Loughborough University

This item was submitted to Loughborough's Research Repository by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

The multinomial logit model revisited: a semi-parametric approach in discrete choice analysis

PLEASE CITE THE PUBLISHED VERSION

http://dx.doi.org/10.1016/j.trb.2010.09.007

PUBLISHER

© Elsevier

VERSION

AM (Accepted Manuscript)

LICENCE

CC BY-NC-ND 4.0

REPOSITORY RECORD

Li, Baibing. 2019. “The Multinomial Logit Model Revisited: A Semi-parametric Approach in Discrete Choice Analysis”. figshare. https://hdl.handle.net/2134/9235.

Loughborough University

This item was submitted to Loughborough's Institutional Repository (https://dspace.lboro.ac.uk/) by the author and is made available under the following Creative Commons Licence conditions.

creative commons

COMMONS DEED

Attribution-NonCommercial-NoDerivs 2.5

You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor.

Noncommercial. You may not use this work for commercial purposes.

No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/

The Multinomial Logit Model Revisited:

A Semiparametric Approach in Discrete Choice Analysis

Baibing Li
Business School
Loughborough University
Loughborough LE11 3TU, United Kingdom

ABSTRACT

The multinomial logit model in discrete choice analysis is widely used in transport research. It has long been known that the Gumbel distribution forms the basis of the multinomial logit model. Although the Gumbel distribution is a good approximation in some applications such as route choice problems, it is chosen mainly for mathematical convenience. This can be restrictive in many other scenarios in practice. In this paper we show that the assumption of the Gumbel distribution can be substantially relaxed to include a large class of distributions that is stable with respect to the minimum operation. The distributions in the class allow heteroscedastic variances. We then seek a transformation that stabilizes the heteroscedastic variances. We show that this leads to a semiparametric choice model which links the linear combination of travel-related attributes to the choice probabilities via an unknown sensitivity function. This sensitivity function reflects the degree of travelers' sensitivity to the changes in the combined travel cost. The estimation of the semiparametric choice model is also investigated and empirical studies are used to illustrate the developed method.

Keywords: Discrete choice model; Gumbel distribution; Multinomial logit model; Semiparametric model; Variance stabilization

---

Tel: 44-1509 228841. *E-mail address: b.li2@lboro.ac.uk*

1. Introduction

The multinomial logit model is widely used in a variety of transport-related choice contexts. Compared with the other choice models, the multinomial loigt model is particularly attractive in many modeling scenarios due to the nature that it is linked to the decision-making behavior via the maximising (minimising) the utility (cost). In the derivation of the closed-form multinomial logit model, there are three underlying assumptions, i.e., the random variables of interest are assumed (a) to be independent of each other; (b) to have an equal variability across cases; and (c) to follow the Gumbel (Type I extreme value) distribution (McFadden, 1978; Ben-Akiva and Lerman, 1985; Train, 2003; Bhat et al., 2008; Koppelman, 2008). In practice, these assumptions may be violated in many choice contexts. To address this issue, much attention has been paid to the relaxation of these assumptions in the last three decades. Because these assumptions are related to each other to some extents, relaxing one assumption may affect the others (see e.g. Castillo et al., 2008; Fosgerau and Bierlaire, 2009). In general, however, the extensions of the multinomial logit model may be classified into two different categories: open-form and closed-form choice models.

Closed-form choice models have several advantages over open-form models. They are usually simpler both conceptually and computationally. Consequently it is usually easier to specify a closed-form model and interpret the obtained results. In this paper, we will focus on closed-form models. See Bhat et al. (2008) for an overview of various open-form choice models developed in recent years.

Several important approaches were developed in the 1970s to increase the flexibility of the multinomial logit model by relaxing the assumption on the independence of alternative outcomes while still retaining the choice models in a closed form. They include the nested logit model and a more general approach: the generalized extreme value (GEV) family

(McFadden, 1978). Since the 1970s, this area has attracted a large number of researchers and many useful models have been proposed, such as paired combinatorial logit (PCL), cross-nested logit (CNL), and generalized nested logit (GNL). These approaches allow dependence or correlation among the random variables by relaxing the cross-elasticity restrictions. See Train (2003) and Koppelman (2008) for recent overviews.

Relaxation of the equality of the error variance structure across cases has been investigated by Swait and Adamowicz (1996), Bhat (1997), and many others. Swait and Adamowicz (1996) developed the heteroscedastic multinomial logit (HMNL) model that allows the random error variances to be non-identical across individuals/cases. On the other hand, Bhat (1997) proposed the covariance heterogeneous nested logit model (COVNL). The COVNL model was developed on the basis of the nested logit model and it allows heterogeneity across cases in the covariance of nested alternatives.

Now we turn to the assumption on the functional form of the underlying distributions. Lee (1983) in his pioneering work explored relaxing the assumption of the underlying distributions by an arbitrary *pre-specified* distribution. Recently Castillo et al. (2008) have proposed using the Weibull distribution as an alternative to the Gumbel distribution to derive a multinomial choice model. They show that the Weibull distribution may provide a better approximation for some route choice problems than the Gumbel in practice. Further they demonstrate that if the random variables for different alternatives follow the Weibull distribution, then a closed-form expression for the choice probabilities can be obtained from the utility-maximizing behavior. Furthermore, Fosgerau and Bierlaire (2009) show that the assumption of the Weibull distribution is associated with the discrete choice model having multiplicative error terms, and the log-transformation links the multiplicative model to the additive model for which the Gumbel distribution is assumed. In addition, Castillo et al. (2008) and Fosgerau and Bierlaire (2009) find that the Weibull-distribution-based model

allows random variables with heteroscedastic variances. As a consequence, performing the log-transformation can achieve two goals simultaneously: on the one hand it stabilizes variances, and on the other it specifies the Weibull distribution as the underlying distribution instead of the conventional Gumbel distribution.

From a practical perspective, assuming a particular functional form such as the Gumbel or Weibull distribution for the underlying distribution of random variables for different alternatives is restrictive in many applications. This is because discrete choice analysis is used in a variety of the problems in transport research (Bhat et al., 2008). It is hard to believe that a single statistical distribution can accommodate such a variety of applications. In this paper we shall present empirical evidence that the actual underlying distribution indeed differs from both the Gumbel and Weibull distributions in some applications.

In addition, as demonstrated in this paper, the assumption of underlying distributions for different alternatives is linked to a sensitivity function which reflects how sensitive a traveler is to the changes in a linear combination of travel-related attributes such as travel time, travel expenses, etc. Consequently, specifying an underlying distribution implicitly stipulates a sensitivity function. Empirical results in this paper show that people may have different sensitivities to the same amount of change when using different transportation modes. Hence the issue of travelers' sensitivity has to be taken into consideration during modeling.

The purpose of this paper is to extend the Weibull-distribution results obtained by Castillo et al. (2008) and Fosgerau and Bierlaire (2009) to a more general situation. We will show that to derive a closed-form discrete choice model from the cost-minimization (or utility-maximizing) behavior, the actual functional form of the underlying distributions does not have to be explicitly pre-specified provided that they belong to a certain class of distributions that is stable with respect to the minimum operation. The model with an unspecified underlying distribution allows researchers considerable flexibility in model

specification, which is particularly important in practice because the discrete choice model is applied in different areas.

We will show that the proposed distribution family allows heteroscedastic variances. Hence, it has also relaxed the assumption of homoscedastic variances made for the multinomial logit model. We will seek a transformation that stabilizes the heteroscedastic variances of the underlying distributions. We will show that this leads to a semiparametric choice model which links the linear combination of travel-related attributes to the choice probabilities via an unknown sensitivity function. We will also investigate the estimation of the unknown sensitivity function and discuss practical implications of the sensitivity function.

This paper is organised as follows. In the next section the assumption of the Gumbel distribution is relaxed to include a large class of distributions. In Section 3 a semiparametric choice model is investigated. Section 4 is devoted to the estimation of the unknown sensitivity function and the coefficients of the attributes in the semiparametric choice model. The developed method is illustrated in Section 5 using two datasets from Danish value-of-time study. Finally discussion and conclusions are given in Section 6.

2. Underlying distributions in discrete choice analysis

2.1. A stable class of distributions with respect to the minimum operation

Discrete choice models may be investigated in various transport-related contexts. In this paper we consider this problem from the perspective of individual choice behavior where a traveler wishes to minimize his/her travel cost among several alternatives (routes, transportation modes etc.). Note that for the problems of random utility maximization, the results can be applied straightforwardly by considering the corresponding negative utilities.

Let C_n denote the feasible choice set of each individual n (n=1,…,N) and ξ_{in} denote the random travel cost for traveler n when choosing alternative i. We assume throughout this paper that the random costs ξ_{in} ($i \in C_n$ and for all n) are independent of each other.

Rather than assuming a particular distribution (such as the Gumbel or Weibull) for the random costs ξ_{in}, we suppose that the distributions of ξ_{in} are from a large class of distributions with the following functional form of cumulative distribution function (CDF):

$$F_{in}(t) = \Pr\{\xi_{in} < t\} = 1 - [1 - F(t)]^{\alpha_{in}}, \quad (1)$$

where the base distribution function $F(t)$ is left unspecified. The parameters α_{in} are assumed to be associated with individual alternative i and traveler n. From the perspective of statistical inference, the assumption that the random costs ξ_{in} follow any distribution from distribution family (1) with an unspecified base function $F(t)$ allows researchers great flexibility to accommodate different problems. Table 1 displays some special cases of distributions from this distribution family.

(Table 1 is here)

Let V_{in} and σ_{in}^2 denote the expectations and variances of ξ_{in}, i.e., $E\xi_{in} = V_{in}$ and $\text{var}(\xi_{in}) = \sigma_{in}^2$. The variances σ_{in}^2 may depend on the expectations V_{in} so in general they are heteroscedastic. We suppose that the expectations V_{in} are linked to a linear function of a q-vector of attributes $\mathbf{x}_{in}$ (usually including attributes for alternative i as viewed by traveler n and characteristics of traveler n such as income, gender and age) that influences specific discrete outcomes:

$$V_{in} = \mathbf{x}_{in}^T \boldsymbol{\beta}, \quad (2)$$

where **β** is a vector of parameters to be estimated. As Fosgerau and Bierlaire (2009) have pointed out, if the coefficient of the travel expenses is normalized to one unit, then other coefficients in vector **β** can be interpreted as willingness-to-pay indicators.

Now we show that distribution family (1) is closed under the minimum operation. Suppose that the random costs ξ_{in} follow the distribution $F_{in}(t) = 1 - [1 - F(t)]^{\alpha_{in}}$ for any base function $F(t)$. Under the assumption of independence we have

$$\Pr\{\min_{i\in C_n} \xi_{in} < t\} = 1 - \Pr\{\min_{i\in C_n} \xi_{in} \geq t\}$$

$$= 1 - \prod_{i\in C_n} \Pr\{\xi_{in} \geq t\} = 1 - \prod_{i\in C_n}\{1 - F_{in}(t)\} = 1 - [1 - F(t)]^{\alpha_{0n}},$$

where $\alpha_{0n} = \sum_{i\in C_n} \alpha_{in}$. Hence, the minimum cost $\min_{i\in C_n} \xi_{in}$ belongs to the same distribution family as the individual random costs ξ_{in} ($i \in C_n$ and for all n) do.

We also note that under the assumption of independence, any distribution family that is closed under the minimum operation must have the functional form given in (1). Hence the family (1) is the most general class that is stable with respect to the minimum operation. As shown later in Section 3, the stability with respect to the minimum operation is crucial for the derivation of choice probabilities.

2.2 Variance-stabilizing transformations

In general, for a given base function $F(t)$, the variances of the distributions in family (1), $\text{var}(\xi_{in}) = \sigma_{in}^2$, are heteroscedastic (see, e.g., Table 1). Hence, the distribution family (1) does not restrict the random costs to be homoscedastic. In this subsection we show that the variances can be stabilized via a suitable transformation. First we state a theorem that links distribution family (1) to the Gumbel distribution.

Theorem 1. Suppose that random variables X_j (j=1,…,m) have the following CDFs:

$$\Pr\{X_j < x\} = 1 - [1 - F(x)]^{\alpha_j} \quad \text{with} \quad \alpha_j > 0 \quad (j=1,\ldots,m),$$

where $F(x)$ is any chosen CDF. Then $h(x) = \theta^{-1}\log\{-\log[1 - F(x)]\}$ is a monotonically increasing transformation and the transformed random variables $Z_j = h(X_j)$ follow the Gumbel distribution $G(z;\theta,\alpha_j) = 1 - \exp[-\alpha_j \exp(\theta z)]$ with a common scale parameter θ.

From Theorem 1, $h(.)$ transforms ξ_{in} to Gumbel-distributed variates $\eta_{in} = h(\xi_{in})$ with CDFs of

$$G(t;\theta,\alpha_{in}) = 1 - \exp[-\alpha_{in}\exp(\theta t)]\,. \tag{3}$$

The means and variances of η_{in} are given by $E\eta_{in} = -\{\log(\alpha_{in}) + \gamma\}/\theta$ and $\text{var}(\eta_{in}) = \pi^2/(6\theta^2)$ respectively, where γ is the Euler constant.

Theorem 1 also indicates that $h(t) = \theta^{-1}\log\{-\log[1 - F(t)]\}$ transforms heteroscedastic variances $\text{var}(\xi_{in}) = \sigma_{in}^2$ to a constant value $\pi^2/(6\theta^2)$. Consequently $h(t)$ is a variance-stabilizing transformation. Hence, the relaxation of the Gumbel distribution to the distribution family (1) is also linked to the relaxation of the second assumption for multinomial logit models, i.e., homoscedastic variances. Castillo et al. (2008) and Fosgerau and Bierlaire (2009) have considered a special case where the transformation function $h(t)$ is taken as the log-transformation for variance stabilization. Table 2 displays $h(t)$ for some commonly used distributions.

(Table 2 is here)

In practice, variance stabilization is an important issue and has been investigated intensively in the literature. A general asymptotic variance-stabilizing transformation for a random variable X with a mean of μ and variance $\sigma^2(\mu)$ can be shown to be $\tilde{h}(t) =$

$\int^t \sigma^{-1}(\mu)d\mu$ (see, e.g., Tibshirani, 1988). For the exponential distribution, for instance, $\sigma^2(\mu) = \mu^2$ (see Table 1). Hence the asymptotic variance-stabilizing transformation for the exponential distribution is $\tilde{h}(t) = \log(t)$ which is identical to $h(t)$, up to a scale/level constant.

In general, however, $\tilde{h}(.)$ and $h(.)$ differ from each other. Unlike the asymptotic variance-stabilizing transformation $\tilde{h}(.)$, $h(.)$ is an accurate transformation for variance stabilization for the distribution family (1).

2.3 Identifiability

Any discrete choice model must address the issue of identifiability since the level and scale of utility are irrelevant (Ben-Akiva and Lerman, 1985; Train, 2003). When the variance-stabilizing transformation $h(t)$ is replaced by $a + bh(t)$ with two constants a and $b > 0$, the Gumbel distribution (3) is replaced by:

$$G(t; \theta, \alpha_{in}) = 1 - \exp\left[-\alpha_{in}\exp\left(\theta t/b - \theta a/b\right)\right]$$

with mean $-\{\log(\alpha_{in}) + \gamma\}b/\theta + a/b$ and variance $b^2\pi^2/(6\theta^2)$ respectively. Hence, the transformation function $h(t)$ is not uniquely defined. In practice, for identification purposes, some restrictions on the level constant and scale constant have to be imposed to ensure that $h(t)$ is identifiable. We shall return to this issue in Sections 3 and 4.

2.4 The mean function

In this subsection we will derive a mean function that links the means before and after the transformation. Let $h^{-1}(t)$ denote the inverse function of $h(t)$, i.e., $h^{-1}(t) = F^{-1}\{1 -$

exp [−exp (θt)]}, where $F^{-1}(t)$ is the inverse function of $F(t)$. Then the expectations of ξ_{in} may be evaluated as follows:

$$V_{in} = E(\xi_{in}) = \int h^{-1}(t)\, dG(t; \theta, \alpha_{in}). \quad (4)$$

Theorem 2. Let $F(t)$ be any chosen CDF and $h(t) = \theta^{-1}\log\{-\log[1 - F(t)]\}$. Then for the random variables X_j having the CDFs given by $\Pr\{X_j < x\} = 1 - [1 - F(x)]^{\alpha_j}$ with $\alpha_j > 0$, the expectations $E(X_j) = \int h^{-1}(t)\, dG(t; \theta, \alpha_j)$ are monotonically decreasing functions of α_j (j=1,…,m).

From Theorem 2, an implicit mean function $H(.)$ is derived from the variance-stabilizing transformation $h(.)$:

$$\alpha_{in} = H(V_{in}), \quad (5)$$

where $H(.) > 0$ is monotonically decreasing. This has established a link between the parameter α_{in} and the expectation V_{in} of a random variable ξ_{in}. Since $E\eta_{in} = -\{\log(\alpha_{in}) + \gamma/\theta$ and θ is constant, $H(.)$ captures the relationship between the two means obtained before and after transformation $h(.)$ is applied. In the case of the exponential distribution $F_{in}(t) = 1 - \exp\{-\alpha_{in}t\}$, for instance, we have $H(.) = 1/t$.

In general, the relationship between $H(.)$ and $h(.)$ can be complicated. Specifically, let $H^{-1}(t)$ be the inverse function of $H(.)$. From the proof of Theorem 2, we have

$$H^{-1}[\exp(-\theta s)] = \theta \int h^{-1}(t) \exp[\theta(t - s)] \exp\{-\exp[\theta(t - s)]\}\, dt.$$

Hence, $H^{-1}[\exp(-\theta t)]$ is a convolution of $h^{-1}(t)$ and the density function of the Gumbel distribution $g(t; \theta, 1) = \theta \exp(\theta t) \exp[-\exp(\theta t)]$. Clearly under some mild conditions $H(.)$ is uniquely determined by $h(.)$, and vice versa. The complexity of the relationship between $H(.)$ and $h(.)$ can be seen from the case of the Type II logistic

distribution $F_{in}(t) = 1 - [1 + \exp(t)]^{-\alpha_{in}}$ with a mean of $V_{in} = \psi(1) - \psi(\alpha_{in})$, where $\psi(x)$ is the first derivative of the function $\log\Gamma(x)$ and $\Gamma(x)$ is the gamma function. By defining $\psi^{-1}(x)$ as the inverse function of $\psi(x)$, we obtain $H(t) = \psi^{-1}(\psi(1) - t)$, which does not have a simple link to the corresponding $h(t) = \theta^{-1}\log\{\log[1 + \exp(t)]\}$. Table 2 displays the mean function $H(t)$ for some commonly used distributions.

Under certain conditions there exists a simple approximate relation between the two functions $h(t)$ and $H(t)$. Specifically, noting that η_{in} follow the Gumbel distribution (3) with $E\eta_{in} = -\{\log(\alpha_{in}) + \gamma\}/\theta$, we have

$$\alpha_{in} = \exp\{-\gamma - \theta E(\eta_{in})\} = \exp\{-\gamma)\,/\exp\,[\theta E(\eta_{in})].$$

By approximating $E(\eta_{in}) = E[h(\xi_{in})]$ by $h[E(\xi_{in})] = h(V_{in})$, we obtain $\alpha_{in} \approx \exp\{-\gamma)\,/\exp\,[\theta h(V_{in})]$. Hence the mean function $H(t)$ can be approximated by $1/\exp\,[\theta h(t)]$, up to a constant. In addition, noting that $h(t) = \theta^{-1}\log\{-\log[1 - F(t)]\}$, we can further obtain an approximate relationship between the mean function $H(t)$ and the unspecified base distribution $F(t)$:

$$H(t) \approx 1/\{-\log\,[1 - F(t)]\},$$

up to a constant. Note that for the exponential distribution, the above relation holds exactly, i.e., $H(t) = 1/\{-\log\,[1 - F(t)]\}$.

3. A semiparametric choice model

3.1. A semiparametric single-index choice model

In this section we show that the assumption of distribution family (1) leads to a semi-parametric single-index choice model.

According to the theory of individual choice behavior (see, e.g., Ben-Akiva and Lerman, 1985), the probability that any alternative i in C_n is chosen by traveler n is $P_n(i) = \Pr\{\xi_{in} \leq \xi_{jn}, \forall j \in C_n \text{ and } j \neq i\}$. Then from the total probability theorem, we obtain:

$$P_n(i) = \Pr\{\xi_{in} \leq \min_{j \in C_n, j \neq i} \xi_{jn}\} = \int \Pr\{\min_{j \in C_n, j \neq i} \xi_{jn} \geq t | \xi_{in} = t\} \, dF_{in}(t).$$

Since $F_{in}(t) = \Pr\{\xi_{in} < t\} = \Pr\{\eta_{in} < h(t)\} = 1 - \exp[-\alpha_{in} \exp(\theta h(t))]$, we have

$$dF_{in}(t) = \theta \alpha_{in} \exp\{-\alpha_{in} \exp[\theta h(t)]\} \exp[\theta h(t)] dh(t).$$

Hence, by defining $\alpha_{0n} = \sum_{j \in C_n} \alpha_{jn}$, we obtain

$$P_n(i) = \theta \alpha_{in} \int \exp\{-\alpha_{0n} \exp[\theta h(t)]\} \exp[\theta h(t)] dh(t) = \frac{\alpha_{in}}{\alpha_{0n}} = \frac{H(V_{in})}{\sum_{j \in C_n} H(V_{jn})}.$$

When the expectations $E\xi_{in} = V_{in}$ are linked to a linear function of a q-vector of attributes $\mathbf{x}_{in}$ via equation (2), this leads to the following choice model:

$$P_n(i) = \frac{H(\mathbf{x}_{in}^T \boldsymbol{\beta})}{\sum_{j \in C_n} H(\mathbf{x}_{jn}^T \boldsymbol{\beta}))}. \quad (6)$$

Clearly equation (6) generalizes the multinomial logit model by using the unknown mean function $H(.)$ to replace the exponential function. In addition, although the random costs of interest have heteroscedastic variances as assumed in equation (1), the variances are stabilized via $h(.)$ so that the scale parameter in (6) is constant across all alternatives and travelers. This scale parameter is absorbed into $H(.)$ so it is not identifiable. Hence, extending the multinomial logit model by allowing an unspecified functional form $H(.)$ can address both the issue of nonlinearity in the mean function and the issue of variance stabilization.

Equation (6) belongs to semiparametric single-index models. In statistics and econometrics, a model is termed a single index model if it only depends on the vector $\mathbf{x}$ through a single linear combination, i.e. $\mathbf{x}^T \boldsymbol{\beta}$. In a semiparametric single index model, the model depends on $\mathbf{x}$ through an unknown function $H(.)$, i.e. $H(\mathbf{x}^T \boldsymbol{\beta})$ (see, e.g., Stoker 1986;

Ichimura, 1993). Note also that in semiparametric single-index models, there is only one nonparametric dimension, thus these methods fall into the class of dimension reduction techniques. Consequently although both $\boldsymbol{\beta}$ and $H(.)$ are unknown, only $H(.)$ is nonparametrically estimated.

In recent years attention has been paid to semiparametric approaches in the transport literature. For instance, Fosgerau (2006) has investigated the distribution of the value of travel time savings, and Fosgerau and Bierlaire (2007) have investigated mixing distributions in discrete choice analysis, both using a semiparametric approach.

The semiparametric single index model (6) is a special case of the more general nonparametric choice model investigated in Huang and Nychka (2000) where $H(\mathbf{x}^T\boldsymbol{\beta})$ is further extended to a general nonlinear function $H(x_1, \dots, x_q)$ of q variables. From a computational perspective, a major advantage of semiparametric single-index models is that they avoid the so-called "curse of dimensionality" by reducing the nonparametric dimensionality from q to one.

Due to the issue of identifiability of $H(.)$ and $\boldsymbol{\beta}$, it is required in this paper that the linear combination of attributes $\mathbf{x}^T\boldsymbol{\beta}$ does not include an intercept, and that $\boldsymbol{\beta}$ has unit length and one of its entry (say the first one) has a positive sign. Following Ichimura (1993), some further conditions need to be imposed for the identification of $H(.)$ and $\boldsymbol{\beta}$. In particular $H(.)$ is required not to be constant on the support of $\mathbf{x}^T\boldsymbol{\beta}$. The vector of attributes $\mathbf{x}$ should also admit at least one continuously distributed component. See Ichimura (1993) for details.

3.2. Sensitivity function

From equation (5) the mean function $H(.)$ is non-negative. Now define $H(t) = \exp\{S(t)\}$ so that the range of $S(t)$ is the whole real line. Equation (6) becomes

$$P_n(i) = \frac{\exp\{S(\mathbf{x}_{in}^T\boldsymbol{\beta})\}}{\sum_{j\in C_n}\exp\{S(\mathbf{x}_{jn}^T\boldsymbol{\beta})\}}. \quad (7)$$

In this paper $S(t)$ is termed sensitivity function. It reflects how sensitive a traveler is to the changes in the combined travel cost (including travel time, travel expenses, etc.). Table 2 displays the sensitivity function for some commonly used distributions. When the sensitivity function is linear, $S(t) = -\theta t$ with a scalar parameter $\theta > 0$, model (7) reduces to the ordinary multinomial logit model and the corresponding underlying distribution is the Gumbel. A linear sensitivity function thus provides a benchmark for comparison. This is illustrated in Figure 1. For the sensitivity function represented by the dotted line in Figure 1, for instance, travelers are more sensitive to one unit increment in the combined travel cost in the area where the combined travel cost is high. In contrast, the sensitivity function of the broken line represents the scenario where travelers are more tolerable to the increment in the combined travel cost. In the multiplicative choice model developed in Castillo et al. (2008) and Fosgerau and Bierlaire (2009), the logarithm sensitivity function is used. It is worth noting that for the log-function, there not exist a point with respect to which it is symmetric. Hence it is suitable to such a scenario where travelers are more sensitive to one extreme end of the combined travel cost but less sensitive to the other.

(Figure 1 is about here)

From a practical perspective, a very important issue is model selection: how do we discriminate among several competing choice models, including the ordinary multinomial logit model where no transformation is applied, the multiplicative choice model with the log-transformation, and the more general semi-parametric model (7)? In this paper we incorporate the well-known deviance information criterion (DIC) as a measure of goodness-

of-fit for model comparison. The DIC is a hierarchical modeling generalization of the AIC (Akaike information criterion) and BIC (Bayesian information criterion). Similar to AIC and BIC, it takes into consideration both the model accuracy and the degree of model parsimony. See Spiegelhalter (2002) for the details on the DIC.

3.3. Further extensions

The multinomial logit model is frequently used as a building block in discrete choice analysis to handle more complex scenarios. Potentially the semiparametric choice model could also be combined with some existing approaches in discrete choice analysis to deal with complicated scenarios in practice. A thorough investigation for such extensions is beyond the scope of this paper and would admit a separate paper. In this subsection, we simply demonstrate how it can be combined with an existing approach, mixed multinomial logit model.

The mixed logit is a generic approach that is conceptually appealing and also computationally efficient. It can be derived by allowing a random-coefficients structure (Train, 2003; Bhat et al., 2008). Consequently it can address the issue of heterogeneity across travelers and does not exhibit the property of independence from irrelevant alternatives (IIA).

Now for the semiparametric choice model, we follow Train (2003) and Bhat et al. (2008), and assume that the coefficients vary across travelers in the population with density $q(\boldsymbol{\beta})$ so that the heterogeneity across travelers can be taken into account. For each traveler, however, it is assumed that the semiparametric choice probability $L_{ni}(\boldsymbol{\beta}) = \frac{\exp\{S(\mathbf{x}_{in}^T\boldsymbol{\beta})\}}{\sum_{j\in C_n}\exp\{S(\mathbf{x}_{jn}^T\boldsymbol{\beta})\}}$ still holds. Since the researcher observes $\mathbf{x}_{in}$ but not $\boldsymbol{\beta}$, the unconditional choice probability is the integral of over all possible variable of $\boldsymbol{\beta}$:

$$P_n(i) = \int L_{ni}(\boldsymbol{\beta})q(\boldsymbol{\beta})\,d\boldsymbol{\beta}.$$

A commonly used model for $q(\boldsymbol{\beta})$ is the normal distribution (Train, 2003). Fosgerau and Bierlaire (2007) have investigated a practical test for the choice of the mixing distribution.

It is of interest to compare the mixed logit and the above mixed semiparametric choice model. In the existing mixed logit model, the heterogeneity is modeled solely by the mixing distribution because the ordinary multinomial logit assumes homogeneity across observations. In contrast, in the above mixed semiparametric choice model, the heterogeneity across alternatives and the heterogeneity across travelers are dealt with separately: the former is addressed via the variance-stabilizing transformation $h(.)$, whereas the latter is modeled by the mixing distribution $q(\boldsymbol{\beta})$. Since different sources of variability are modeled separately, it is more straightforward for model specification and interpretation in the mixed semiparametric choice model.

It is also worth noting that although the semiparametric choice model has relaxed, to some extents, the assumptions (b) and (c) as mentioned in Section 1, it still retains the IIA property which may sometimes be restrictive in practice. The above mixed semiparametric model, however, does not exhibit the IIA property and thus is more flexible to accommodate the nature of complicated problems in practice.

4. Bayesian inference

In this section we investigate the estimation of the unknown sensitivity function $S(t)$ and coefficient vector $\boldsymbol{\beta}$. Ichimura (1993), Horowitz (2001), and Fosgerau (2006) investigated statistical inference for semiparametric models using the kernel density estimation method. On the other hand, Fosgerau and Bierlaire (2007) investigated approximating the unknown mixing distribution using Legendre polynomials as a basis.

In this paper, we use P-splines to approximate $S(t)$ and perform Bayesian analysis to draw statistical inference. The approach of P-splines has been widely used in statistics in recent years. The idea of the P-splines is quite similar to that used in Fosgerau and Bierlaire (2007), i.e. B-splines are used as the basis functions to approximate an unknown function of interest.

4.1. The Bayesian P-splines approach

The P-splines approach was developed by Eilers and Marx (1996). Compared to smoothing splines, the P-splines approach usually leads to a more parsimonious parameterization. Lang and Brezger (2004) have recently considered Bayesian inference for additive nonparametric regression models using the P-splines approach. The major advantage of using a Bayesian approach, rather than the maximum likelihood method, is that it is still applicable even if the sample size in an analysis is relatively small (Gelman et al., 2003). In addition, the smoothing parameter can be determined straightforwardly in the Bayesian analysis. In contrast, the smoothing parameter has to be determined via cross-validation when using the maximum likelihood method.

In the P-splines approach in Eilers and Marx (1996), it is assumed that an unknown function $S(t)$ can be approximated by splines of degree l with $r+1$ equally spaced knots over the domain of t. The unknown function $S(t)$ is written in terms of a linear combination of $m = r + l$ B-spline basis functions $B_j(t)$ (j=1,…,m), i.e. $S(t) = \sum_{j=1}^{m} \omega_j B_j(t) = \mathbf{B}^T(t)\boldsymbol{\omega}$, where $\mathbf{B}^T(t) = [B_1(t), \dots, B_m(t)]$ and $\boldsymbol{\omega}^{\mathbf{T}} = [\omega_1, \dots, \omega_m]$. Since the vector of the basis functions $\mathbf{B}(t)$ is given and fixed, the estimation of the unknown function $S(t)$ reduces to the estimation of the unknown parameter vector $\boldsymbol{\omega}$. See De Boor (1978) for the details of B-splines.

Now we focus on the estimation for model (7). As mentioned earlier, it is assumed that vector $\boldsymbol{\beta}$ has unit length with a positive sign of its first entry due to the issue of identifiability. In addition, during the estimation, $S(t)$ is expressed as $S(u+vt)$ so that $S(t)$ has a fixed support, say on [0, 1], where *u* and *v* are two scaling parameters.

Let y_{in} be 1 if traveler *n* chose alternative *i* and 0 otherwise (*n*=1,...,*N*). Let $\mathbf{X}$ and $\mathbf{Y}$ denote the data matrices comprising $\mathbf{x}_{in}$ and y_{in} for all *i* and *n*. Let $t_{in} = \mathbf{x}_{in}^T\boldsymbol{\beta}$. Then the likelihood is

$$L(\mathbf{Y};\boldsymbol{\beta},\boldsymbol{\omega},\mathbf{X}) = \prod_{n=1}^{N}\prod_{i\in C_n}\{P_i(i)\}^{y_{in}} = \prod_{n=1}^{N}\prod_{i\in C_n}\left\{\frac{\exp\{\mathbf{B}^T(t_{in})\boldsymbol{\omega}\}}{\sum_{j\in C_n}\exp\{\mathbf{B}^T(t_{jn})\boldsymbol{\omega}\}}\right\}^{y_{in}} .$$

Eilers and Marx (1996) suggested a moderately large number of knots in B-splines approximation to ensure enough flexibility, and defined a roughness penalty based on differences of adjacent B-spline coefficients to guarantee sufficient smoothness of the fitted curves. This leads to a penalised log-likelihood given by $\log L(\mathbf{Y};\boldsymbol{\beta},\boldsymbol{\omega},\mathbf{X}) - \boldsymbol{\omega}^T\mathbf{K}\boldsymbol{\omega}/(2\tau^2)$, where τ^2 is a smoothing parameter to be determined, and $\mathbf{K}$ is a given penalty matrix. $\mathbf{K}$ is chosen as a symmetric tri-diagonal matrix whose main diagonal is 1, 2, 2, …, 2, 2, 1. The diagonal entries immediately above and below the main diagonal are all equal to -1 (Eilers and Marx, 1996; Lang and Brezger, 2004).

In Bayesian analysis, the penalty term $-\boldsymbol{\omega}^T\mathbf{K}\boldsymbol{\omega}/(2\tau^2)$ can be treated as a prior of $\boldsymbol{\omega}$ for given τ: $p(\boldsymbol{\omega}|\tau) \propto \exp\{-\boldsymbol{\omega}^T\mathbf{K}\boldsymbol{\omega}/(2\tau^2)\}$. The prior of the smoothing parameter τ^2 is usually assumed to follow an inverse gamma distribution $IG(a,b)$: $p(\tau^2) \propto (\tau^2)^{-(a+1)}\exp\{-b/\tau^2\}$, where *a* and *b* are two hyper-parameters (Lang and Brezger, 2004). The prior of $\boldsymbol{\beta}$ is chosen as non-informative: $p(\boldsymbol{\beta}) \propto 1$. Combining the likelihood and the priors, the posterior distribution is given by

$$p(\boldsymbol{\beta},\boldsymbol{\omega},\tau|\mathbf{Y}) \propto L(\mathbf{Y};\boldsymbol{\beta},\boldsymbol{\omega},\mathbf{X})\, p(\boldsymbol{\beta})p(\boldsymbol{\omega}|\tau)\, p(\tau^2).$$

4.2. Markov chain Monte Carlo simulation

In this subsection, we investigate using Markov chain Monte Carlo (MCMC) simulation to draw samples from the posterior distribution of the unknown parameters, $p(\boldsymbol{\beta}, \boldsymbol{\omega}, \tau|\mathbf{Y})$. The algorithm used below is a mixture of the Gibbs sampler and Metropolis-Hasting algorithm where the simulation is carried out block-wise, drawing each of $\boldsymbol{\beta}$, $\boldsymbol{\omega}$ and τ in turn. See Gelman et al. (2003, Chapter 11) for an overview on MCMC.

Specifically, let $\boldsymbol{\beta}^c$, $\boldsymbol{\omega}^c$ and τ^c be the values of the parameters in the current iteration. The initial guess of $\boldsymbol{\beta}$, $\boldsymbol{\beta}^0$, can be obtained using the ordinary multinomial logit model and then is normalized to ensure it has unit length with a positive first entry. For given $\boldsymbol{\beta}^0$, the initial guess of $\boldsymbol{\omega}$, $\boldsymbol{\omega}^0$, can be obtained by maximising the penalised likelihood.

In each subsequent iteration, it is easy to show that, for given $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ at the current iteration, i.e. $\boldsymbol{\beta}^c$ and $\boldsymbol{\omega}^c$, the full conditional distribution of τ^2 is an inverse gamma distribution, $IG(\tilde{a}, \tilde{b})$ with $\tilde{a} = a + rank(\mathbf{K})/2$ and $\tilde{b} = b + \boldsymbol{\omega}^{cT}\mathbf{K}\boldsymbol{\omega}^c/2$. Hence, τ can be drawn straightforwardly.

However, parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ cannot be drawn directly from their conditional posterior distributions. Hence, the Metropolis-Hasting algorithm is used below to draw vectors $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$.

First, we consider sampling $\boldsymbol{\beta}$ for fixed $\boldsymbol{\omega}^c$ and τ^c. We draw a proposal $\boldsymbol{\beta}^p$ from the proposal distribution $N(\boldsymbol{\beta}^c, \rho_1^2\boldsymbol{\Sigma}_1)$, where ρ_1 is a tuning parameter and $\boldsymbol{\Sigma}_1$ is a diagonal matrix whose diagonal entries are the squared values of the entries of $\boldsymbol{\beta}^0$. The sampling of proposal $\boldsymbol{\beta}^p$ from the proposal distribution can be carried out block-wise or component-wise. Then $\boldsymbol{\beta}^p$ is normalized to ensure it has unit length with a positive first entry. We calculate the acceptance rate as follows:

$$r_1 = \min\{1, p(\boldsymbol{\beta}^p, \boldsymbol{\omega}^c, \tau^c|\mathbf{Y})/p(\boldsymbol{\beta}^c, \boldsymbol{\omega}^c, \tau^c|\mathbf{Y})\}.$$

The proposal $\boldsymbol{\beta}^p$ is accepted with probability r_1. $\boldsymbol{\beta}^c$ is then replaced by $\boldsymbol{\beta}^p$ if $\boldsymbol{\beta}^p$ is accepted; otherwise $\boldsymbol{\beta}^c$ remains unchanged.

Similarly, we can draw a sample of $\boldsymbol{\omega}$ for fixed $\boldsymbol{\beta}^c$ and τ^c. We draw a proposal $\boldsymbol{\omega}^p$ from the proposal distribution $N(\boldsymbol{\omega}^c, \rho_2^2\boldsymbol{\Sigma}_2)$, where ρ_2 is a tuning parameter and $\boldsymbol{\Sigma}_2$ is a diagonal matrix whose diagonal entries are the squared values of the entries of $\boldsymbol{\omega}^0$. Note that to ensure the support of the function $S(t)$ is on [0, 1], $t_{in} = \mathbf{x}_{in}^T\boldsymbol{\beta}$ is scaled via two scalar parameters u and v such that $u + vt_{in} = u + v\mathbf{x}_{in}^T\boldsymbol{\beta} \in [0, 1]$ for all i and n. The sampling of proposal $\boldsymbol{\omega}^p$ can be carried out block-wise or component-wise. We then calculate the acceptance rate as follows:

$$r_2 = \min\{1, p(\boldsymbol{\beta}^c, \boldsymbol{\omega}^p, \tau^c|\mathbf{Y})/p(\boldsymbol{\beta}^c, \boldsymbol{\omega}^c, \tau^c|\mathbf{Y})\}.$$

The proposal $\boldsymbol{\omega}^p$ is accepted with probability r_2. $\boldsymbol{\omega}^c$ is then replaced by $\boldsymbol{\omega}^p$ if $\boldsymbol{\omega}^p$ is accepted; otherwise $\boldsymbol{\omega}^c$ remains unchanged.

Our numerical experience shows that the acceptance rates r_1 and r_2 for $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ are quite high by appropriately choosing the tuning parameters.

5. Empirical applications

Fosgerau et al. (2006) carried out a large-scale Danish value-of-time study comprising several surveys, two of which involved single mode public transport experiments where interchanges between two vehicles in the experiments were restricted to be of the same type, trains or buses. To illustrate the developed method, we apply the developed method in this section to analyze these two datasets. The two datasets involved stated preferences about two train-related alternatives and two bus-related alternatives respectively. Travel time for public transport users in the study was broken down into four components: (a) access/egress time (other modes than public transport, including walking, cycling, taxi, etc.); (b) in-vehicle time;

(c) headway of the first used mode; and (d) interchange waiting time. The attributes considered in their study included these four travel time components, plus the number of interchanges and travel expenses. The travelers' time values were inferred from binary alternative routes characterised by these attributes. Fosgerau and Bierlaire (2009) also analyzed these two datasets using the multiplicative choice model.

Throughout this section, we set l=3 and r=4, thus the B-splines used in the following analyses included seven cubic basis functions $B_j(t)$ (j=1,…,7) on the support [0, 1]. Following the suggestion in Lang and Brezger (2004), we chose $a = b = 0.001$ for the prior of τ^2. The total number of iterations in the MCMC simulation was set as 10,000. The first 5,000 iterations were considered as burnt-in period and the corresponding draws were discarded. The results are reported below using the remaining 5,000 draws. Following Gelman et al. (2003), we calculate the posterior means (used as estimates), posterior standard deviations, and 95% credible intervals of the unknown parameters to summarise the results of the obtained posterior distributions. The original stated preferences are panel data. For illustration purposes, we selected only N=100 different travelers from each dataset, and then randomly chose one observation for each traveler (hereafter referred to as 'train data' and 'bus data' respectively) in the following analyses.

5.1. Analysis for the train data

We first consider the ordinary multinomial logit model where no transformation is involved:

$$S(\mathbf{x}^T\boldsymbol{\beta}) = -\theta(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6)\,, \tag{8}$$

where $x_1, \ldots, x_6$ represent the six attributes: access-egress time, headway, in-vehicle-time, waiting time, number of interchanges, and travel expenses. This model is a special case of the

more general semiparametric choice model (7) where the sensitivity function is taken as $S(t) = -\theta t$ with θ a scaling parameter.

The ordinary multinomial logit model was fitted and the results are displayed in the top panel of Table 3 where following Fosgerau and Bierlaire (2009), the coefficient of travel expenses was normalized to unit so that other coefficients can be interpreted as willingness-to-pay indicators. It can be seen that all attributes, except for the headway, were significant at 5% level. The value of the DIC for the ordinary multinomial logit model was 120.2.

(Table 3 is about here)

Next, the multiplicative choice model developed in Castillo et al. (2008) and Fosgerau and Bierlaire (2009) was used to fit the data. This model is a special case of model (7) where the sensitivity function is taken as $S(t) = -\theta \log(t)$:

$$S(\mathbf{x}^T\boldsymbol{\beta}) = -\theta \log(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6). \quad (9)$$

The mid-panel of Table 3 displays the estimates of the coefficients $\boldsymbol{\beta}$ in model (9). It can be seen that all the estimates, except for the headway, were significant at 5% level. The DIC of model (9), 122.6, was slightly higher than that of model (8), indicating that overall the data were not better fitted via the log-transformation.

Finally, the semi-parametric model developed in this paper was applied to fit the data:

$$S(\mathbf{x}^T\boldsymbol{\beta}) = S(u + v(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6)), \quad (10)$$

where no particular functional form of the sensitivity function $S(t)$ was imposed *a priori*. Instead it was determined by the data. The bottom panel of Table 3 displays the estimates of the coefficients $\boldsymbol{\beta}$. Clearly this model outperformed both models (8) and (9): it had the lowest value of the DIC, 83.1, thus providing much better fitting to the data. Figure 2 displays the obtained sensitivity function $S(t)$ on its support [0, 1]. It can be seen that in the middle part

of the support, the sensitivity function $S(t)$ is not sensitive to the change of the combined travel cost. Towards to the both extreme ends of the support, however, it increases (or decreases) rapidly. This suggests that each unit increment in the combined travel cost does not impact on the train users equally.

(Figure 2 is about here)

Figure 2. The estimated sensitivity function $S(t)$ for the train data

Similar to the other two models, all attributes, except for the headway, were significant at 5% level. The results obtained by model (10) were robust in the sense that the sensitivity function in model (10) was data-driven and no particular functional form was imposed simply due to mathematical convenience.

We also note that $S(t)$ in Figure 2 substantially differs from the sensitivity function $-\theta t$ of the Gumbel distribution and from the sensitivity function $-\theta \log(t)$ of the Weibull distribution (see Table 2). Hence, it suggests that the underlying distribution for the train data is neither the Gumbel nor the Weibull. This provides empirical evidence that the underlying assumption of the Gumbel distribution can be restrictive in some applications.

5.2. Analysis for the bus data

Next we briefly discuss the analysis for the bus data. Again we consider three different models: (a) the ordinary multinomial logit model without transformation, equation (8); (b) the multiplicative choice model with the log-transformation, equation (9); and (c) the semiparametric choice model (10). The estimation results are displayed in Table 4.

(Table 4 is about here)

It can be seen that all three models provided comparable fittings to the data in terms of likelihood, DIC and ρ^2. Figure 3 displays the sensitivity function $S(t)$ on its support of [0, 1] obtained using the semiparametric choice model. It is clear that the obtained sensitivity function is quite close to a linear function, and thus not surprisingly it produced similar estimates to that of the ordinary multinomial logit model for this particular data. Due to its simplicity, it seems that the ordinary multinomial logit model is a sensible choice.

(Figure 3 is about here)

Figure 3. The estimated sensitivity function $S(t)$ for the bus data

From this example it can be seen that when the actual underlying distribution is close to the Gumbel, the semiparametric model can automatically adapt its sensitivity function to produce a result similar to that of the ordinary multinomial logit model.

6. Discussion and conclusions

This paper has investigated the assumption of the underlying distributions of the random terms in the multinomial logit model. The research on this topic can be dated back to the early work of Lee (1983) who explored relaxing the assumption of underlying distributions by an arbitrary pre-specified distribution. On the other hand, Castillo et al. (2008) and Fosgerau and Bierlaire (2009) focused on one particular distribution, the Weibull, and used the Weibull distribution as an alternative to the Gumbel distribution to derive a choice model from the utility-maximizing behavior.

This paper has proposed relaxing the assumption of underlying distributions from the Gumbel and Weibull distributions to a wider distribution class, the distribution family (1). In comparison with the work of Lee (1983), the assumption of underlying distributions in this paper, i.e., the distribution family (1), is slightly more restrictive but still quite flexible to accommodate problems arising in different areas. More importantly, unlike Lee (1983), the underlying distribution in this paper is not required to be pre-specified in the stage of modeling. It also retains a crucial property in discrete decision analysis, i.e., it is closed under the minimum operation. Hence, similar to the multinomial logit model, the developed semi-parametric choice model can be derived from the individual choice behavior via the random cost minimization (or utility maximization). In addition, the distributions in family (1) do not require the random costs of interest to have homoscedastic variances. The proposed distribution family leads to a semiparametric choice model which links the linear combination of travel-related attributes to the choice probabilities via an unknown sensitivity function.

This paper has also shown that the sensitivity function plays an important role. Travelers may have different sensitivities to different transportation modes. When the sensitivity function is nonlinear, it indicates that travelers' reaction to the combined travel cost does not change in a proportionate manner. Clearly this has practical implications for the policy makers of public transportation systems.

Of the three assumptions made for the multinomial logit model as mentioned in the beginning of this paper, the semiparametric choice model has not only substantially relaxed the assumption of the Gumbel distribution but also to some extents relaxed the assumption of homoscedastic variances, and has addressed the issue of heteroscedastic variances via the variance-stabilizing transformation.

This paper has mainly focused on the closed form discrete choice models. One of the referees has pointed out that another approach to relax the underlying distributions of the multinomial logit model is via an open form choice model, i.e. mixed logit (see, e.g. Train 2003). The mixed logit uses the multinomial logit $L_{ni}(\boldsymbol{\beta}) = \frac{\exp\{\mathbf{x}_{in}^T\boldsymbol{\beta}\}}{\sum_{j\in C_n}\exp\{\mathbf{x}_{jn}^T\boldsymbol{\beta}\}}$ as a kernel, and all the remaining variability that cannot be accounted for is captured by a mixing distribution $q(\boldsymbol{\beta})$. In practice it is crucial to specify the right mixing distribution so that the variability that the multinomial logit kernel cannot explain is modeled by the mixing distribution. This may make the specification of the mixing distribution difficult. Although a commonly used mixing distribution is normal, Hess et al. (2005) show that it may lead to misleading interpretation of the results if the normal distribution is blindly used. In the recent years a considerable attention has been paid to this research issue. Fosgerau (2006) considers a number of distributions and concludes that a bad choice of the mixing distribution may lead to extreme bias. Hess and Axhausen (2005) look at a wealth of parametric distributions to investigate if they can reproduce a given target mixing distribution. Fosgerau and Bierlaire (2007) develop a practical test for the choice of mixing distribution. As a generic approach, the mixed logit may also cause difficulties in interpretation of results because there may be more than one source of variability that are modeled using a single mixing distribution. It could be hard for a researcher to distinguish between the different sources of variability from the obtained mixing distribution.

This paper tries to derive a flexible model that still maintains the closed form type of multinomial logit model. Instead of using a mixing distribution to capture all remaining variability, the semiparametric choice model in this paper explains different sources of variability more explicitly: it addresses the issue of heteroscedastic variances via the transformation and the issue of the nonlinear utility via the sensitivity function. Hence, model

specification is more straightforward and consequently interpretation of results is much easier. In practice it is up to the researcher to choose between a more generic or a more specific modeling approach on the basis of the purpose of analysis and his/her personal preference.

Finally, it should be noted that this paper has focused on data analysis and modelling. As one of the referees pointed out, one issue that the paper does not discuss is forecasting. In some applications, forecasting is even more important than modelling. The issue of forecasting for the semiparametric choice model will be investigated in the future research.

Appendix. Proofs of theorems

Proof of Theorem 1. It is trivial to show that $h(x)$ is monotonically increasing. The CDF of Z_j are given by $\Pr\{Z_j < z\} = \Pr\{X_j < h^{-1}(z)\}$. Since $\Pr\{X_j < x\} = 1 - [1 - F(x)]^{\alpha_j}$, we obtain $\Pr\{Z_j < z\} = 1 - [1 - F(h^{-1}(z))]^{\alpha_j} = 1 - \exp[-\alpha_j \exp(\theta z)]$. This completes the proof.

Proof of Theorem 2. We suppress the subscript j in equation (4) and let $V(\alpha) = \alpha\theta \int h^{-1}(t) \exp(\theta t) \exp\{-\alpha \exp(\theta t)\}\, dt$ denote the expectation EX which is regarded as a function of α. For α being increased to $\delta\alpha$, where $\delta = \exp(-\theta\lambda) > 1$ with $\theta > 0$ and $\lambda < 0$, we consider

$$V(\delta\alpha) = \delta\alpha\theta \int h^{-1}(t) \exp(\theta t) \exp\{-\delta\alpha \exp(\theta t)\}\, dt$$

$$= \alpha\theta \int h^{-1}(t) \exp[\theta(t-\lambda)] \exp\{-\alpha \exp[\theta(t-\lambda)]\}\, dt.$$

Let $u = t - \lambda$. Then

$$V(\delta\alpha) = \alpha\theta \int h^{-1}(u+\lambda) \exp(\theta u) \exp\{-\alpha \exp(\theta u)\}\, du.$$

Since $h^{-1}(.)$ is increasing and $\lambda < 0$, we have $h^{-1}(u+\lambda) - h^{-1}(u) \leq 0$. This implies that $V(\delta\alpha) \leq V(\alpha)$ for any $\delta > 1$ and $\alpha > 0$. Hence $V(\alpha)$ is a decreasing function of α. This completes the proof.

Acknowledgements

The author would like to thank the referees for their constructive comments on the earlier versions of this paper, and also to thank Dr Michel Bierlaire, Ecole Polytechnique Fédérale de Lausanne, Transport and Mobility Laboratory, Switzerland, for kindly providing the data analyzed in this paper.

References

Ben-Akiva, M., Lerman, S. R., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. The MIT Press, London.

Bhat, C. R., 1997. Covariance heterogeneity in nested logit models: Econometric structure and application to intercity travel. Transportation Research Part B 31 (1), 11-21.

Bhat, C. R., N. Eluru, R. B. Copperman, 2008. Flexible model structures for discrete choice analysis, in: D. A. Hensher, K. J. Button, (Eds.), Handbook of Transport Modelling, 2nd ed. Elsevier, Oxford, 75-104.

Castillo, E., Menendez, J. M., Jiménez, P., Rivas, A., 2008. Closed form expressions for choice probabilities in the Weibull case. Transportation Research Part B 42 (4), 373-380.

Eilers, P. H. C., Marx, B. D., 1996. Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). Statist. Sci. 11 (2), 89–121.

Fosgerau, M., 2006. Investigating the distribution of the value of travel time savings. Transportation Research Part B 40 (8), 688–707.

Fosgerau, M., Bierlaire, M., 2007. A practical test for the choice of mixing distribution in discrete choice models. Transportation Research Part B 41 (7), 784–794.

Fosgerau, M., Bierlaire, M., 2009. Discrete choice models with multiplicative error terms. Transportation Research Part B 43 (5), 494–505.

Fosgerau, M., Hjorth, K., Vincent Lyk-Jensen, S., 2006. The Danish Value of Time Study. Final report. DTF Report.www.transport.dtu.dk.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., 2003. Bayesian Data Analysis. Chapman & Hall, London.

Hess, S. Axhausen, K. W., 2005. Distributional assumptions in the representation of random taste heterogeneity. In: Proceedings of the Fifth Swiss Transportation Research Conference.

Hess, S., Bierlaire, M. and Polak, J., 2005. Estimation of value of travel-time saving using mixed logit models. Transportation Research Part A 39 (3), 221-236.

Horowitz, J. L., 2001. Nonparametric estimation of a generalized additive model with an unknown link function. Econometrica 69 (2), 499-513.

Huang, J.-C., Nychka, D. W. 2000. A nonparametric multiple choice method within the random utility framework. Journal of Econometrics 97 (2), 207-225.

Ichimura, H., 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. Journal of Econometrics 58 (1-2), 71-120.

Koppelman, F. S., 2008. Closed form discrete choice models, in: D. A. Hensher, K. J. Button, (Eds.), Handbook of Transport Modelling, 2nd ed. Elsevier, Oxford, 257-278.

Lang, S., Brezger, A., 2004. Bayesian P-splines. J. Comput. Graphical Statist. 13 (1), 183–212.

Lee, L.- F., 1983. Generalized econometric models with selectivity. Econometrica 51 (2), 507-512.

McFadden, D., 1978. Modelling the choice of residential location. In: A. Karlquist, L. Lundqvist, F. Snickars, J. Weibull, (Eds.), Spatial Interaction Theory and Residential Location. North-Holland, Amsterdam, 75-96.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., van der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society Series B, 64 (4), 583–639.

Stoker, T. M., 1986. Consistent estimation of scaled coefficients. Econometrica 54 (6), 1461-1481.

Swait, J., Adamowicz, W., 1996. The effect of choice environment and task demands on consumer behavior: Discriminating between contribution and confusion. Working Paper, Department of Rural Economy, University of Alberta.

Train, K., 2003. Discrete Choice Methods with Simulation. Cambridge University Press, New York.

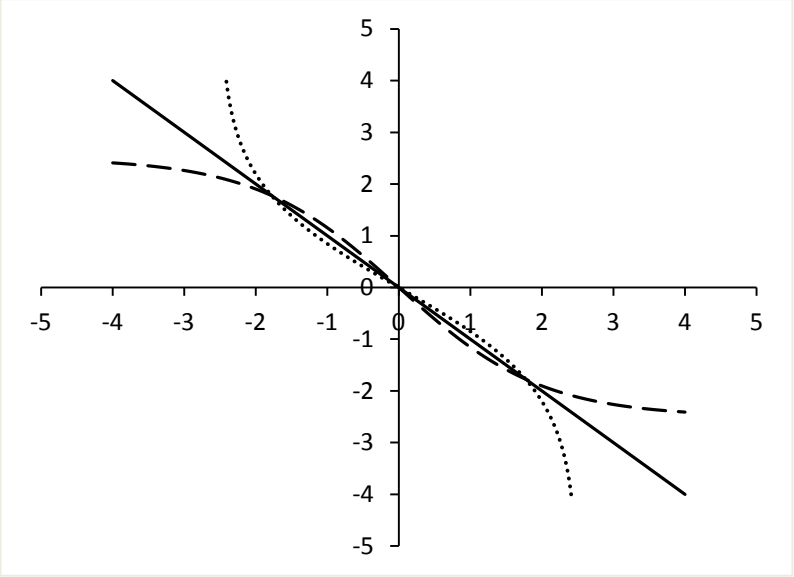

Figure 1. Illustration of sensitivity functions

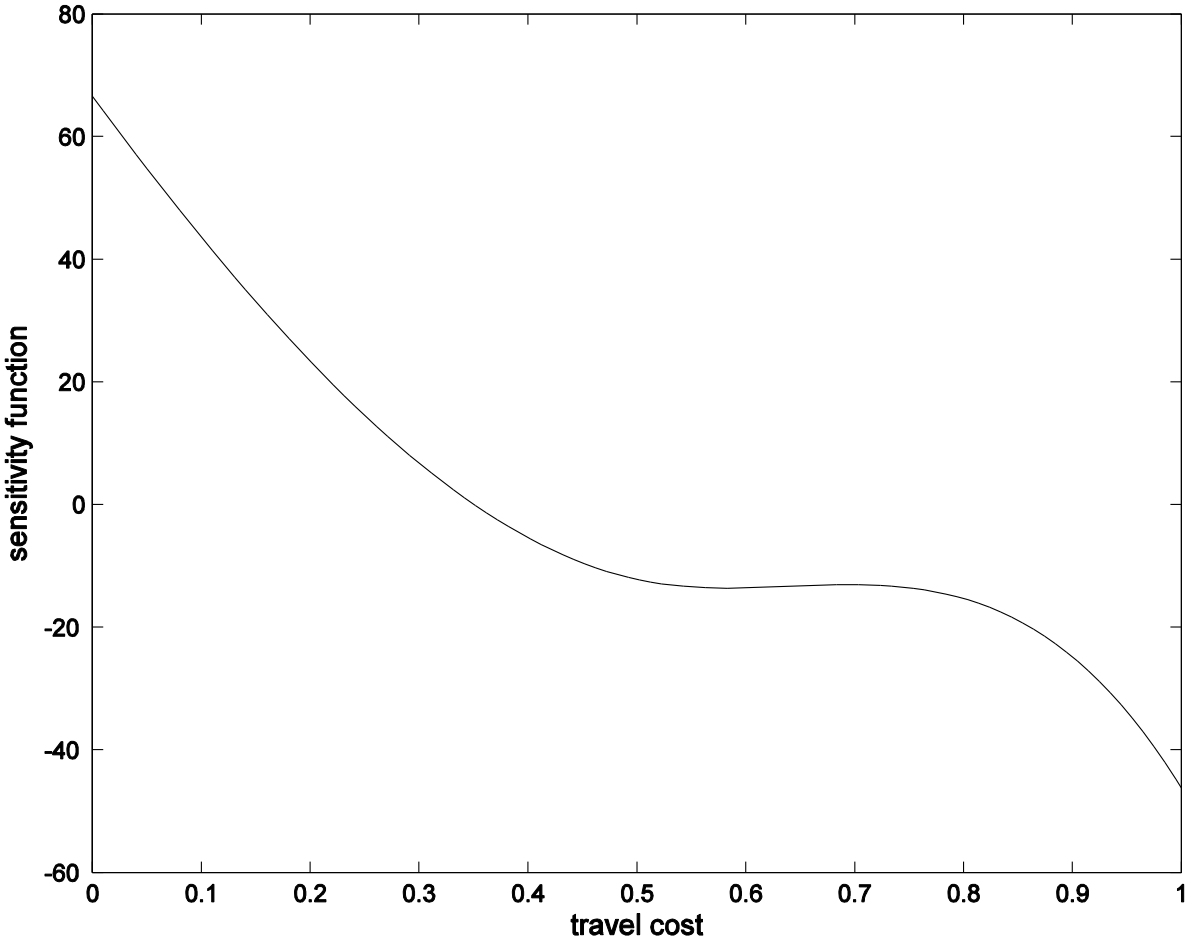

Figure 2. The estimated sensitivity function $S(t)$ for the train data

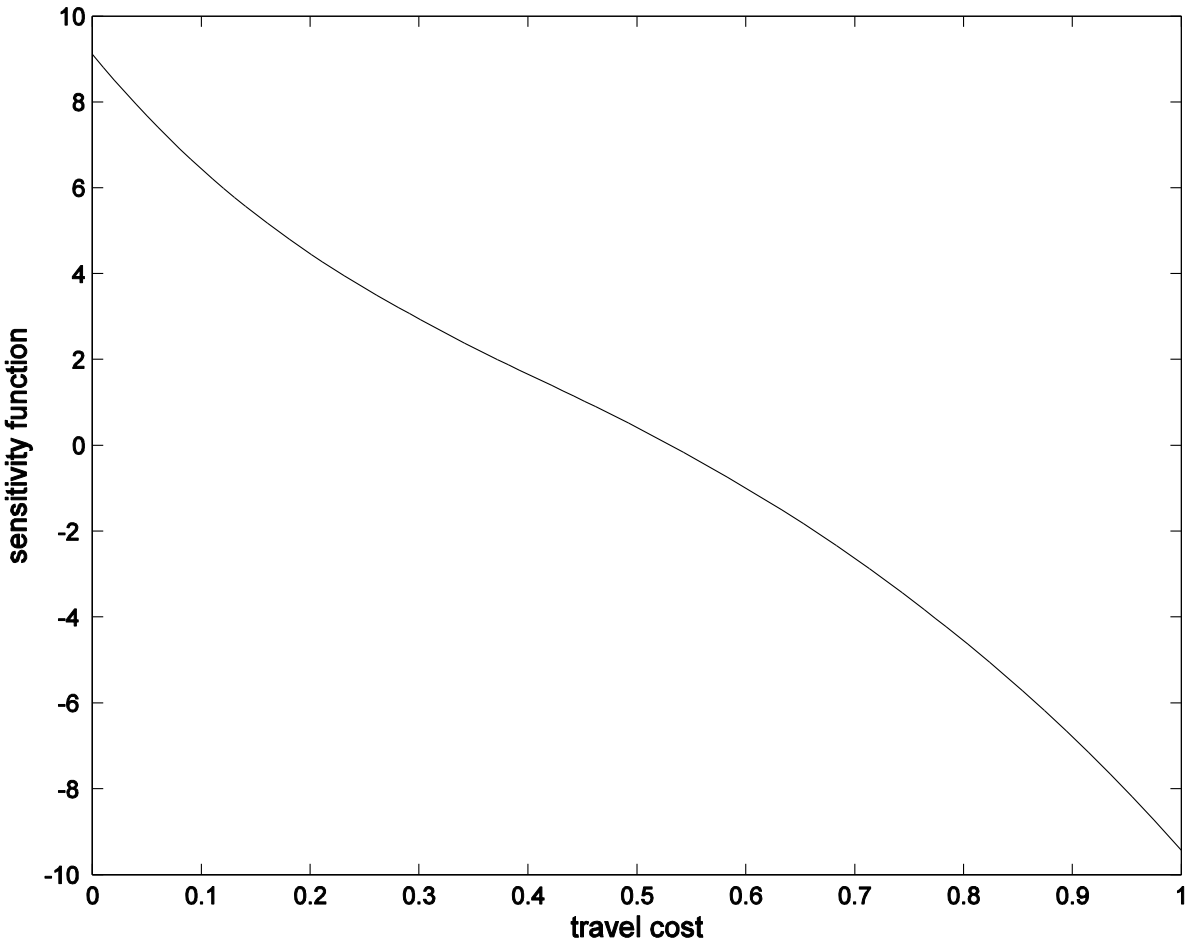

Figure 3. The estimated sensitivity function $S(t)$ for the bus data

Table 1. Special cases of the distribution family (1)

	Underlying distribution $F_{in}(t)$	Base distribution $F(t)$	Expectation V_{in}	Variance σ_{in}^2
Exponential	$1-\exp\{-\alpha_{in}t\}$	$1-\exp\{-t\}$	α_{in}^{-1}	α_{in}^{-2}
Pareto	$1-t^{-\alpha_{in}}\ \ (t\geq 1)$	$1-t^{-1}$	$\alpha_{in}/(\alpha_{in}-1)$	$\alpha_{in}/[(\alpha_{in}-1)^2(\alpha_{in}-2)]$
Type II generalized logistic	$1-[1+\exp(t)]^{-\alpha_{in}}$	$1-1/[1+\exp(t)]$	$\psi(1)-\psi(\alpha_{in})$	$\psi'(1)-\psi'(\alpha_{in})$
Gompertz	$1-\exp\{-\alpha_{in}[\exp(\theta t)-1]\}$	$1-\exp\{-[\exp(\theta t)-1]$		
Rayleigh	$1-\exp\{-\alpha_{in}t^2/2\}$	$1-\exp\{-t^2/2\}$	$[\pi/(2\alpha_{in})]^{1/2}$	$(4-\pi)/(2\alpha_{in})$
Weibull	$1-\exp\{-\alpha_{in}\,t^{\theta}\}$	$1-\exp\{-t^{\theta}\}$	$\alpha_{in}^{-1/\theta}\Gamma(1+1/\theta)$	$\alpha_{in}^{-2/\theta}\{\Gamma\left(1+\frac{2}{\theta}\right)-[\,\Gamma(1+\frac{1}{\theta})]^2\}$
Gumbel	$1-\exp\{-\alpha_{in}\exp(\theta t)\}$	$1-\exp\{-\exp(\theta t)\}$	$-\{\log(\alpha_{in})+\gamma\}/\theta$	$\pi^2/(6\theta^2)$

Table 2. The variance-stabilizing transformations, mean functions, and sensitivity functions for some distributions in family (1)

	Variance-stabilizing Transformation $h(t)$	Mean function $H(t)$	Sensitivity function $S(t)$
Exponential	$\theta^{-1}\log$ (t)	t^{-1}	$-\log(t)$
Pareto	$\theta^{-1}\log\{\log(t)\}$	$t/(t-1)$	$\log(t)-\log(t-1)$
Type II generalized logistic	$\theta^{-1}\log\{\log[1+\exp(t)]\}$	$\psi^{-1}(\psi(1)-\psi(t))$	$\log\{\psi^{-1}(\psi(1)-\psi(t)\}$
Gompertz	$\theta^{-1}\log\{\exp(\theta t)-1\}$		
Rayleigh	$\theta^{-1}\log(t^2)$	$\pi/(2t^2)$	$-2\log(t)$
Weibull	$\log(t)$	$\{\Gamma(1+1/\theta)/t\}^{\theta}$	$-\theta\log(t)$
Gumbel	t	$\exp(-\gamma-\theta t)$	$-\theta t$

Table 3. Estimates using different models for the train data*

	Multinominal logit model		
	log-likelihood= -54.8 $\rho^2 = 0.21$ DIC= 120.2		
Attributes	Posterior mean	Posterior standard deviation	95% credible interval
Access-egress time	0.70	0.39	(0.11, 1.63)
Headway	-1.26	7.23	(-14.00, 14.49)
In-vehicle time	0.32	0.15	(0.04, 0.63)
Waiting time	0.78	0.30	(0.27, 1.41)
Interchange	4.41	1.92	(0.68, 8.72)

	Multiplicative choice model		
	log-likelihood= -57.2 $\rho^2 = 0.17$ DIC=122.6		
Attributes	Posterior mean	Posterior standard deviation	95% credible interval
Access-egress time	1.32	0.49	(0.43, 2.38)
Headway	22.78	29.90	(-30.69, 74.40)
In-vehicle time	0.59	0.34	(0.10, 1.50)
Waiting time	0.97	0.12	(0.77, 1.17)
Interchange	6.35	3.11	(0.88, 13.63)

	Semiparametric choice model		
	log-likelihood= -34.4 $\rho^2 = 0.50$ DIC=83.1		
Attributes	Posterior mean	Posterior standard deviation	95% credible interval
Access-egress time	0.36	0.11	(0.16, 0.61)
Headway	-0.23	3.11	(-5.99, 6.27)
In-vehicle time	0.07	0.03	(0.01, 0.13)
Waiting time	0.39	0.11	(0.17, 0.62)
Interchange	0.87	0.39	(0.13, 1.65)

* The value of log-likelihood is -69.3 when all the parameters are set equal to zero.

Table 4. Estimates using different models for the bus data*

	Multinominal logit model		
	log-likelihood= -52.2 $\rho^2 = 0.24$ DIC=116.4		
Attributes	Posterior mean	Posterior standard deviation	95% credible interval
Access-egress time	0.30	0.17	(0.05, 0.68)
Headway	-4.92	3.53	(-11.54, 1.98)
In-vehicle time	0.01	0.03	(-0.04, 0.07)
Waiting time	0.47	0.20	(0.11, 0.88)
Interchange	1.37	0.51	(0.45, 2.42)

	Multiplicative choice model		
	log-likelihood= -52.5 $\rho^2 = 0.24$ DIC=114.3		
Attributes	Posterior mean	Posterior standard deviation	95% credible interval
Access-egress time	0.37	0.17	(0.05, 0.66)
Headway	-1.47	1.21	(-2.99, 0.97)
In-vehicle time	0.13	0.03	(0.09, 0.18)
Waiting time	0.41	0.12	(0.19, 0.62)
Interchange	0.97	0.53	(0.03, 1.76)

	Semiparametric choice model		
	log-likelihood= -52.5 $\rho^2 = 0.24$ DIC=114.6		
Attributes	Posterior mean	Posterior standard deviation	95% credible interval
Access-egress time	0.38	0.21	(0.04, 0.88)
Headway	-2.84	4.66	(-14.05, 4.47)
In-vehicle time	0.01	0.06	(-0.15, 0.09)
Waiting time	0.42	0.22	(0.04, 0.91)
Interchange	1.20	0.74	(0.11, 3.11)

* The value of log-likelihood is -69.3 when all the parameters are set equal to zero.