Gradient-Assisted Model Specification

Stephane Hess*[1], Georges Sfeir[2], and David Bunch[3]

[1]Professor, Choice Modelling Centre, Institute for Transport Studies, University of Leeds, United Kingdom
[2]Research Fellow, Choice Modelling Centre, Institute for Transport Studies, University of Leeds, United Kingdom
[3]Professor Emeritus, Graduate School of Management, UC Davis

Short summary

This paper proposes a gradient-assisted approach to model specification in discrete choice analysis. The key idea is that, although the sample-level gradient of the log-likelihood is zero at the maximum likelihood optimum, respondent-level gradients are not, and therefore contain valuable information about underlying heterogeneity. Using a well known Stated Preference (SP) dataset, we show that these gradients reveal systematic variation in preferences and can be exploited in several ways. First, they help identify omitted systematic heterogeneity and guide improved multinomial logit specifications. Second, in mixed logit models, gradients are strongly related to posterior means and can be introduced directly into random coefficient distributions, leading to major improvements in fit and a substantial reduction in residual heterogeneity. Third, gradients provide a natural basis for clustering respondents and improving class allocation in latent class models. Across all model structures, the use of gradients yields better fit, more behaviourally meaningful heterogeneity patterns, and lower willingness-to-pay measures. The paper also highlights endogeneity concerns as an important next step for future research.
Keywords: discrete choice models; gradients; heterogeneity; model specification; willingness to pay

1 Introduction

Model specification in discrete choice modelling remains a complex and often iterative process. While analysts typically rely on prior theory, empirical intuition, and diagnostic indicators such as model fit, convergence behaviour, parameter estimates, and standard errors, the process is still largely characterised by trial and error. A key piece of information that is routinely generated during estimation but rarely exploited explicitly is the gradient of the log-likelihood function. This paper proposes a systematic framework for using gradients as a tool for model specification, interpretation of heterogeneity, and structural refinement in discrete choice models.

2 Methodology

The starting point of the proposed approach is the fundamental role of gradients in maximum likelihood estimation (MLE). Consider a sample of N individuals, each contributing a sequence of observed choices Y_n. The likelihood of observing Y_n given parameters β is denoted by $L_n(\beta)$, and the sample log-likelihood is constructed as the sum of individual contributions:

$$LL(\beta) = \sum_{n=1}^{N} \log\big(L_n(\beta)\big). \tag{1}$$

The maximum likelihood estimator is then defined as:

$$\widehat{\beta} = \arg\max_{\beta} LL(\beta). \tag{2}$$

MLE relies on first-order conditions expressed in terms of gradients of the log-likelihood function.

At any candidate parameter vector β_s, the gradient is given by:

$$g_s = \left(\frac{\partial LL(\beta)}{\partial \beta} \right)_{\beta_s}. \tag{3}$$

At the optimum, the gradient of the sample log-likelihood is equal to zero:

$$\left(\frac{\partial LL(\beta)}{\partial \beta} \right)_{\widehat{\beta}} = 0. \tag{4}$$

Crucially, however, this optimality condition applies only at the aggregate (sample) level. Because the log-likelihood is additive across individuals, the gradient can be decomposed as:

$$\frac{\partial LL(\beta)}{\partial \beta} = \sum_{n=1}^{N} \frac{\partial LL_n(\beta)}{\partial \beta}. \tag{5}$$

As a result, the first-order condition at the optimum implies:

$$\left(\frac{\partial LL(\beta)}{\partial \beta} \right)_{\widehat{\beta}} = 0, \tag{6}$$

but this does not imply that:

$$\left(\frac{\partial LL_n(\beta)}{\partial \beta} \right)_{\widehat{\beta}} = 0 \quad \forall n \,(\text{or any}). \tag{7}$$

In other words, while the sum of individual gradients is zero at the optimum, individual contributions are generally not. This distinction between sample-level and individual-level gradients is central to our analysis. It implies that different individuals exert opposing "forces" on the parameter estimates, with some pushing estimates upwards and others downwards. The MLE corresponds to the equilibrium of these opposing influences, rather than a point that is optimal for every individual.

This observation implies that the distribution of individual gradients contains rich information about underlying preference heterogeneity. Rather than being merely a computational tool used internally by optimisation algorithms, gradients can be interpreted behaviourally: they indicate how each individual would locally adjust the parameter vector to better fit their observed choices. The sample optimum is not a universal optimum across individuals, but rather the point at which opposing gradients cancel out in aggregate.

This discrepancy between sample-level optimality and individual-level gradients forms the basis of the proposed gradient-assisted specification approach. By analysing the distribution of individual gradients at the MLE, we obtain direct insights into how preferences vary across the sample. In the following sections, we show how this information can be systematically exploited to improve model specification, explain heterogeneity, and enhance behavioural interpretation across a range of discrete choice models.

3 Results

Illustration on Swiss VTT data

We illustrate the proposed approach using Swiss value-of-travel-time (VTT) data for a binary public transport choice setting. The utility specification for alternative j for respondent n in task t is written as:

$$V_{jnt} = \delta_j + \beta_{tt} TT_{jnt} + \beta_{tc} TC_{jnt} + \beta_{hw} HW_{jnt} + \beta_{ch} CH_{jnt}, \tag{8}$$

with normalisation $\delta_2 = 0$. The specification is thus linear in travel time (TT), travel cost (TC), headway (HW), and the number of interchanges (CH).

Estimation of this base multinomial logit model already yields sensible signs and strong statistical significance for most coefficients. However, the key issue for the present paper is not whether the sample-level model is acceptable in the conventional sense, but whether the gradient structure at the optimum reveals systematic information that is left unused by the analyst. Table 1 reports the estimation results together with the sample-level gradient at the optimum and the minimum and maximum values of the individual-level gradients across respondents.

Table 1: Base MNL estimation results and spread in individual-level gradients at the optimum.

	Estimate	Rob. t-rat	$g_{\hat{\beta}}$	$\min\left[\left(\frac{\partial LL_n(\beta)}{\partial \beta}\right)_{\hat{\beta}}\right]$	$\max\left[\left(\frac{\partial LL_n(\beta)}{\partial \beta}\right)_{\hat{\beta}}\right]$
δ_1	-0.0159	-0.35	-3.80E-05	-5.33	3.80
β_{tt}	-0.0598	-8.86	3.45E-03	-128.12	218.47
β_{tc}	-0.1317	-5.57	-3.91E-04	-65.13	92.30
β_{hw}	-0.0375	-16.16	1.94E-03	-123.41	109.74
β_{ch}	-1.1521	-18.77	1.30E-05	-2.65	6.57
Final log-likelihood	-1,665.62				

Table 1 makes the central point of the paper immediately visible. Even though the sample-level gradients are all essentially equal to zero at the optimum, the individual-level gradients are spread widely around zero.
This finding is illustrated more directly in Figure 1, which plots the distributions of the individual gradient contributions for the four key parameters.

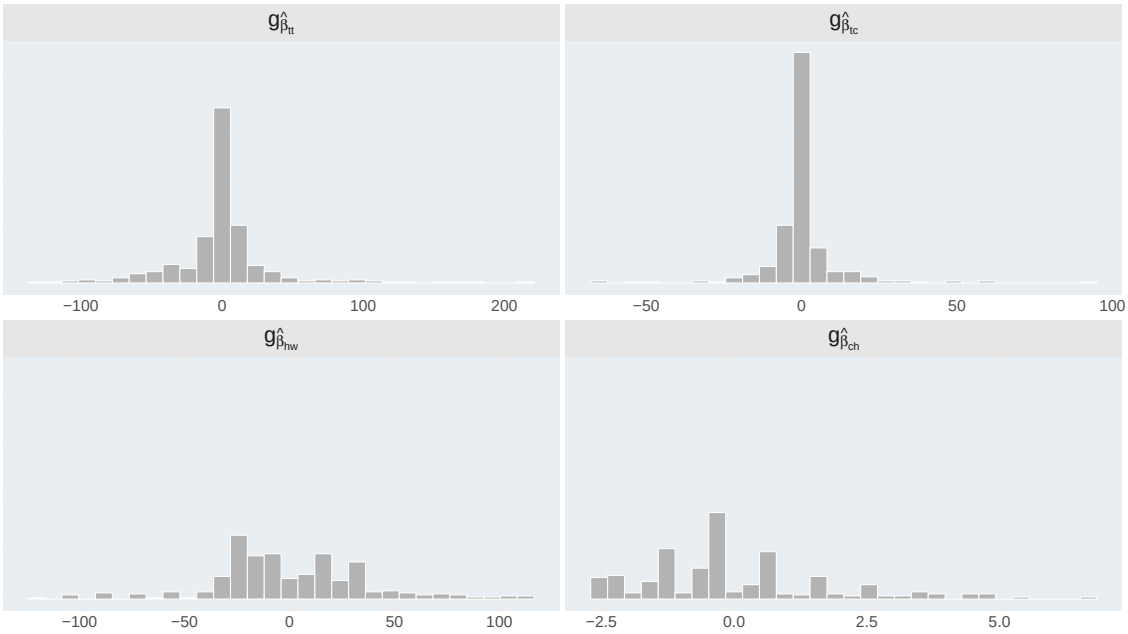


Figure 1: Histograms of respondent-level gradient contributions at the MNL optimum

The behavioural interpretation is immediate. If a given respondent has a strongly negative gradient for β_{tt} at the sample optimum, this suggests that, for that respondent, a more negative time coefficient would locally improve model fit. Conversely, a positive gradient indicates that for that respondent, a less negative coefficient would work better. Gradients therefore provide a respondent-specific directional diagnostic. Rather than simply indicating that the optimiser has converged, they describe the residual tension between the common parameter vector and the heterogeneity present in the data.

Explaining gradients and improving the MNL specification

To investigate whether these gradients are systematically related to observed characteristics, we model them using a structural equation framework. Specifically, we explain the respondent-level gradients on the basis of household income, car availability, journey purpose (commute, shopping, business, leisure), and reference levels for travel time, travel cost, headway, and interchanges. The purpose of this exercise is not merely descriptive. If gradients are systematically associated with observed covariates, then they point directly to omitted heterogeneity that can be brought back into the deterministic utility specification.
The structural equation results are reported in Table 2. The table also indicates, in brackets, the

direction in which the associated sensitivity should change (in absolute terms) according to the sign of the gradient relationship.

Table 2: Structural equation model results for respondent-level gradients. Brackets indicate the implied change in sensitivity.

Explanator	$g_{\hat{\beta}_{tt}}$			$g_{\hat{\beta}_{tc}}$			$g_{\hat{\beta}_{hw}}$			$g_{\hat{\beta}_{ch}}$		
income	-2.631			0.999	(↘)	.	-2.841			0.023		
car availability	0.471			-0.309			-4.509	(↗)	*	0.193	(↘)	*
commute	-3.223	(↗)	*	1.085	(↘)	**	-4.858	(↗)	*	0.237	(↘)	*
shopping	0.824			-0.155			-3.564	(↗)	.	0.307	(↘)	***
business	-10.109	(↗)	***	2.658	(↘)	**	1.476			0.169	(↘)	.
reference tt	31.475	(↘)	*	-17.344	(↗)	*	1.750			0.371		
reference tc	-26.276	(↗)	.	17.616	(↘)	*	-0.479			-0.053		
reference hw	2.328			0.297			-2.472			0.050		
reference ch	3.672			-0.661			-2.848			0.032		

The results in Table 2 suggest that the gradients are far from random noise. Commute and business trips are associated with stronger time sensitivity, while income, reference costs, and reference travel times shape the desired direction of cost sensitivity. Headway and interchange sensitivities are also related to car availability and trip purpose. The message is therefore clear: the gradients reveal omitted systematic heterogeneity that can be brought into the utility specification.
Guided by these results, we estimate an improved MNL model in which the taste coefficients vary systematically with observed variables. The revised utility function is:

$$V_{jnt} = \delta_j + \beta_{tt_{nt}} TT_{jnt} + \beta_{tc_{nt}} TC_{jnt} + \beta_{hw_{nt}} HW_{jnt} + \beta_{ch_{nt}} CH_{jnt}, \tag{9}$$

with $\delta_2 = 0$, where:

$$\beta_{tt_{nt}} = (\beta_{tt} + \gamma_{tt,com} z_{com,n} + \gamma_{tt,bus} z_{bus,n}) \left(\frac{ref_{tt,n}}{\overline{ref_{tt}}}\right)^{\lambda_{tt,\beta_{tt}}} \left(\frac{ref_{tc,n}}{\overline{ref_{tc}}}\right)^{\lambda_{tc,\beta_{tt}}}, \tag{10}$$

$$\beta_{tc_{nt}} = (\beta_{tc} + \gamma_{tc,com} z_{com,n} + \gamma_{tc,bus} z_{bus,n}) \left(\frac{inc_n}{\overline{inc}}\right)^{\lambda_{inc,\beta_{tc}}} \left(\frac{ref_{tt,n}}{\overline{ref_{tt}}}\right)^{\lambda_{tt,\beta_{tc}}} \left(\frac{ref_{tc,n}}{\overline{ref_{tc}}}\right)^{\lambda_{tc,\beta_{tc}}}, \tag{11}$$

$$\beta_{hw_{nt}} = \beta_{hw} + \gamma_{hw,car\,av} z_{car\,av,n} + \gamma_{hw,com} z_{com,n} + \gamma_{hw,bus} z_{bus,n}, \tag{12}$$

$$\beta_{ch_{nt}} = \beta_{ch} + \gamma_{ch,car\,av} z_{car\,av,n} + \gamma_{ch,com} z_{com,n} + \gamma_{ch,shop} z_{shop,n} + \gamma_{ch,bus} z_{bus,n}. \tag{13}$$

The corresponding estimation results are shown in Table 3. Several of the effects suggested by the gradient analysis become directly significant in the improved specification.

Table 3: Improved MNL model informed by gradient analysis.

	Left-hand block						Right-hand block				
	Estimate	Rob. t	Sig.	Grad. sens. ch.	Grad. sig.		Estimate	Rob. t	Sig.	Grad. sens. ch.	Grad. sig.
δ_1	-0.0339	-0.7				β_{hw}	-0.0304	-8.9	***		
β_{tt}	-0.0832	-7.1	***			$\gamma_{hw,car\,av}$	-0.0108	-2.2	*	(↗)	*
$\gamma_{tt,com}$	-0.0609	-2.1	*	(↗)	*	$\gamma_{hw,com}$	-0.0190	-3.2	***	(↗)	*
$\gamma_{tt,bus}$	-0.0673	-3.3	***	(↗)	***	$\gamma_{hw,bus}$	-0.0133	-1.6	.	(↗)	.
$\lambda_{tt,\beta_{tt}}$	0.0506	0.1		(↘)	*	β_{ch}	-1.3402	-13.1	***		
$\lambda_{tc,\beta_{tt}}$	-0.3944	-0.9		(↗)	.	$\gamma_{ch,car\,av}$	0.1998	1.5	.	(↘)	*
β_{tc}	-0.3091	-7.2	***			$\gamma_{ch,com}$	-0.0940	-0.6		(↘)	*
$\gamma_{tc,com}$	-0.0747	-0.9		(↘)	**	$\gamma_{ch,shop}$	0.3341	1.7	*	(↘)	***
$\gamma_{tc,bus}$	0.0607	0.8		(↘)	**	$\gamma_{ch,bus}$	0.1503	0.7		(↘)	.
$\lambda_{inc,\beta_{tc}}$	-0.0522	-0.7		(↘)	.						
$\lambda_{tt,\beta_{tc}}$	1.1253	1.9	*	(↗)	*						
$\lambda_{tc,\beta_{tc}}$	-1.7482	-3.1	***	(↘)	*						

The improved specification leads to a sizeable increase in model fit, as shown in Table 4. The log-likelihood improves from −1,665.62 to −1,541.65, and the likelihood ratio test strongly rejects the base model. At the same time, the implied sample-level valuations fall substantially.
The drop in WTP indicators is substantively important. It suggests that a specification that ignores gradient-based evidence on heterogeneity can overstate the average valuation of service improvements. Put differently, the gradients do not merely improve fit; they alter behavioural interpretation.

Table 4: Comparison between the base and improved MNL models.

	Base model	Improved model
LL	-1,665.62	-1,541.65
pars	5	21
LR-test p-value	1.369×10^{-43}	
Sample level value of time (CHF/hr)	27.21	18.64
Sample level value of headway (CHF/hr)	17.06	6.88
Sample level value of interchanges (CHF/ch)	8.75	3.67

Gradient information and the shape of continuous heterogeneity

The next question is whether gradients can also inform the treatment of unobserved heterogeneity in mixed logit models. We begin with a standard MMNL specification:

$$V_{jnt} = \delta_j + \beta_{tt_n} TT_{jnt} + \beta_{tc_n} TC_{jnt} + \beta_{hw_n} HW_{jnt} + \beta_{ch_n} CH_{jnt}, \tag{14}$$

with

$$\beta_{k_n} = -e^{\mu_{\log \beta_k} + \sigma_{\log \beta_k} \xi_{k_n}}, \qquad \xi_{k_n} \sim N(0,1). \tag{15}$$

This specification allows each respondent to have their own sensitivity, drawn from a lognormal distribution to preserve the expected negative sign for disutility coefficients.
The corresponding estimation results are given in Table 5.

Table 5: Base MMNL model with lognormally distributed coefficients.

	Estimate	Rob. t-rat
δ_1	-0.0415	-0.59
$\mu_{\log \beta_{tt}}$	-1.9852	-17.15
$\sigma_{\log \beta_{tt}}$	0.4734	8.73
$\mu_{\log \beta_{tc}}$	-1.0219	-5.49
$\sigma_{\log \beta_{tc}}$	1.0034	9.02
$\mu_{\log \beta_{hw}}$	-2.9328	-32.17
$\sigma_{\log \beta_{hw}}$	0.8055	10.26
$\mu_{\log \beta_{ch}}$	0.6278	7.75
$\sigma_{\log \beta_{ch}}$	0.8459	8.67
Final LL	-1,442.29	
MNL LL	-1,665.62	
LR-test p-value	2.29×10^{-95}	

The base MMNL model strongly improves on the MNL model, as expected. However, the key insight from this model is that the respondent-level gradients from the simple MNL are highly correlated with the posterior means from the MMNL. Figure 2 illustrates this relationship.
This suggests a further refinement: instead of treating gradients merely as an ex post diagnostic, they can be incorporated directly into the random coefficient specification as covariates. We therefore estimate:

$$V_{jnt} = \delta_j + \beta_{tt_n} TT_{jnt} + \beta_{tc_n} TC_{jnt} + \beta_{hw_n} HW_{jnt} + \beta_{ch_n} CH_{jnt}, \tag{16}$$

with

$$\beta_{k_n} = -e^{\mu_{\log \beta_k} + \sigma_{\log \beta_k} \xi_{k_n} + \tau_k g_{\hat{\beta}_k}}, \qquad \xi_{k_n} \sim N(0,1). \tag{17}$$

Here, the gradient term $\tau_k g_{\hat{\beta}_k}$ shifts the position of each respondent within the coefficient distribution as a function of the respondent-specific gradient from the base MNL model.
The estimation results for this gradient-augmented MMNL model are reported in Table 6.
The improvement in fit is striking. More importantly, several of the standard deviations shrink sharply and, for headway and interchanges, are essentially reduced to zero. This implies that once

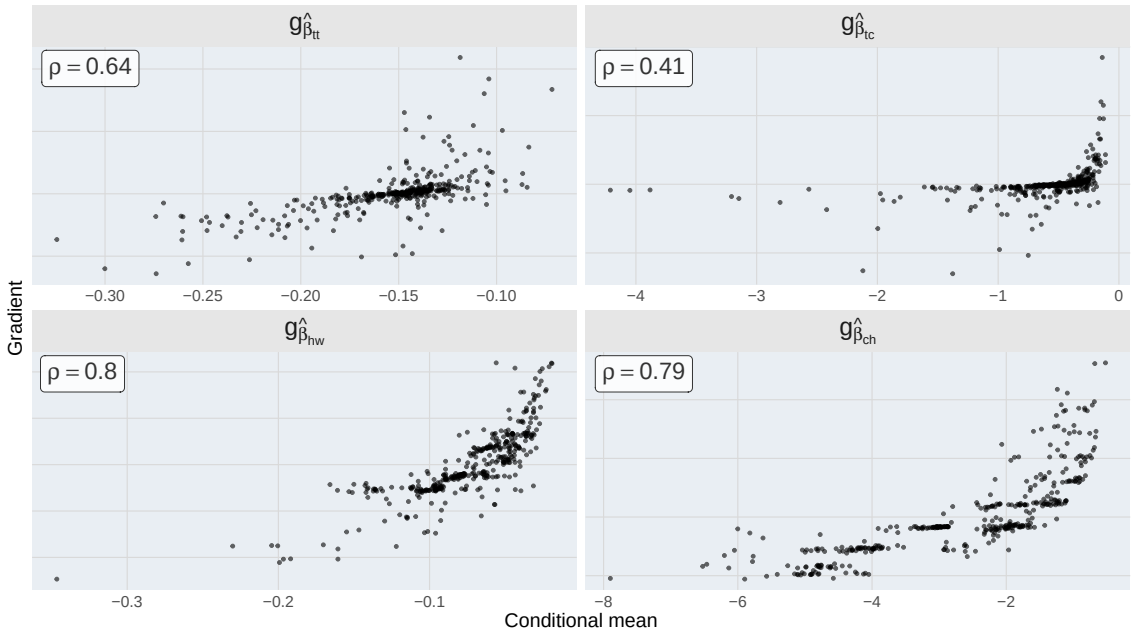


Figure 2: Relationship between MNL respondent-level gradients and MMNL posterior means. The strong correlation indicates that gradients contain direct information about the location of respondents on the parameter distribution.

Table 6: MMNL model with gradients included as covariates in the random coefficient distributions.

	Estimate	Rob. t-rat
δ_1	-0.0308	-0.46
$\mu_{\log \beta_{tt}}$	-1.9805	-20.85
$\sigma_{\log \beta_{tt}}$	0.1057	1.72
τ_{tt}	-0.0064	-3.81
$\mu_{\log \beta_{tc}}$	-0.9094	-7.12
$\sigma_{\log \beta_{tc}}$	-0.5757	-6.41
τ_{tc}	-0.0225	-5.29
$\mu_{\log \beta_{hw}}$	-2.9483	-53.19
$\sigma_{\log \beta_{hw}}$	-0.0006	-0.10
τ_{hw}	-0.0148	-14.85
$\mu_{\log \beta_{ch}}$	0.5248	11.97
$\sigma_{\log \beta_{ch}}$	0.0009	0.21
τ_{ch}	-0.5207	-15.00
Final LL	-1,037.87	
Base MMNL LL	-1,442.29	
LR-test p-value	9.344×10^{-174}	

gradients are introduced explicitly, a large share of what was previously being captured by random heterogeneity is now explained deterministically through the gradient term. Figure 3 shows that the correlation between conditionals and gradients becomes even stronger in this augmented model. The effect on the implied coefficient distributions is shown in Figure 4.

These distributional changes translate into substantial differences in willingness to pay. Table 7 compares the base and improved MMNL models.

Again, the gradient-informed model produces markedly lower valuations and lower dispersion. Figure 5 further explores what happens if residual randomness is removed altogether.

The corresponding comparison is reported in Table 8. Even when moving to a simpler model with greatly reduced or eliminated residual random heterogeneity, the loss in fit is modest relative to

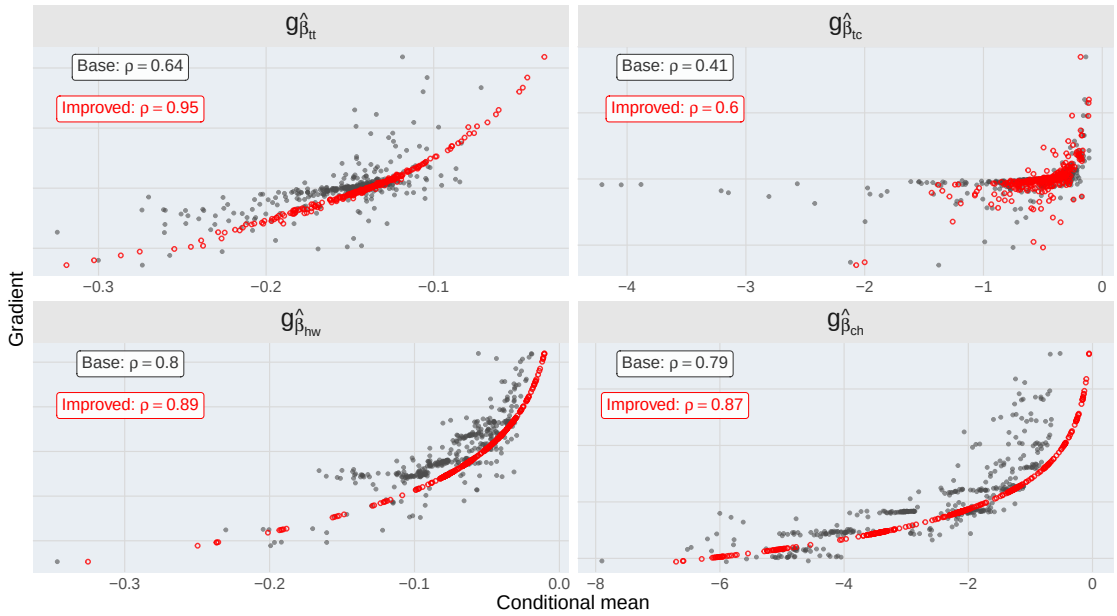


Figure 3: Relationship between conditionals and gradients after incorporating gradients directly into the MMNL specification

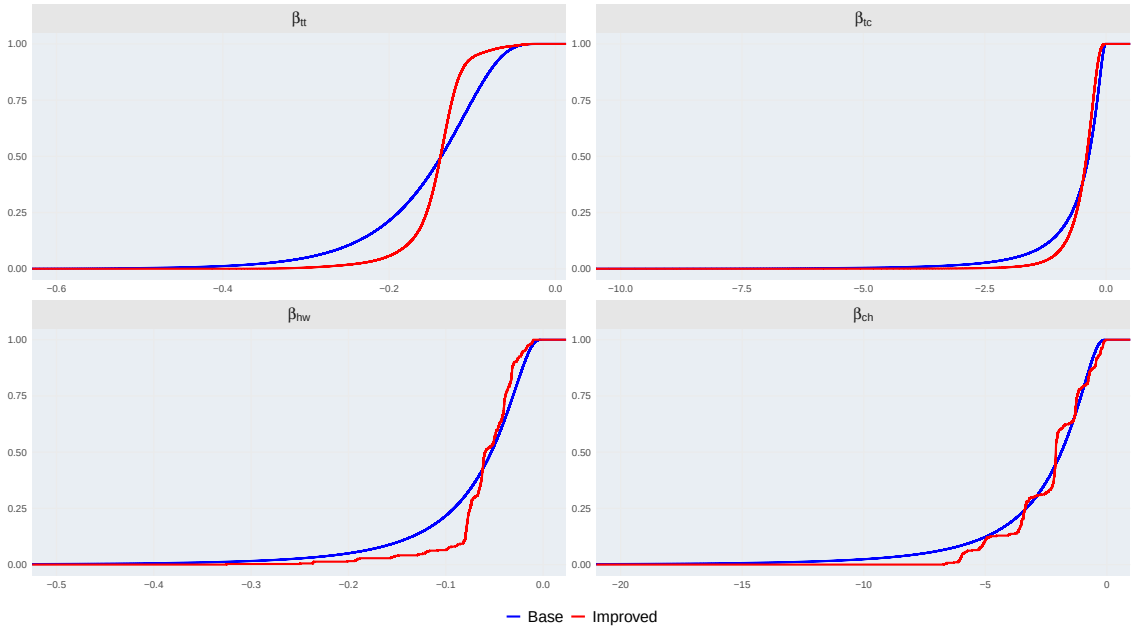


Figure 4: Cumulative distribution functions for the random coefficients before and after incorporating gradients

Table 7: WTP measures for the base and gradient-augmented MMNL models.

	Base model		Improved model	
	Mean	Std. dev.	Mean	Std. dev.
Value of time (CHF/hr)	42.38	66.01	28.01	33.78
Value of headway (CHF/hr)	20.33	43.18	11.15	12.30
Value of interchanges (CHF/ch)	12.31	26.11	6.88	7.88

the large gain achieved through the inclusion of gradients.

Taken together, these results support a strong conclusion: respondent-level gradients explain a

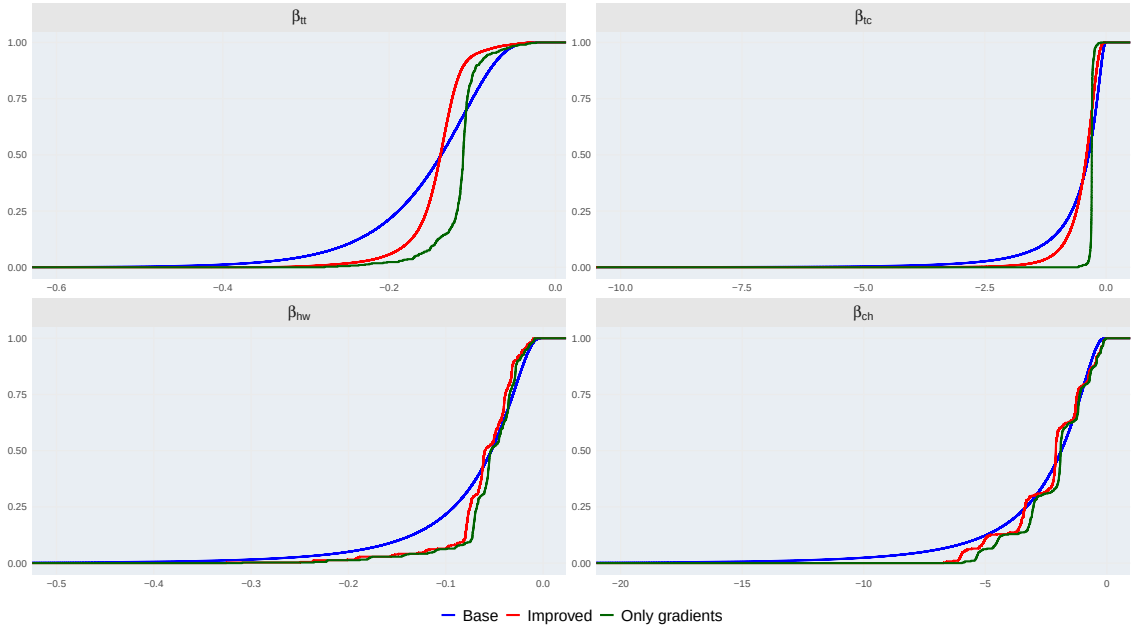


Figure 5: Illustration of the implied distributions when residual randomness is reduced or removed

Table 8: Comparison of WTP distributions and fit across three MMNL-type specifications

	Model 1		Model 2		Model 3	
	Mean	SD	Mean	SD	Mean	SD
Value of time (CHF/hr)	42.38	66.01	28.01	33.78	24.33	10.68
Value of headway (CHF/hr)	20.33	43.18	11.15	12.30	11.29	7.30
Value of interchanges (CHF/ch)	12.31	26.11	6.88	7.88	7.05	4.77
Final LL	-1,442.29		-1,037.87		-1,063.34	

large part of the heterogeneity that mixed logit models typically capture through broad continuous distributions. They therefore provide not only a diagnostic tool, but also a structural route towards more parsimonious and behaviourally transparent specifications.

Clustering gradients for latent class modelling

A further application of the same idea arises in latent class models. We begin with standard latent class specifications with two and three classes. The alternative-specific constant is kept generic across classes, and class allocation does not initially use covariates:

$$\pi_c = \frac{e^{\kappa_c}}{\sum_{s=1}^{S} e^{\kappa_s}}, \qquad \kappa_1 = 0. \tag{18}$$

The two-class and three-class results are shown in Tables 9 and 10. Both models improve substantially on the base MNL model, and they reveal strong segmentation in valuations.
To connect these latent classes with the gradient information, we compute expected class-specific gradients as:

$$\hat{g}_{\hat{\beta}_{tt},c} = \frac{\sum_{n=1}^{N} \hat{\pi}_{c,n} \hat{g}_{\hat{\beta}_{tt},n}}{\sum_{n=1}^{N} \hat{\pi}_{c,n}}, \tag{19}$$

and similarly for the other coefficients. The results, summarised in Table 11, show that high-VTT classes tend to have expected negative gradients for time and positive gradients for cost, consistent with the interpretation developed earlier.
This result motivates a clustering exercise directly on the respondent-level gradients. Since gradients are available for each respondent and each coefficient, they define a natural feature space for

Table 9: Baseline two-class latent class model

	Two-class model			
	Class 1		Class 2	
	Estimate	Rob. t-rat	Estimate	Rob. t-rat
β_{tt}	-0.0628	-5.93	-0.2804	-3.23
β_{tc}	-0.0877	-4.41	-1.9030	-4.83
β_{hw}	-0.0432	-10.80	-0.0511	-4.62
β_{ch}	-1.0440	-9.48	-2.4868	-4.82
κ_c	0	–	-0.8216	-4.27
δ_1 (generic)	-0.0523	-1.03	idem	
π_c	0.69		0.31	
Value of time (CHF/hr)	42.98		8.84	
Value of headway (CHF/hr)	29.59		1.61	
Value of interchanges (CHF/ch)	11.91		1.31	
Final log-likelihood	-1,551.94			
MNL log-likelihood	-1,665.62			
LR-test p-value	3.936×10^{-47}			

Table 10: Baseline three-class latent class model

	Three-class model					
	Class 1		Class 2		Class 3	
	Estimate	Rob. t-rat	Estimate	Rob. t-rat	Estimate	Rob. t-rat
β_{tt}	-0.2284	-3.95	-0.0480	-5.75	-0.3392	-4.18
β_{tc}	-1.8362	-5.38	-0.0524	-3.39	-1.0368	-4.29
β_{hw}	-0.0471	-3.00	-0.0410	-7.87	-0.0689	-7.10
β_{ch}	-1.7191	-4.00	-0.7618	-5.86	-3.9239	-7.59
κ_c	0	–	0.9139	3.56	0.6745	2.38
δ_1 (generic)	-0.0383	-0.67	idem		idem	
π_c	0.18		0.46		0.36	
Value of time (CHF/hr)	7.46		54.95		19.63	
Value of headway (CHF/hr)	1.54		46.95		3.99	
Value of interchanges (CHF/ch)	0.94		14.54		3.78	
Final log-likelihood	-1,489.94					
Two-class log-likelihood	-1,551.94					
B&S test p-value	5.90×10^{-28}					

unsupervised segmentation. We therefore apply K-means clustering. Figures 6, 7, and 8 present the principal-component visualisations for two and three clusters, together with the silhouette analysis suggesting that a fourth cluster adds little additional structure.

Rather than using these clusters deterministically, we incorporate the standardised distance from each respondent to each centroid as a covariate in the class allocation model:

$$\pi_{n,c} = \frac{e^{\zeta_c D_{n,c}}}{\sum_{s=1}^{S} e^{\zeta_s D_{n,s}}}. \tag{20}$$

This maintains probabilistic class membership while allowing gradient-based segmentation to inform the allocation process.

The resulting two-class and three-class models are reported in Tables 12 and 13. The distance effects are significant, and the fit improvements over the baseline latent class models are sizeable. Finally, Table 14 shows that the gradient-informed class allocation improves the alignment between expected income and class-specific value of time. This is important because one persistent challenge

Table 11: Class-specific valuations and expected gradients in the baseline latent class models.

	Two-class model		Three-class model		
	Class 1	Class 2	Class 1	Class 2	Class 3
Value of time (CHF/hr)	42.98	8.84	7.46	54.95	19.63
Value of headway (CHF/hr)	29.59	1.61	1.54	46.95	3.99
Value of interchanges (CHF/ch)	11.91	1.31	0.94	14.54	3.78
$\hat{g}_{\hat{\beta}_{tt},c}$	-6.80	15.46	25.09	-7.78	-2.89
$\hat{g}_{\hat{\beta}_{tc},c}$	2.58	-5.87	-8.06	4.12	-1.13
$\hat{g}_{\hat{\beta}_{hw},c}$	-5.59	12.71	11.48	-8.47	4.91
$\hat{g}_{\hat{\beta}_{ch},c}$	0.31	-0.70	0.01	0.80	-1.02

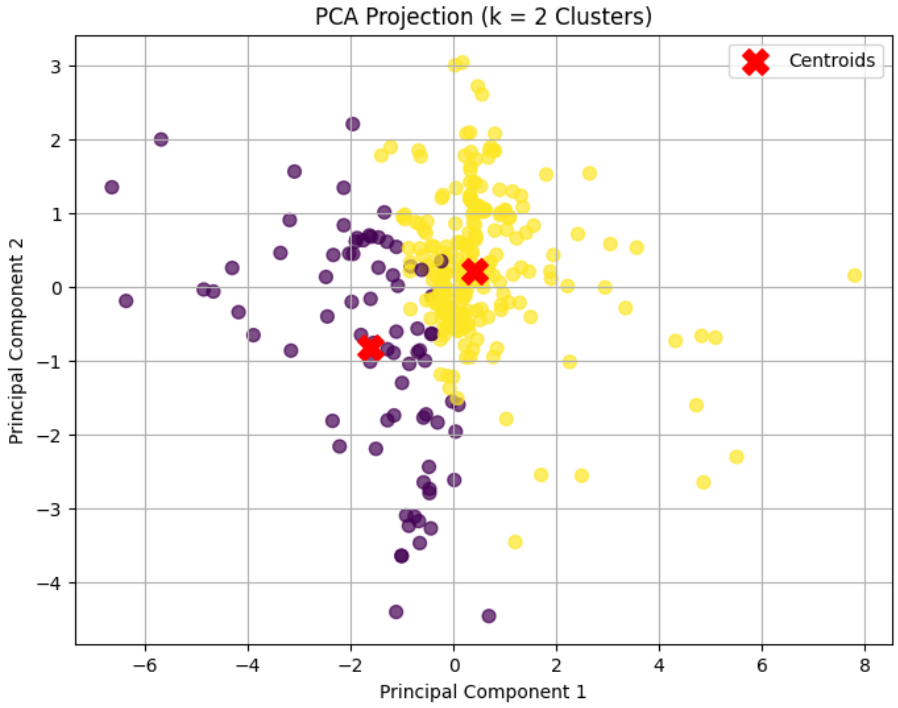


Figure 6: Principal component projection of respondent-level gradients under a two-cluster K-means solution.

in latent class models is that the behavioural interpretation of the classes may not line up well with observed socioeconomic patterns. The gradient-based approach improves that linkage.

Summary of findings

Across all three modelling families, the evidence points in the same direction. First, respondent-level gradients from a simple MNL model reveal strong and systematic heterogeneity that is invisible in the aggregate first-order conditions. Second, this information can be linked to observed covariates and used to guide richer deterministic specifications. Third, the same gradients are strongly related to posterior means in mixed logit models and can be used directly to explain variation that would otherwise be captured by broad continuous random distributions. Fourth, gradients provide a powerful basis for clustering respondents and improving latent class allocation.

A consistent empirical result across the paper is that gradient-informed models not only fit substantially better, but also produce lower willingness-to-pay measures. This applies to the improved MNL model, the gradient-augmented MMNL model, and the gradient-informed latent class models. Figure 9 summarises the valuation findings visually.

The overarching conclusion is therefore that gradients should not be treated merely as internal

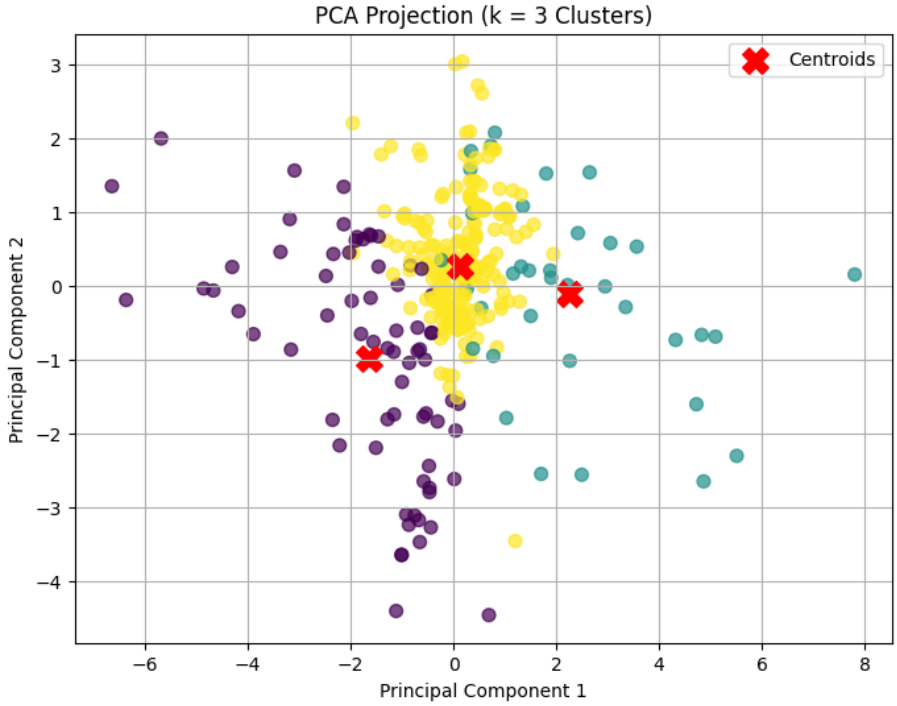


Figure 7: Principal component projection of respondent-level gradients under a three-cluster K-means solution.

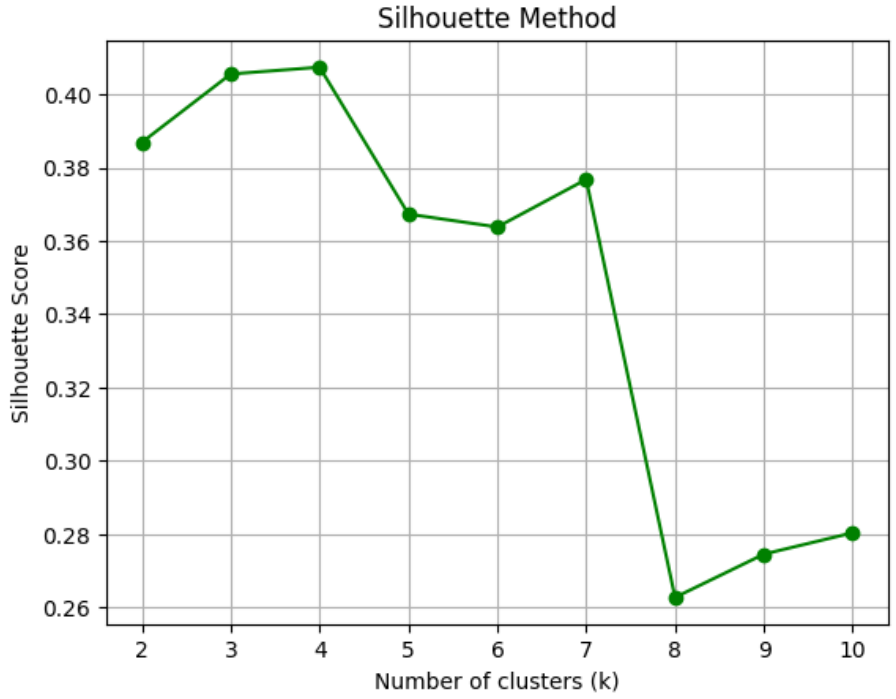


Figure 8: Silhouette analysis for gradient-based clustering. The results suggest that moving beyond three clusters brings only limited gains in separation.

optimisation objects. They are a rich behavioural signal. At the sample optimum, they tell us which respondents are insufficiently represented by the common parameter vector and in which direction that mismatch operates. Exploiting this information leads to better fitting models, more transparent representations of heterogeneity, and more moderate implied valuations. In this sense, gradient-assisted specification offers a systematic complement to the traditional trial-and-error approach to discrete choice model development.

Table 12: Two-class latent class model with gradient-based distance covariates in class allocation.

	Two-class model			
	Class 1		Class 2	
	Estimate	Rob. t-rat	Estimate	Rob. t-rat
β_{tt}	-0.0751	-5.59	-0.0783	-7.36
β_{tc}	-0.3160	-5.39	-0.0852	-5.11
β_{hw}	-0.0451	-10.39	-0.0348	-7.02
β_{ch}	-1.9061	-14.58	-0.3490	-4.45
ζ_c	5.7964	3.00	7.1158	3.29
δ_1 (generic)	-0.0407	-0.79	idem	
π_c	0.73		0.27	
Value of time (CHF/hr)	14.26		55.12	
Value of headway (CHF/hr)	8.56		24.50	
Value of interchanges (CHF/ch)	6.03		4.09	
Final log-likelihood	-1,425.26			
Base LC log-likelihood	-1,551.94			

Table 13: Three-class latent class model with gradient-based distance covariates in class allocation.

	Three-class model					
	Class 1		Class 2		Class 3	
	Estimate	Rob. t-rat	Estimate	Rob. t-rat	Estimate	Rob. t-rat
β_{tt}	-0.1067	-6.49	-0.1431	-12.07	-0.0280	-3.44
β_{tc}	-0.5167	-5.65	-0.2147	-6.52	-0.0689	-2.40
β_{hw}	-0.0312	-5.50	-0.0666	-14.28	-0.0239	-5.56
β_{ch}	-2.6366	-12.02	-1.7386	-16.07	0.0530	0.80
ζ_c	8.5355	3.06	7.6575	3.08	12.2611	2.78
δ_1 (generic)	-0.0397	-0.70	idem		idem	
Value of time (CHF/hr)	12.39		40.00		24.36	
Value of headway (CHF/hr)	3.62		18.61		20.85	
Value of interchanges (CHF/ch)	5.10		8.09		-0.77	
Final log-likelihood	-1,316.26					
Base LC log-likelihood	-1,489.94					

Table 14: Expected income and value of time across classes, with and without gradient-based class allocation.

		Two-class model		Three-class model		
		Class 1	Class 2	Class 1	Class 2	Class 3
2*Without gradients	Value of time (CHF/hr)	42.98	8.84	7.46	54.95	19.63
	Expected income (CHF/annum)	77,389.04	74,503.50	72,074.93	75,344.80	80,243.30
2*With gradients	Value of time (CHF/hr)	14.26	55.12	12.39	40.00	24.36
	Expected income (CHF/annum)	76,211.34	77,252.86	74,524.48	78,401.65	75,095.67

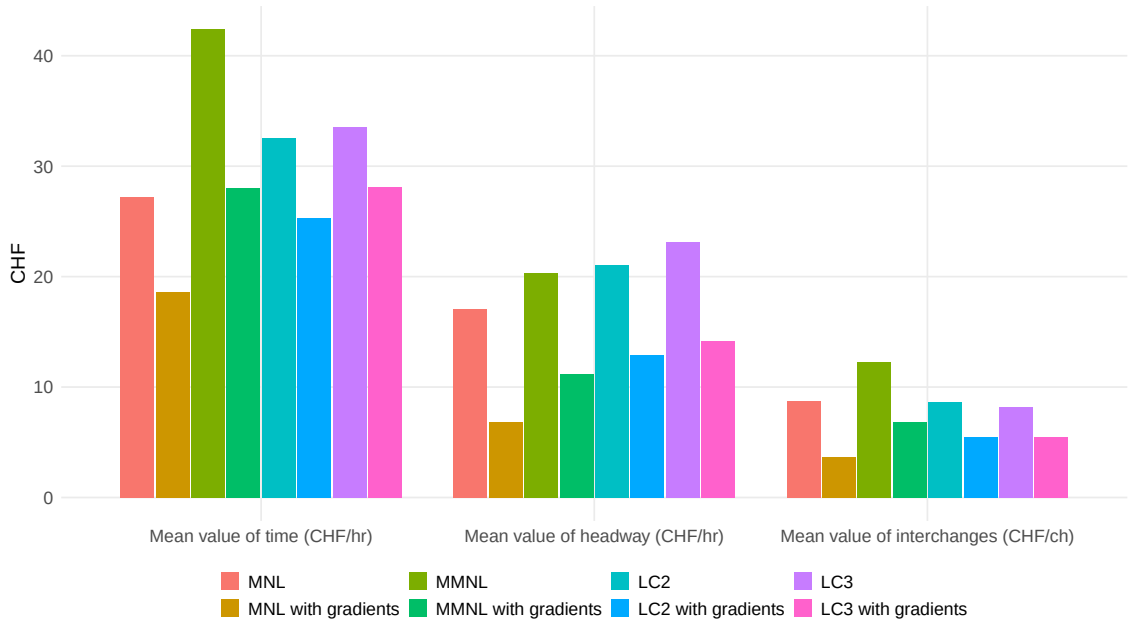


Figure 9: Comparison of implied valuations across model specifications. Incorporating gradient information consistently lowers WTP measures relative to baseline specifications.

4 Conclusion

This paper demonstrates that gradients of the log-likelihood function provide a powerful and largely untapped source of behavioural information for discrete choice modelling. While traditionally used only for optimisation, gradients reveal how individual observations locally "pull" parameter estimates in different directions, offering direct insight into underlying heterogeneity. Across MNL, MMNL, and latent class models, we show that exploiting this information leads to substantial improvements in model fit, more behaviourally meaningful segmentation, and systematically lower willingness-to-pay estimates.

Importantly, gradients help bridge the gap between observed and unobserved heterogeneity, allowing part of what is typically captured through random parameters or latent classes to be explained in a more structured and interpretable way

At the same time, we acknowledge potential concerns regarding endogeneity, as gradients are derived from the same data for which they are then used to help inform the model specification. Addressing this issue is a key priority for our future work. In particular, we plan to explore the use of instrumental variables, as well as hybrid modelling approaches in which gradients are treated as indicators rather than explanatory variables. These extensions will help establish a more robust theoretical and econometric foundation for gradient-assisted specification.