha_{\text{cycling-time}}$	fixed
Walking time	t_{n_w}	$\alpha_{\text{walking-time}}$	fixed
Traffic variability	ν_{n_d}	α_{ν}	fixed
Constant of the Public transit mode	-	C_{p}	fixed
Constant of the Cycling mode	-	C_{c}	fixed
Constant of the Walking mode	-	C_{w}	fixed

[3]available on https://www.icevirtuallibrary.com/doi/suppl/10.1680/jsmic.17.00018.

[4]We use β_{n_cost} to denote $(\bar{\beta}_{n\_\text{cost}} + \sigma_{n\_\text{cost}}\zeta_{n\_\text{cost}})$ in the rest of this paper.

To test the estimation ability of the ESBDA, BDA, and HB approaches at different samples sizes, tests are conducted with four subsamples of the dataset, as shown in Table 3. In addition, the levels of modelling also provide a platform to illustrate our idea of building a modelling system for hierarchical model segmentation, using layers of ESBDA.

Given the sufficiently large full dataset, there is no problem of over-fitting or non-convergence in modelling by any alternative estimator at Level 0. The purpose of Level 0 modelling is to derive a 'mother model' to feed *conjugate prior* parameters to the next level modelling. Level 1-3 Models use corresponding population sample data as they are developed to investigate travel behaviour of a certain sub-population.

Table 3: Modelling levels, and the corresponding modelling objective and sample size of each level

Sample size	Modelling object	Number of choice samples	
		Training	Validation
Level 0	All journeys, regardless of travel purpose time period of travelling or the traveller's attributes, income, age, etc.	8331	7817
Level 1	General home-office journeys, regardless of time period of travelling or the traveller's attributes, income, age, etc.	613	735
Level 2	Home-office journeys during morning peak-time, regardless of the traveller's attributes, income, age, etc.	266	264
Level 3	Home-office journeys during morning peak-time; the 26-35-year-old people whose household income is between £25,000-£49,999.	26	27

3.4. *Experimental Setup*

The *early stopping* criteria used for the modelling are: $L(T)$, UP and a complementary criterion which terminate modelling anyway after 5,000 epochs. The threshold values of $L(T)$ and UP are set to $\alpha = 2$ and $s = 100$. And we set $T_1 = 10$ and $T_2 = 20$.

4. Results and Discussion

Performance of alternative estimators is analysed on the grounds of (i) Statistics and behavioural consistency of parameter estimates (Table 5 to 7) and (ii) the steadiness of modelling progress (Fig. 4 to 6).

Estimated parameters with statistical insignificance and sign error are highlighted in Table 6 to 7. We omit the plot of level-0 modelling as the modelling progresses of all the three estimators are steady and ESBDA does not undergo *early stopping*. Three estimators finally converge to indistinguishable estimates, with majority estimates being statistically significant.

The sample size at Level-1 is still relatively large. CELs of all the three estimators approach their asymptotes. ESBDA early stops at the 2540th epoch. None of the estimators encounters sign error. But some time-cost ratios (e.g., driving time/cost, as highlighted in Table 3) produced by the two benchmark estimators are found to have deviated from the corresponding empirical values at Level 0. The proposed estimator has no such problems and is thus favoured over the benchmark estimators. While the models estimated by the other estimators may mislead behaviour interpretations, the model estimated by ESBDA still maintains strong explanatory power that an appropriate MMNL model is supposed to have.

Table 4: Estimates of the Level 0 modelling

coefficient	value	coefficient	value	coefficient	value
β_{n_cost}	-0.1605	$\alpha_{\text{driving-time}}$	-3.4996	$\alpha_{\text{access-time}}$	-3.4173
$\alpha_{\text{bus-time}}$	-2.2110	$\alpha_{\text{rail-time}}$	-2.3821	$\alpha_{\text{change-walking-time}}$	-1.9474
$\alpha_{\text{change-waiting-time}}$	-2.6313	$\alpha_{\text{cycling-time}}$	-4.6405	$\alpha_{\text{walking-time}}$	-6.2339
α_{ν}	-5.1859	C_{p}	1.7403	C_{c}	0.2730
C_{w}	3.5505				

Table 5: Estimates of the Level 1 modeling through alternative estimators
(* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ^{e} sign error; $^{!}$ unreasonable time-cost ratio)

	ESBDA Estimator (not early stopped)		BDA Estimator		*Non-conjugate-prior* Estimator	
	Mean	StDv	Mean	StDv	Mean	StDv
μ_{n_cost}	-0.1637***	0.0075	-0.1211**	0.0409	-0.1275***	0.0369
σ_{n_cost}	0.0094***	0.0017	0.0486	0.0275	0.0384	0.0237
β_{n_cost}	-0.1725	0.0957	-0.1411	0.2182	-0.1453	0.1941
$\alpha_{\text{driving-time}}$	-3.6692***	0.7421	-0.3705$^{!}$	1.9873	-3.0117	3.0168
$\alpha_{\text{access-time}}$	-2.2671*	0.9753	–5.7970*	2.4072	-5.1126	3.2635
$\alpha_{\text{bus-time}}$	-1.4018	0.7978	-2.3862	1.5479	-2.6516	1.5919
$\alpha_{\text{rail-time}}$	-1.2581	1.1941	-1.4583	2.4762	-2.6057	2.8412
$\alpha_{\text{change-walking-time}}$	-1.4994	0.8110	-2.5526	2.6399	-0.1235$^{!}$	1.1013
$\alpha_{\text{change-waiting-time}}$	-1.5642	1.5574	-4.9852	2.5574	-6.3279	4.2701
$\alpha_{\text{cycling-time}}$	-5.0625***	0.5994	-5.7544**	1.8166	-7.0572*	2.9924
$\alpha_{\text{walking-time}}$	-7.6346***	0.5759	-8.8903***	1.1969	-8.9231***	2.2502
α_{ν}	-6.2389**	1.9229	-11.9053***	2.7490	-13.8196	5.4239
C_{p}	1.7216***	0.4506	2.5713*	1.2857	1.7251	1.0461
C_{c}	1.1120	0.3734	1.2281	1.0263	0.7673	0.9277
C_{w}	4.6604***	0.5723	5.4952***	0.9213	0.5723***	1.0875

As the sample size continues to shrink, the benchmark estimators both see an inevitably increased fluctuation of CELs during modelling, in particular, the Non-*conjugate-prior* estimator. Given a handful of sample data, as the plots show, the Level-2/3 models are unlikely to see convergence on the training set or validation set, even with a set of informative *conjugate-prior* parameters. Estimates generated by one epoch can change massively within just several epochs under the unsteady modelling process. Not surprisingly, obvious sign errors occur in the estimations by the two reference estimators.

Table 6: Estimates of the Level 2 modeling through alternative estimators
(* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ^{e} sign error; $^{!}$ unreasonable time-cost ratio)

	ESBDA Estimator (1480 epochs)		BDA Estimator		Non-*conjugate-prior* Estimator	
	Mean	StDv	Mean	StDv	Mean	StDv
μ_{n_cost}	-0.1723***	0.0497	-0.4450	0.3775	-0.0974	0.0965
σ_{n_cost}	0.0713***	0.0180	0.5705	0.4513	0.2249	0.1385
β_{*_cost}	-0.1965	0.2644	-0.5134	0.7478	-0.1403	0.4695
$\alpha_{\text{driving-time}}$	-3.1287***	0.7790	-4.5707	1.8986	-5.4834*	2.8758
$\alpha_{\text{access-time}}$	-7.1108***	1.8538	-5.6903	3.0392	4.2442^{e}	3.4498
$\alpha_{\text{bus-time}}$	-4.2545***	1.0576	-7.3937*	2.8542	-1.4524	2.2295
$\alpha_{\text{rail-time}}$	-2.0064	1.2996	-5.8100	4.4338	-0.0960$^{!}$	3.6980
$\alpha_{\text{change-walking-time}}$	-2.7981*	1.1139	-2.7262	2.1047	1.3032^{e}	2.7747
$\alpha_{\text{change-waiting-time}}$	-1.5147	1.1692	-1.2664	2.0883	3.9445^{e}	2.3518
$\alpha_{\text{cycling-time}}$	-5.3324***	0.5847	-12.4342*	5.6856	-5.2187*	2.4230
$\alpha_{\text{walking-time}}$	-9.1773***	1.9352	-12.7062***	2.0968	-8.4056***	2.3010
α_{u}	-8.6779***	1.6804	-10.4169***	2.3457	-4.8123	2.9461
C_{p}	1.7203	1.2760	8.2318	4.1037	2.5942*	1.3842
C_{c}	-0.3207	0.9438	7.2235	4.6470	3.7693	1.9513
C_{w}	4.2865*	1.7358	12.1655**	4.0600	7.5329***	2.1980

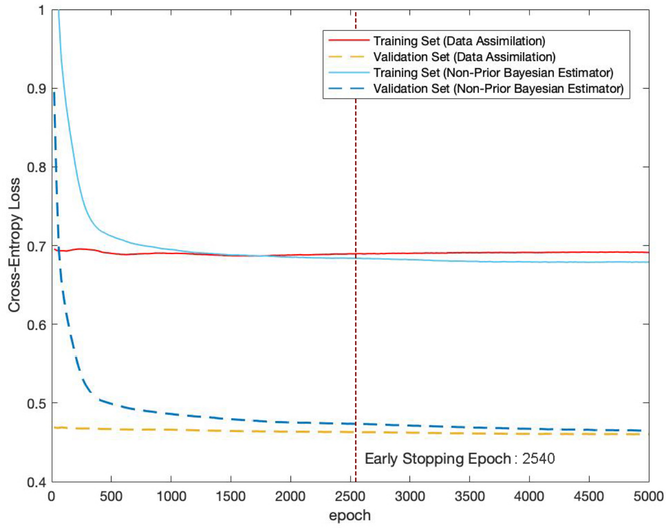

Figure 4: Comparison of Cross-Entropy Loss (CEL) of the *conjugate-prior*-based BDA and the Non-*conjugate-prior* Bayesian Estimator (Level 1 modelling)

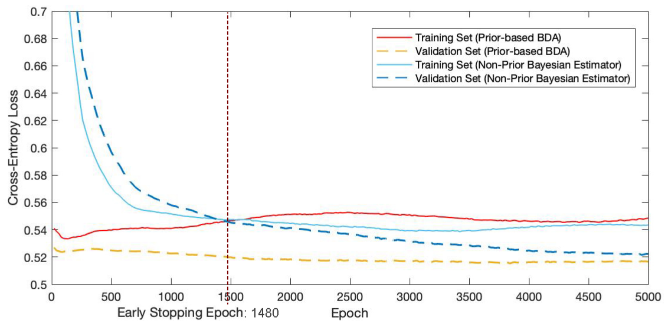

Figure 5: Comparison of Cross-Entropy Loss (CEL) of the *conjugate-prior*-based BDA and the Non-*conjugate-prior* Bayesian Estimator (Level 2 Modelling)

In contrast, ESBDA still arrives at acceptable estimates and maintains strong interpretability. This is attributed to the *early stopping* procedures which effectively terminates the modelling before the estimation diverges under the unsteady simulation process.

Overall, levels of modelling experimentation suggests that of the three estimators, ESBDA is superior in terms of quality of estimates, modelling speed, the steadiness of the modelling process, and the trade-off between the *conjugate prior* and the sample data.

Table 7: Estimates of the Level 3 modeling through alternative estimators
(* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ^{e} sign error; $^{!}$ unreasonable time-cost ratio)

	ESBDA Estimator (80 epochs)		BDA Estimator		Non-*conjugate-prior* Estimator	
	Mean	StDv	Mean	StDv	Mean	StDv
μ_{n_cost}	-0.2010	0.0497	-0.3725	0.4018	0.2163	0.5389
σ_{n_cost}	0.0310	0.0180	0.5252	0.4045	0.6860	1.8527
β_{n_cost}	-0.2262	0.2755	-0.4382	0.7175	-0.2914	0.0820
$\alpha_{\text{driving-time}}$	-2.3779***	0.1864	-2.8289	1.9983	-7.7548	7.1594
*** $\alpha_{\text{access-time}}$	-6.6902***	0.3654	-11.6366**	3.4448	-8.1120*	2.9168
$\alpha_{\text{bus-time}}$	-4.5864***	0.4108	-7.0441	4.8814	-4.1359*	1.7178
$\alpha_{\text{rail-time}}$	-1.5990***	0.2956	-0.9152	2.6612	-1.5551	1.9935
$\alpha_{\text{change-walking-time}}$	-2.5548***	0.0567	6.2294^{e}	5.5857	0.5287^{e}	3.1137
$\alpha_{\text{change-waiting-time}}$	-1.9652***	0.2486	-2.2056	1.9487	-3.1581	3.1191
$\alpha_{\text{cycling-time}}$	-5.4081***	0.1340	-9.1297*	3.2975	-1.9244	3.2986
$\alpha_{\text{walking-time}}$	-8.4161***	0.5026	-5.9748**	1.3551	-5.8217*	1.9707
α_{ν}	-8.5447***	0.1241	-10.4248*	3.6031	-3.4708	2.1917
C_{p}	1.9808***	0.1699	3.1394	1.9598	5.2596	2.5588
C_{c}	-0.6847	0.3522	-0.2329	1.8483	0.3442	2.0902
C_{w}	4.2805***	0.1956	2.1939	2.1759	5.0985*	1.8686

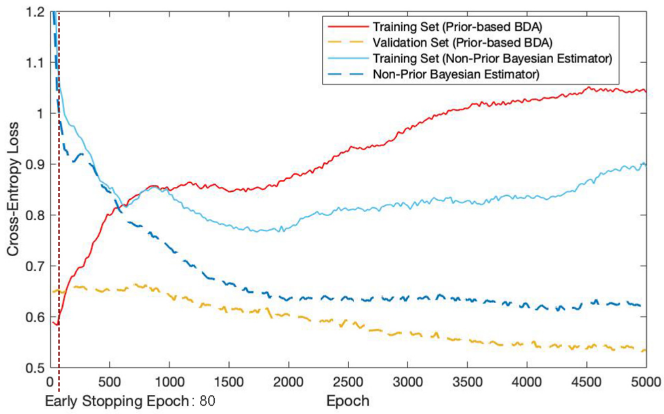

Figure 6: Comparison of Cross-Entropy Loss (CEL) of the *conjugate-prior*-based BDA and the Non-*conjugate-prior* Bayesian Estimator (Level 3 modelling)

5. Conclusions and Future Work

This paper presents a new ESBDA estimator which provides a practical approach to build new MMNLs by transferring/updating from previously estimated model parameters through BDA. Data assimilation is processed using iterative *Bayesian inference*. The estimator is equipped with a lightweight MSL analogue to complement the *Bayes procedures* through inserting the *early stopping* procedures.

The proposed estimator is tested in modelling experiments at three levels of sample size. the model estimated by the ESBDA outperforms its counterparts of the benchmark estimators in each of the three considered dimensions for each experiment. In all experiments, the proposed estimator appears to be the only estimator which yields a decent MMNL model with interpretable coefficients.

Experimental results suggest that the ESBDA estimator is superior over the plain Bayesian estimator and the *conjugate-prior*-based BDA estimator. ESBDA inherits the merits of the two most prominent MMNL estimators — the MSL and the HB procedure, as it is essentially a hybrid. Data assimilation prevents the resultant model from over-dependence on the previously established model, which is not tailored to the modelling target. The model also has addressed the problem of over-fitting to the sample data which may not be sufficiently representative of the modelling target. Meanwhile, ESBDA can effectively prevent non-convergence, which has been a recurrent problem when modelling with little sample data.

The study has several limitations which point to anticipatory yet challenging future research directions to achieving the full potential of the ESBDA Estimator. Planned research work includes: (i) Further comparing ESBDA to Maximum Simulated Likelihood (MSL); (ii) Investigating *Cross-Validation* to substitute for *early stopping* and *Hamiltonian Monte Carlo* for *Random Walk M-H* (iii) Testing the proposed estimator on multiple modelling scenarious with multiple models.

Overall, ESBDA shows great promise as a practical, economical and relatively time-saving tool to assist in analysing choice behaviour, particularly for less wealthy or of specific population groups with lower data availability.

References

Allenby, G., 1997. An introduction to hierarchical bayesian modeling (tutorial notes, advanced research techniques forum). Chicago, IL: American Marketing Association .

Ben-Akiva, M., Bolduc, D., 1987. Approaches to model transferability and updating: the combined transfer estimator. Transp. Res. Rec. , 1–7.

Ben-Akiva, M., Bolduc, D., 1996. Multinomial Probit with a Logit Kernel and a General Parametric Specification of the Covariatice Structure. Technical Report. MIT Working Paper.

Bhat, C.R., 1997. An endogenous segmentation mode choice model with an application to intercity travel. Transportation science 31, 34–48.

Cardell, N., Reddy, B., 1977. A multinomial logit model which permits variations in tastes across individuals. Charles Rivers Associates Inc .

Greene, W.H., Hensher, D.A., 2003. A latent class model for discrete choice analysis: Contrasts with mixed logit. Transp. Res. Part B Methodol. 37, 681–698. doi:10.1016/S0191-2615(02)00046-2.

Hillel, T., 2019. Understanding travel mode choice: A new approach for city scale simulation. Ph.D. thesis. University of Cambridge.

Jazwinski, A., 1970. 1970, stochastic processes and filtering theory. new york: Academic press .

McFadden, D., Train, K., 2000. Mixed mnl models for discrete response. Journal of applied Econometrics 15, 447–470.

Precheit, L., 1998. Early Stopping - But When? Neural Networks: Tricks of the trade , 55–69doi:10.1007/3-540-49430-8.

Reich, S., Cotter, C., 2015. Probabilistic forecasting and Bayesian data assimilation. Cambridge University Press.

Rossi, P.E., McCulloch, R.E., Allenby, G.M., 1996. The value of purchase history data in target marketing. Marketing Science 15, 321–340.

Train, K.E., 2003. Discrete choice methods with simulation. Cambridge university press.

Train, K.E., 2006. Mixed Logit Estimation by Hierarchical Bayes.