Disentangling the effects of unobserved factors on seatbelt use choices in multi-occupant vehicles

Afghari, Amir Pooyan; Faghih Imani, Ahmadreza; Papadimitriou, Eleonora; van Gelder, Pieter; Hezaveh, Amin Mohamadi

DOI
10.1016/j.jocm.2021.100324

Publication date
2021

Document Version
Final published version

Published in
Journal of Choice Modelling

Journal of Choice Modelling 41 (2021) 100324

Contents lists available at ScienceDirect

Journal of Choice Modelling

journal homepage: www.elsevier.com/locate/jocm

Disentangling the effects of unobserved factors on seatbelt use choices in multi-occupant vehicles

Amir Pooyan Afghari [a,*], Ahmadreza Faghih Imani [b], Eleonora Papadimitriou [a], Pieter van Gelder [a], Amin Mohamadi Hezaveh [c]

[a] Safety and Security Science Section, Faculty of Technology, Policy and Management, Delft University of Technology, 2628BX, Delft, the Netherlands
[b] Department of Civil and Environmental Engineering, Imperial College London, London, UK
[c] Civil and Environmental Engineering, University of Tennessee, Knoxville, TN, United States

ARTICLE INFO

Keywords:
Seatbelt use
Endogeneity
Integrated choice and latent variable
Bayesian inference

ABSTRACT

Despite the proven effectiveness of seatbelt use in reducing traffic casualties, not wearing a seatbelt still contributes to a substantial proportion of fatal crashes worldwide. This problem has raised the need to better understand factors contributing to seatbelt use, particularly in multi-occupant vehicles. Among these factors, behavioural determinants of seatbelt use are difficult to measure and their data are not readily available. These behavioural factors may have shared influences on vehicle occupants, causing their seatbelt use choices to be correlated. These complexities have prevented a comprehensive understanding of seatbelt use choices in the literature.

This study aims to fill this gap by developing an econometric model that explains seatbelt use choices in multi-occupant vehicles. A set of binary logit models are constructed for seatbelt use choices and their utilities are correlated across vehicle occupants. A new latent variable representing the unobserved factors or 'atmosphere' of the vehicle is then incorporated into the model. The model is empirically tested using seatbelt use data from Tennessee, United States. Results indicate that vehicle body type and time of the day are significantly associated with seatbelt use. In addition, the collective seatbelt use in a vehicle is influenced by the unobserved atmosphere in the vehicle. Age, alcohol and drug consumption, higher proportion of old population and white racial mix, higher income per capita, and higher education levels are factors contributing to this latent atmosphere.

1. Introduction

The effects of wearing a seatbelt in reducing roadway injuries and fatalities have long been documented in the road safety literature (Knapper et al., 1976; Hodson-Walker, 1970). Previous studies have shown that proper use of seatbelt can decrease the likelihood of fatality in a traffic crash between 44% and 73%, depending on the seating position and vehicle type (Blincoe et al., 2015). However, factors such as perceptions of safety, discomfort and social influence may affect self-protective behaviour of vehicle occupants (Cunill et al., 2004) and thus enforcing seatbelt use is considered to be one of the most effective measures in increasing seatbelt use rates (Dee,

* Corresponding author.
E-mail addresses: a.p.afghari-1@tudelft.nl (A.P. Afghari), s.faghih-imani@imperial.ac.uk (A. Faghih Imani), E.Papadimitriou@tudelft.nl (E. Papadimitriou), P.H.A.J.M.vanGelder@tudelft.nl (P. van Gelder), amohamad@vols.utk.edu (A.M. Hezaveh).

https://doi.org/10.1016/j.jocm.2021.100324
Received 21 November 2020; Received in revised form 21 September 2021; Accepted 22 September 2021
Available online 25 September 2021

1998; Eby et al., 2000). Yet, non-compliance with seatbelt use rules still accounts for a considerable proportion of roadway casualties (Shakya et al., 2020). For example, previous studies have reported that a substantial portion of drivers who died in traffic crashes in Tennessee, United States, failed to wear their seatbelt properly at the time of the crash (Cherry et al., 2018; Hezaveh and Cherry, 2019b; Hezaveh et al., 2019b). Similar findings have been reported in Canada (Jonah and Grant, 1985), China (Passmore and Ozanne-Smith, 2006), Malaysia (Kulanthayan et al., 2004), Pakistan (Klair and Arfan, 2014) and other parts of the world. These findings indicate that some individuals choose not to wear a seatbelt despite its proven effectiveness, raising the need to better understand factors influencing seatbelt use among vehicle occupants. Such an understanding can, in turn, lead to the deployment of more effective countermeasures to further increase seatbelt use rates.

Past research has shown that seatbelt use is associated with a multitude of factors arising from spatial and temporal characteristics of driving, type of vehicle, demographic and socioeconomic attributes of the vehicle occupants and their behaviours, attitudes and social norms (Lund, 1986; Ali et al., 2011; Okamura et al., 2012; Hezaveh and Cherry, 2019b; Afghari et al., 2020). While extensive research has been dedicated to the understanding of the effects of the above factors on seatbelt use, the complexities of vehicle occupants' choices in using seatbelt have been largely unexplored.

1.1. Complexities of seatbelt use choices in multi-occupant vehicles

One of the major challenges in understanding the seatbelt use behaviour of vehicle occupants is that a complete list of the factors contributing to such behaviour may not be readily available, especially from crash reports. In addition, the psychological and behavioural determinants of seatbelt use are very difficult to capture and the available data are usually limited. Recently, Afghari et al. (2020) incorporated residential location characteristics of vehicle occupants into their seatbelt use choices to overcome this challenge. They showed that 'home-based' characteristics of vehicle occupants may serve as a proxy for those attributes whose data are not available. In addition, it is intuitive to postulate that vehicle occupants' seatbelt use is influenced by the attributes of geographic location where they come from because geography highly influences behaviours, attitudes, and social norms (Rentfrow, 2010).

Seatbelt use choices of drivers and passengers in multi-occupant vehicles may be interrelated. This interrelationship may be due to the omission of psychological and behavioural factors from the analysis and can be manifested in unobserved errors with shared influences on the seatbelt use choices of vehicle occupants. For example, safety-conscious drivers may refuse to drive while the passengers are not wearing seatbelt. In the same fashion, safety-conscious passengers may force the driver to fasten the seatbelt. As a result, seatbelt use of front and rear passengers may be influenced not only by the seatbelt use of the driver but also by the seatbelt use of each other.

In addition, the collective character of a vehicle may alter individual seatbelt use choices. For example, the seatbelt use choice of an occupant in a vehicle in which the other occupants are conformist may be completely different than the seatbelt use choice of the same occupant in another vehicle in which the other occupants are eccentric and non-obedient. This collective character of individuals have been shown to be significantly associated with and intuitively represent behavioural ambiance or atmosphere of a place (Lopez-Pintado and Watts, 2008). As such, in this study, we refer to the collective (unobserved) character of vehicle occupants as "the atmosphere of the vehicle".

Finally, the above complexities are even more acute noting that the differences in the psychological and behavioural attributes of individuals may result in the varying effects of other external factors on their seatbelt use choices (Eluru and Bhat, 2007). For example, while many studies have shown that male drivers are less likely to wear seatbelt compared to female drivers (e.g., Pickrell and Ye, 2009; Gkritza and Mannering, 2008; Hezaveh and Cherry, 2019a), other studies have shown that some male drivers may be more safety-conscious than female drivers and thus may be more likely to wear seatbelt (Abay et al., 2013). This heterogeneity in the effect of gender on seatbelt use arises from unobserved behavioural factors (i.e. safety-consciousness) and thus may result in erroneous inferences about the effect of gender if not accounted for in modelling seatbelt use.

The above complexities associated with the seatbelt use choices of vehicle occupants have largely prevented a comprehensive understanding of such choices in the literature. One way of explaining these complexities is to develop a comprehensive econometric model that is capable of properly explaining the seatbelt use choices of driver and passengers in multi-occupant vehicles.

1.2. Econometric modelling of seatbelt use choices in multi-occupant vehicles

Seatbelt use choices of vehicle occupants are often obtained from crash reports or roadway observations in the format of a binary variable (i.e. zero for not wearing a seatbelt, and one for wearing a seatbelt). Univariate binary logit models have been widely used in the statistical literature as the modelling approach to estimate the effects of exogenous factors (e.g. the sociodemographic attributes, alcohol consumption, vehicle type) on individuals' binary choices (Washington et al., 2020). These models are based on the random utility theory, according to which the utility of a choice consists of a linear combination of a deterministic and a stochastic term. The deterministic term indicates the systematic effects of observed factors whereas the stochastic term indicates the random error caused by unobserved factors on individual choices. An important limitation of the univariate binary logit model is that it does not account for the correlation between binary choices of multiple individuals. As a result, the multivariate version of this model is more suited to capture the possible correlations between seatbelt use choices of multiple occupants in one vehicle. Multivariate models have been extensively used in the transport and discrete choice model literature (Ravulaparthy et al., 2013; Golob and Regan, 2002; Bhat and Srinivasan, 2005; Russo et al., 2014) to capture the shared influence of unobserved factors on multiple dependent variables. In addition, the random parameters specification has been used in choice models to capture unobserved heterogeneity in data and account for the varied effects of explanatory variables on the dependent variable (Mannering et al., 2016).

Another limitation of the binary logit model is the assumption that the dependent variable (seatbelt use choice in this context) must not affect the explanatory variables. When this assumption is not met (for example, if the atmosphere of the vehicle is to be included in the model as an explanatory variable), the dependent variable would be endogenous with the explanatory variable. Simultaneous equation models have been largely used in the statistical literature to address the endogeneity problem in statistical and econometric models (Washington et al., 2020). These models are divided into two general categories of single-equation methods (indirect least squares, instrumental variables, two-stage least squares and limited information maximum likelihood) and system equation methods (three-stage least square and full information maximum likelihood) with the latter providing consistent and more efficient estimates (Washington et al., 2020). Among system equation methods, the full information maximum likelihood approach is the most appropriate for modelling discrete choice data (Guevara and Hess, 2019; Guevara and Ben-Akiva, 2010). In particular, the latent variable approach has been shown to be a promising alternative to address endogeneity but it comes with high computational cost and identification problems (Guevara, 2015; Afghari et al., 2018, 2019b). Another approach to address endogeneity is to jointly analyse the variables which are endogenous with each other. This approach was first introduced by Eluru and Bhat (2007) in order to investigate the relationship between seatbelt use and crash injury severity. They found that the joint model is able to address the bias in parameter estimates caused by the endogeneity between safety-conscious drivers and their injury severity in traffic crashes.

Despite fairly extensive advancements in statistical modelling of seatbelt use, there is no econometric model that explains the complexities underlying seatbelt use choices of driver and passengers in multi-occupant vehicles. Modelling these choices requires various enhancements (e.g., multivariate setting and latent variables) of the existing statistical and econometric models. These enhancements lead to model estimation complexities, particularly using the maximum likelihood estimation. With the advanced computational capabilities, however, there is a great potential in developing these advanced econometric models and estimating them via simulation-based techniques such as Markov Chain Monte Carlo (MCMC) in the Bayesian framework (Oviedo-Trespalacios et al., 2020).

1.3. Study objectives

While extensive research efforts have been dedicated to understanding the seatbelt use choices of vehicle occupants, a meaningful portion of vehicle occupants still do not wear seatbelt. Previous studies have mostly focused on the seatbelt use choices of drivers and passengers separately and thus the interrelationship between seatbelt use choices of drivers and passengers have not been well studied. This study aims to fill this gap by developing a comprehensive econometric model that systematically considers the seatbelt use choices of drivers and passengers and their interactions. Borrowing from the ecological and epidemiological studies (Suzuki et al., 2012; Kouvonen et al., 2008), we introduce a new latent variable representing the atmosphere of the vehicle and incorporate it into the econometric model. The hypothesized model is then empirically tested using seatbelt use data from crash reports in Tennessee, United States. To minimize the possible sample selection bias resulted by extracting seatbelt use data from crash reports (Peltzman, 1975), the data have been validated by roadside observations in Tennessee (Hezaveh and Cherry, 2019a; Hezaveh et al., 2019b). It is also important to note that the average seatbelt use rates of drivers and passengers in Tennessee are fairly high (around 90%) and thus the main challenge in this study is to formulate an econometric model that is capable of capturing the variation within the small proportion of non-wearing seatbelt individuals.

2. Methodology

It is hypothesized that the combination of the above modelling methodologies (a multivariate binary choice model with latent variables and random parameters) corresponds with the seatbelt use choices of driver and passengers in multi-occupant vehicles. In addition, residential location characteristics of vehicle occupants are used in the analysis as proxies for their behaviours and attitudes. To reduce the dimensions of the analysis and avoid autocorrelation between the residential location variables, principal component analysis is used to summarize these characteristics into orthogonal explanatory variables in the proposed model. The details of these methodological approaches are presented in the following sections.

2.1. Latent variable multivariate binary choice model

Let $Y_i = [Y_{1i}, Y_{2i}, \ldots, Y_{ji}]$ be the vector of j $(j = 1, 2, \ldots, J;$ in our study $J = 3)$ binary dependent variables representing seatbelt use choices ($Y_{ji} = 0$ if not wearing seatbelt, and $Y_{ji} = 1$ if wearing seatbelt) of jth vehicle occupant (e.g. driver, front seat passenger, rear seat passenger) in vehicle i. The utility of using seatbelt for a given vehicle occupant (U_{ji}) is stated as:

$$U_{ji} = \beta_{ji} X_{ji} + \varepsilon_{ji} \tag{1}$$

where β_{ji} are estimable parameters (including the intercept), X_{ji} are explanatory variables (e.g. sociodemographic factors, vehicle type) and ε_{ji} is the random error term assumed to be identically and independently distributed across observations in each equation and describing the random part of the utility. To correlate the utilities of vehicle occupants within the same vehicle, a new random term (η_{ji}) is added to the above utilities such that:

$$\begin{cases} U_{1i} = \beta_{1i}X_{1i} + \varepsilon_{1i} + \eta_{1i} \\ U_{2i} = \beta_{2i}X_{2i} + \varepsilon_{2i} + \eta_{2i} \\ . \\ . \\ . \\ U_{ji} = \beta_{ji}X_{ji} + \varepsilon_{ji} + \eta_{ji} \end{cases} \tag{2}$$

where η_{ji} is the random term and is specified to capture the correlation between seatbelt use of multiple occupants in the same vehicle (i.e. multivariate setting) and thus follows a multivariate normal distribution with mean 0 and variance-covariance matrix Θ:

$$\begin{bmatrix} \eta_{1i} \\ \eta_{2i} \\ . \\ . \\ . \\ \eta_{Ji} \end{bmatrix} \sim MVN(0, \Theta) \text{ and } \Theta = \begin{bmatrix} \sigma_{11} & \sigma_{12} & . & . & . & \sigma_{1J} \\ \sigma_{21} & \sigma_{22} & . & . & . & \sigma_{2J} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ \sigma_{J1} & \sigma_{J2} & . & . & . & \sigma_{JJ} \end{bmatrix} \tag{3}$$

To account for the unobserved heterogeneity in the effects of explanatory variables on the dependent variables, model parameters are allowed to vary across vehicles:

$$\beta_{ji} = \overline{\beta}_j + \delta_{ji} \text{ and } \delta_{ji} \sim \text{Normal}(0, \upsilon_j) \tag{4}$$

where $\overline{\beta}_j$ and υ_j are the mean and standard deviation of parameters across vehicles. If the standard deviation of the Normal distributions are zero ($\upsilon_j = 0$), the model will reduce to the *fixed parameters* model. Assuming that ε_{ji} is generalized extreme value distributed (McFadden, 1981), the probability of jth occupant wearing seatbelt in vehicle i can be presented as:

$$P(Y_{ji} = 1) = \frac{1}{1 + e^{-(\beta_{ji}X_{ji} + \eta_{ji})}} \tag{5}$$

The likelihood of using seatbelt across all individuals can then be determined by the product of equation (4) over the entire observations. This model is referred to as the *multivariate binary choice model* in this manuscript.

The atmosphere of the vehicle is now incorporated into the model as latent explanatory variables in the linear utility functions for the dependent variables (Equation (1)). More specifically, we define a new latent variable (Z_i) for each occupant and insert it in the utility of the binary choice model for that occupant. It is hypothesized that this latent variable representing the atmosphere of the vehicle is a proportion (more on this will be presented later in this section) and thus is assumed to follow a beta distribution as (Afghari et al., 2021):

$$P(Z_{ji} = z) = \Gamma(a+b)\frac{z^{(a-1)}\,(1-z)^{(b-1)}}{\Gamma(a)\Gamma(b)} \tag{6}$$

where Γ is the gamma function and a and b are the parameters of the beta distribution. The expectation of this latent variable is equal to $E[Z_{ji}] = \mu = \frac{a}{a+b}$. Using a logit link function in the structural equation of the latent variable, the expectation of this latent variable can be expressed as a function of exogenous covariates:

$$\mu_{ji} = \frac{1}{1 + e^{-(\lambda_j m_{ji})}} \tag{7}$$

where λ_j are estimable parameters, m_{ji} exogenous covariates, and the rest of notations are as previously stated.

Similar to the approach used by (Sanko et al., 2014), we now employ a measurement equation to help inform the role of the latent variables in the model. More specifically, it is postulated that the latent variable for each vehicle occupant is measured by the proportion of self-excluded seatbelt use –which is the proportion of seatbelt use for all vehicle occupants excluding the vehicle occupant of interest:

$$P(X_{ji}^{(p)}) = g(Z_{ji},\ \gamma_j) \text{ where} X_{ji}^{(p)} = \frac{\left[\sum_{n=1}^{j-1} Y_{ni}\right] + \left[\sum_{n=j+1}^{J} Y_{ni}\right]}{J-1} \tag{8}$$

$X_{ji}^{(p)}$ in the above equation is the proportion of self-excluded seatbelt use in the ith vehicle for the jth vehicle occupant and γ_j is an estimable parameter. It is worth mentioning that while $X_{ji}^{(p)}$ is endogenous with the seatbelt use of each vehicle occupant (Y_{ji}), employing a latent variable in the original utility function that is explained by exogenous covariates (m_{ji} in Equation (7)) and lacks a one-to-one relationship with $X_{ji}^{(p)}$ caused by the errors in $X_{ji}^{(p)}$ in Equation (8), corrects for such endogeneity in the overall model.

Selecting the specific functional form for $g(.)$ is not a trivial task and may require several trial and errors. However, previous studies have shown that choosing the same type of probability distribution for the endogenous and the latent variables (beta distribution in this study) enhances model estimation and provides promising results (Oviedo-Trespalacios et al., 2020). Therefore, the measurement equation is defined as:

$$P\left(X_{ji}^{(p)}\right) = \Gamma(c+d)\frac{\left(X_{ji}^{(p)}\right)^{(c-1)}\left(1-X_{ji}^{(p)}\right)^{(d-1)}}{\Gamma(c)\Gamma(d)} \text{ where } \frac{c}{c+d} = \frac{1}{1+e^{-(\gamma_j Z_{ji})}} \tag{9}$$

It is worth mentioning that the choice of the beta distribution as the distribution for $X_{ji}^{(p)}$ is based not only on its ability to enhance model estimation, but also on the theoretical hypothesis in this study in which the proportion of self-excluded seatbelt use in a vehicle (as a continuous variable) measures the atmosphere of the vehicle for each occupant. This hypothesis is primarily motivated by psychological research showing that becoming attracted to strangers –which is the main element of being influenced by the atmosphere– is better predicted by the proportion of their similar behaviours rather than the number of their similar behaviours (Byrne and Nelson, 1965).

This elaborate model for the seatbelt use choices of vehicle occupants is referred to as the *latent variable* model in the frequentist approach (Sanko et al., 2014) and does not have a closed from to be estimated using the regular maximum likelihood estimation technique. However, it has an elegant hierarchical representation in the Bayesian approach (Oviedo-Trespalacios et al., 2020):

$$P\left(Y_{1i},\ Y_{2i},\ \ldots,\ Y_{ji}\right) \sim \text{ Multivarite binary logit} - \text{normal}\left(\overline{\beta}_j,\ \upsilon_j, \Theta,\ Z_{ji}\right) \tag{10}$$

$Z_{ji} \sim$ beta (a, b) where $\frac{a}{a+b} = \frac{1}{1+e^{-(\lambda_j m_{ji})}}$ *structural equation*

$X_{ji}^{(p)} \sim$ beta (c, d) where $\frac{c}{c+d} = \frac{1}{1+e^{-(\gamma Z_{ji})}}$ *measurement equation*

Bayesian estimation offers a significant advantage over the maximum likelihood estimation in that complicated likelihood functions and posteriors can be considered in model estimation (For more details on the Bayesian estimation of hierarchical models see Bolduc et al. (2005)). As such, the Bayes' theorem is used to estimate the model in which posterior estimates are drawn based on random sampling from the likelihood and the prior (Washington et al., 2020; Lord and Washington, 2018). Standard MCMC sampling method is used to simulate the posterior densities in the above Bayesian framework because the above model is intractable analytically. Within the estimation process and for identification purposes, the parameters of latent variables in the utility functions are fixed ($\upsilon_j = 0$) and one of the shape parameters in each of the beta distributions (b in the structural equation and d in the measurement equation) is set to unity.

2.2. Model selection and goodness-of-fit

The suitability of the hypothesized latent variable multivariate model is tested by applying it on empirical data and comparing its statistical fit with that of the alternative models. In addition, fixed and random parameters variants of the models can be estimated separately and their statistical fit can be compared to ensure whether the parameters are fixed or random.[1]

Deviance Information Criterion (DIC) is used as the measure of fit for model selection in the Bayesian paradigm. DIC is the hierarchical modelling generalization of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) and is defined as (Geedipally et al., 2014):

$$DIC = \overline{D}(\Theta) + P_D \tag{11}$$

where

$$\overline{D}(\Theta) = E[-2\log L]$$

$$P_D = \overline{D}(\Theta) - D(\overline{\Theta})$$

In the formulation above, L is the likelihood of the model at convergence, Θ is the total number of parameters,P_D is the effective number of parameters reflecting model complexity and $D(\overline{\Theta})$ is the deviance evaluated at a posterior summary of Θ. In addition, the McFadden pseudo-rho squared (ρ^2) is used as a widely accepted measure of goodness-of-fit for discrete choice models (McFadden, 1973):

$$\rho^2 = 1 - \left[D_m/D_0\right] \tag{12}$$

where D_m and D_0 are the deviance of the full model and the deviance of the null model, respectively. A model with a lower DIC or a higher ρ^2 is usually preferred over the other models.

Geedipally et al. (2014) have shown that DIC is only comparable for the models that have the same likelihood structure i.e. nested

[1] The random parameters specification in the Bayesian approach is slightly different than the frequentist approach in that it is assumed that the additional variance in the parameters comes from unobserved heterogeneity and thus an additional level of hierarchy is introduced into the Bayesian models (Oviedo-Trespalacios et al., 2020; Rusli et al., 2018). One way to investigate the suitability of random parameters in capturing unobserved heterogeneity, in such a specification, is to compare the statistical fit of the fixed and random parameters models.

models, and leads to erroneous inferences otherwise. As such, Mean Absolute Deviance (MAD) and Mean Squared Predictive Error (MSPE) are also used to compare the performance of the models that are not nested:

$$MAD = \frac{1}{N} \sum_{i=1}^{N} \left| Y_{ji} - P\left(Y_{ji}\right) \right| \tag{13}$$

$$MSPE = \frac{1}{N-P} \sum_{i=1}^{N} \left(Y_{ji} - P\left(Y_{ji}\right) \right)^2 \tag{14}$$

where N is the sample size and P is the number of estimated parameters. The model with smaller MAD and MSPE is usually preferred over the other models.

2.3. Principal component analysis

In estimating the above latent variable model, residential location characteristics of vehicle occupants can be used as proxies for their behaviours and attitudes (Hezaveh and Cherry, 2019b; Afghari et al., 2020). However, these characteristics may have high autocorrelation with each other as previously shown in the statistical analysis of behavioural data (Huitema, 1986; Huitema and McKean, 1991). Principal Component Analysis (PCA) is a common approach used in the statistical literature (Tipping and Bishop, 1999) to summarize data when there are too many variables in the analysis (Henry and Hidy, 1979). The PCA creates a set of new variables, referred to as *principal components (PC)*, each of which is a linear and orthogonal combination of the original variables in such a way that each orthogonal combination captures the maximum variability in the original set of variables and has the minimum autocorrelation with other linear combinations. The principal components can be obtained by applying the orthogonal transformation and finding the Eigenvectors and Eigenvalues of the Spearman correlation matrix of the original set of explanatory variables. The principal components are then arranged based on their decreasing contribution to the total variance of the original set of explanatory variables: the first principal component explains the highest variability in the explanatory variables; the second principal component explains the second highest variability in the explanatory variables, and so forth (the cumulative contribution of all principal components is equal to 1). These principal components can then be used in the analysis as representatives of the original set of variables. The number of principal components to be used in the model depends on the specific research objective, though the common practice is to use all principal components with Eigenvalues greater than one (Tipping and Bishop, 1999).

3. Empirical data

The empirical data in this study were provided by Tennessee Integrated Traffic Analysis Network (TITAN), which is a state-wide repository for traffic crashes and surveillance reports completed by Tennessee law enforcement agencies. For the year 2016, the data include 247,536 crashes and information about 725,388 drivers who were involved in traffic crashes. TITAN also provides information

Table 1
Summary statistics of variables used in the study.

	Mean	St. D.	Min	Max
Driver				
Seatbelt use (dummy)	0.92	0.26	0.00	1.00
Age	37.04	17.54	17.00	92.00
Alcohol consumption (dummy)	0.02	0.16	0.00	1.00
Distraction (dummy)	0.05	0.25	0.00	1.00
Drug consumption (dummy)	0.01	0.12	0.00	1.00
Gender: male	0.55	0.50	0.00	1.00
Front seat passenger				
Seatbelt use (dummy)	0.93	0.26	0.00	1.00
Age	36.84	18.61	17.00	99.00
Alcohol consumption (dummy)	0.02	0.13	0.00	1.00
Drug consumption (dummy)	0.01	0.08	0.00	1.00
Gender: male	0.47	0.50	0.00	1.00
Rear seat passenger				
Seatbelt use (dummy)	0.88	0.32	0.00	1.00
Age	33.52	17.80	17.00	99.00
Alcohol consumption (dummy)	0.02	0.13	0.00	1.00
Drug consumption (dummy)	0.00	0.07	0.00	1.00
Gender: male	0.52	0.50	0.00	1.00
Driving characteristics				
Driving during the day (dummy)	0.69	0.46	0.00	1.00
Driving in clear weather conditions (dummy)	0.75	0.43	0.00	1.00
Driving in rainy weather conditions (dummy)	0.12	0.33	0.00	1.00
Vehicle body type: large vehicles (dummy)	0.17	0.38	0.00	1.00
Speed limit (mph)	40.04	13.78	0.00	70.00

regarding seatbelt use by occupants at the time of the crash. For this study, we defined seatbelt 'non-use' as vehicle occupants who did not restrain both lap and shoulder seat belt at the time of a crash. Furthermore, occupants' home locations were geocoded using only recorded addresses with an accuracy level of the premise (e.g., property name, building name), address-level accuracy, or intersection level accuracy and used in the analysis (For more details about the geocoding process of the home addresses please see Merlin et al. (2020) and Hezaveh et al. (2019a)).

Due to the very few number of observations for vehicles with more than three occupants, the data were further down sampled to only include three-occupant vehicles. As a result and after controlling for the number of occupants in vehicles and cleaning the data (i.e., removing the incomplete records and error entries), 10,950 records (3 650 vehicles each with 3 occupants) with a Tennessee address were selected for assignment to the census tract data. Census tract data from the U.S. survey in 2010 were also used for obtaining sociodemographic and socioeconomic data elements including total population (and its proportions of different age cohorts), percentage of racial mix (white, black, Hispanic, etc.), transport mode share for work-commute, average travel time, average household size, proportion of educational levels, average income, income per capita, average vehicle ownership, and proportion of vacant houses. To prevent outliers, we only considered the census tracts that had more than 20 observations. Table 1 and Table 2 present the summary statistics of the variables and the census tract data considered as input for model estimation in this study.

4. Results

To test the applicability of the proposed latent variable multivariate model, it was estimated against the empirical data. While it is intuitive to assume that the seatbelt use choices of all vehicle occupants are interrelated, the magnitude of this interrelationship may vary across vehicle occupants. For example, one may argue that the seatbelt use choice of the driver may be only slightly associated with the seatbelt use choices of the front-seat and the rear-seat passengers due to the heavy enforcement of the seatbelt use for the former occupant. As a result, a restricted variant of the latent variable multivariate model was also estimated in which the latent variable is only included in the utilities of the front-seat and rear-seat passengers. Further, two additional model specifications were also estimated to serve as the benchmark for testing the performance of the proposed latent variable model in this study. These two models are (i) univariate joint models with contemporaneous seatbelt use choices of vehicle occupants as explanatory variables (Eluru and Bhat, 2007), and (ii) a multivariate model without the latent variable. Fig. 1 presents the schematic diagram of the modelling framework in this study. The proposed multivariate model with latent variables is obtained when the observed effects among seatbelt uses are set to zero (J arrows are removed). The multivariate model without latent variables is obtained when the observed effects among seatbelt uses and the observed effects of latent variables are set to zero (J and L arrows are removed); and the univariate joint models are obtained if the observed effects of latent variables and the unobserved effects among seatbelt uses are set to zero (L and the C arrows are removed).

In all of these models, explanatory variables were tested for multicollinearity by computing the Pearson correlation coefficients,

Table 2
Summary statistics of census tract information used in the study.

	Driver				Front seat passenger				Rear seat passenger			
	Mean	St. D.	Min	Max	Mean	St. D.	Min	Max	Mean	St. D.	Min	Max
Total population density (1 000 person/km2)	0.731	1.009	0.000	33.099	0.734	0.870	0.000	10.513	0.754	0.879	0.000	10.513
Percentage of population under 16 years old	0.241	0.079	0.000	0.602	0.240	0.081	0.000	0.572	0.242	0.083	0.000	0.713
Percentage of population above 60	0.185	0.092	0.000	0.894	0.184	0.091	0.000	0.894	0.182	0.090	0.000	0.746
Percentage of white race	0.704	0.326	0.000	1.000	0.695	0.328	0.000	1.000	0.689	0.332	0.000	1.000
Percentage of commuting to work by car	0.933	0.073	0.000	1.000	0.927	0.090	0.000	1.000	0.926	0.091	0.000	1.000
Percentage of commuting to work by car pool	0.113	0.084	0.000	0.817	0.112	0.085	0.000	0.817	0.112	0.085	0.000	0.678
Percentage of commuting to work by bus	0.013	0.038	0.000	0.385	0.016	0.047	0.000	0.500	0.017	0.048	0.000	0.617
Percentage of commuting to work by bike	0.001	0.006	0.000	0.104	0.001	0.008	0.000	0.185	0.001	0.007	0.000	0.104
Percentage of commuting to work by walk	0.014	0.035	0.000	0.419	0.016	0.043	0.000	0.500	0.016	0.043	0.000	0.500
Average travel time (hours)	0.429	0.102	0.000	1.098	0.425	0.099	0.000	0.882	0.424	0.101	0.000	0.882
Average household size	2.610	0.528	0.000	16.233	2.609	0.756	0.000	37.675	2.613	0.688	0.000	30.723
Percentage of high school education	0.525	0.188	0.000	0.970	0.530	0.186	0.000	0.970	0.533	0.185	0.000	0.970
Percentage of college education	0.214	0.079	0.000	0.642	0.211	0.077	0.000	0.642	0.211	0.078	0.000	0.541
Percentage of post-secondary education	0.192	0.112	0.000	0.637	0.191	0.112	0.000	0.677	0.188	0.112	0.000	1.000
Average household income ($100,000)	0.451	0.223	0.000	2.147	0.440	0.218	0.000	2.012	0.434	0.217	0.000	2.383
Income per capita	0.223	0.102	0.000	1.202	0.219	0.097	0.000	1.054	0.216	0.097	0.000	1.054
Number of vacant houses	0.117	0.099	0.000	0.925	0.121	0.099	0.000	0.676	0.123	0.098	0.000	0.688
Proportion of households with 0 vehicle	0.073	0.093	0.000	0.733	0.080	0.102	0.000	0.733	0.083	0.106	0.000	0.733
Proportion of households with 1 vehicle	0.339	0.143	0.000	0.892	0.344	0.146	0.000	0.806	0.349	0.146	0.000	0.880
Proportion of households with 2 vehicles	0.372	0.125	0.000	0.786	0.363	0.129	0.000	0.746	0.361	0.132	0.000	0.786
Proportion of households with 3 or more vehicles	0.216	0.119	0.000	0.694	0.210	0.121	0.000	0.756	0.205	0.120	0.000	0.694

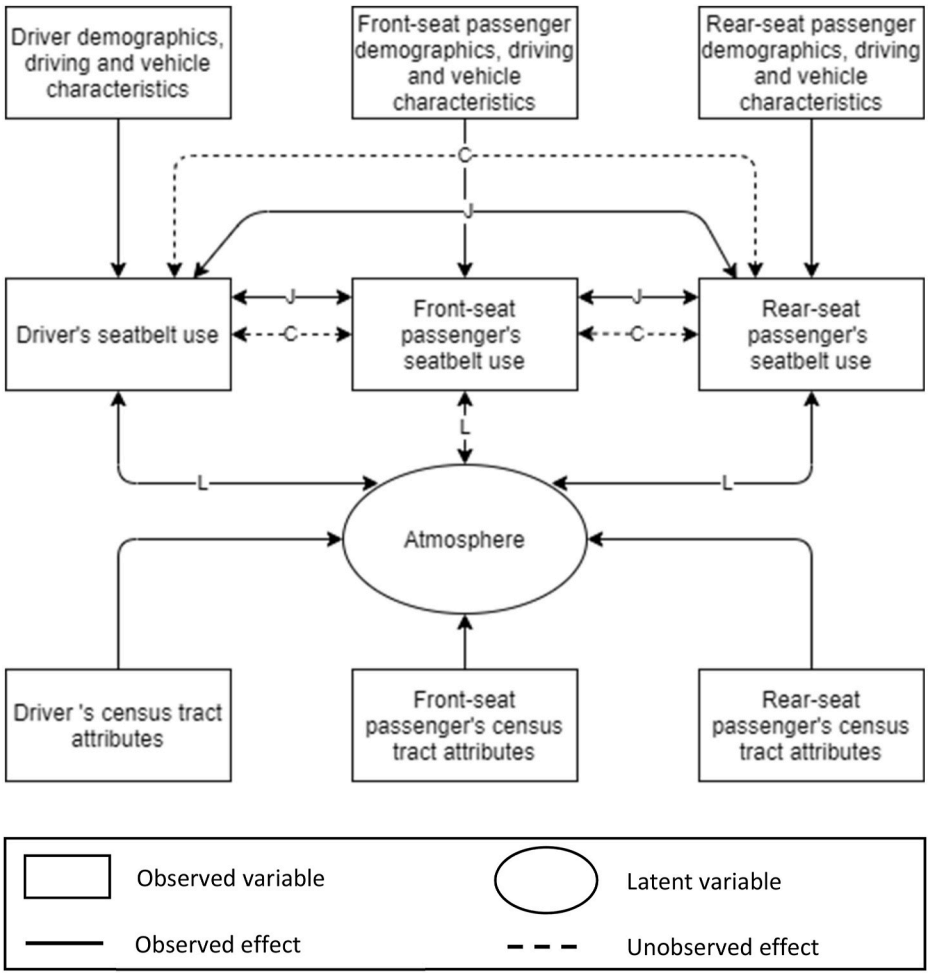

Fig. 1. Schematic diagram of the modelling framework for seatbelt use in three-occupant vehicles (*C*: unobserved factors shared among seatbelt uses, *J*: observed effects of seatbelt uses on one another; *L*: observed effects of latent variables on seatbelt uses).

and the variables with correlation coefficients higher than 0.7 were excluded from the models. The explanatory variables were inserted into the models using stepwise variable selection criterion. All models were estimated in the Bayesian framework. In the absence of solid prior information, non-informative priors were assigned to all parameters in the form of normal distribution with mean 0 and standard deviation 100. In addition, non-informative priors in the form of exponential distributions (with parameter 1) were assigned to the standard deviation of random parameters (Afghari et al., 2019a). Finally, non-informative priors in the form of a Wishart distribution ($\begin{bmatrix} 0.1 & 0.005 & 0.005 \\ 0.005 & 0.1 & 0.005 \\ 0.005 & 0.005 & 0.1 \end{bmatrix}$, was assumed as the hyper-prior for the variance-covariance matrix of the random error terms (Afghari et al., 2016). MCMC simulation was used to estimate the posterior in all models which resulted in two Markov chains converging after 50,000 iterations. The convergence was ensured by visual monitoring (obtaining stabilized and well-mixed chains) as well as assessing the Gelman-Rubin statistics ($R_{\text{Gelman-Rubin}} \rightarrow 1$). The simulation process was continued for additional 10,000 iterations in order to generate posterior samples for developing inferences about parameters.

During the estimation process, we found that while census tract variables were not statistically significant in any of the models, their principal components were indeed statistically significant in the latent variable model and provided more insight about the effects of residential location characteristics on seatbelt use (more on this will be presented later), hence justifying the need to use the principal component analysis in this study. The results of the principal component analysis i.e. the Eigenvalues, proportion of explained variability, and cumulative proportion of explained variability are provided in the appendix.

4.1. Model selection

The goodness-of-fit measures were calculated for all model candidates in order to select the model with the superior statistical fit. The results are reported in Table 3.

Table 3
Results of goodness-of-fit measures for model candidates.

	Joint models			Multivariate model without latent variables			Multivariate with restricted latent variables			Multivariate with unrestricted latent variables		
	Driver	Front-seat passenger	Rear-seat passenger	Driver	Front-seat passenger	Rear-seat passenger	Driver	Front-seat passenger	Rear-seat passenger	Driver	Front-seat passenger	Rear-seat passenger
MAD	0.057	0.044	0.108	0.028	0.035	0.053	0.083	0.042	0.036	0.007	0.010	0.011
MAD_N	0.378	0.296	0.462	0.184	0.235	0.227	0.551	0.282	0.151	0.004	0.005	0.005
MSPE	0.028	0.021	0.054	0.006	0.011	0.013	0.028	0.019	0.006	0.000	0.001	0.000
$MSPE_N$	0.275	0.214	0.406	0.064	0.115	0.088	0.221	0.084	0.032	0.004	0.008	0.003
D_0	1948.0	1928.0	2 640.0	506.3	514.4	831.9	506.3	514.4	831.9	506.3	514.4	831.9
D	867.1	663.1	1 576.0	1 414.6			1999.0			336.8		
ρ^2	0.555	0.656	0.403	0.236			−0.079			0.818		
DIC	875.8	671.1	1 587.0	2 443.0			2 909.0			671.5		

MAD_N: Mean absolute deviance for individuals who have not worn seatbelt, **$MSPE_N$**: mean squared predictive error for individuals who have not worn seatbelt, **D_0**: Null deviance, **D**: Deviance at convergence, ρ^2: McFadden rho squared.

The multivariate model with unrestricted latent variables has lower MAD for drivers, front-seat passengers and rear-seat passengers (0.007, 0.010, and 0.011 respectively) in comparison with the joint model (0.057, 0.044 and 0.108, respectively) and the multivariate model without latent variables (0.028, 0.035, and 0.053, respectively). This difference in the mean absolute deviances significantly increases for the vehicle occupants who have not used seatbelt (second row in Table 3). The same results were found for MSPE across the model candidates. In addition, the multivariate model with unrestricted latent variables has a substantial higher ρ^2 (0.818) compared to all other model candidates. Finally, the multivariate model with unrestricted latent variables has a substantial lower DIC (671.5) compared to the multivariate model with restricted latent variables (2 909.0). These findings indicate that the proposed multivariate model with unrestricted latent variables has consistently improved statistical fit in comparison with the alternative models. This superiority in the statistical fit indicates that the correlation between seatbelt use choices of vehicle occupants arises from two distinct sources, (1) the latent 'atmosphere' measured by the self-excluded seatbelt use choices of vehicle occupants and predicted by their underlying socioeconomic attributes, and (2) other unobserved factors captured by the contemporaneous error terms. The proposed latent variable is a better predictor of seatbelt use choices in multi-occupant vehicles in comparison with the sole use of observed seatbelt use choices (in the joint model and in the multivariate model with endogenous variables) or the sole use of unobserved correlations between seatbelt use choices (in the multivariate model). It also supports our hypothesis that the atmosphere of the

Table 4
Result of the latent variable multivariate model of seatbelt use choices in multi-occupant vehicles.

Vehicle occupant	Variable	Mean	St. D.	95%Credible Interval	
				2.5%	97.5%
Driver	Constant	−13.860	5.301	−22.420	−6.748
	Vehicle body type: large vehicles	4.276	2.527	0.482	10.030
	Time of day - daytime	3.433	2.094	0.263	8.304
	Latent variable for the driver	62.880	23.700	33.010	93.340
	Structural equation				
	Constant	0.627	0.021	0.585	0.667
	First PC[a] of the front-seat passenger	−0.053	0.010	−0.074	−0.033
	First PC of the rear-seat passenger	−0.047	0.011	−0.068	−0.026
	Front-seat age	0.282	0.026	0.232	0.335
	Rear-seat age	−0.065	0.023	−0.113	−0.021
	Front-seat alcohol consumption	−1.191	0.225	−1.653	−0.764
	Rear-seat alcohol consumption	−0.671	0.226	−1.096	−0.246
	Front-seat drug consumption	−2.218	0.369	−2.970	−1.569
	Measurement equation				
	γ (coefficient of the latent variable)	9.913	0.029	9.857	9.972
Front-seat passenger	Constant	−6.196	1.340	−9.076	−4.170
	Latent variable for the front-seat passenger	38.750	7.083	28.560	52.590
	Structural equation				
	Constant	0.668	0.022	0.626	0.712
	First PC of the driver	−0.083	0.009	−0.100	−0.062
	First PC of the rear-seat passenger	−0.021	0.010	−0.041	−0.002
	Driver age	0.203	0.019	0.165	0.238
	Driver alcohol consumption	−1.355	0.161	−1.697	−1.047
	Rear alcohol consumption	−0.888	0.206	−1.297	−0.526
	Driver drug consumption	−1.728	0.229	−2.157	−1.303
	Rear drug consumption	−0.532	0.312	−1.114	0.092
	Measurement equation				
	γ (coefficient of the latent variable)	9.903	0.029	9.848	9.963
Rear-seat passenger	Constant	−25.120	9.669	−42.240	−13.640
	Latent variable for the rear-seat passenger	71.420	27.410	39.050	117.100
	Structural equation				
	Constant	0.984	0.022	0.943	1.027
	First PC of the driver	−0.093	0.010	−0.113	−0.074
	First PC of the front-seat passenger	−0.061	0.010	−0.078	−0.040
	Driver age	0.155	0.027	0.101	0.207
	Front age	0.173	0.028	0.119	0.235
	Driver alcohol consumption	−1.535	0.185	−1.925	−1.206
	Rear alcohol consumption	−0.626	0.180	−0.996	−0.265
	Driver drug consumption	−1.295	0.193	−1.643	−0.902
	Rear drug consumption	−1.882	0.361	−2.594	−1.226
	Measurement equation				
	γ (coefficient of the latent variable)	9.798	0.026	9.747	9.850
Variance-covariance	σ_{11} (driver)	5.789	5.689	1.097	23.820
	$\sigma_{12} = \sigma_{21}$ (driver/front-seat passenger)	8.369	6.857	1.722	27.670
	$\sigma_{13} = \sigma_{13}$ (driver/rear-seat passenger)	6.235	5.208	0.685	18.900
	σ_{22} (front-seat passenger)	12.970	9.912	2.133	35.170
	$\sigma_{23} = \sigma_{23}$ (front-seat/rear-seat passengers)	10.910	10.430	0.815	37.020
	σ_{33} (rear-seat passenger)	10.920	12.550	0.316	42.230

[a] PC: Principal component.

vehicle influences the occupants' seatbelt use choices.

It is worth mentioning that while estimating the models, we found that the fixed parameters variants of all models had lower DIC compared to their random parameters variants implying that the unobserved heterogeneity in the effects of explanatory variables on seatbelt use choices is not statistically significant for this sample data. However, this finding may be an artefact of the random parameters specification in Bayesian statistics in which the Bayesian inference, by definition, accounts for the uncertainty in parameters via Bayes theorem. Further specifying the parameters to be random is just adding another level of hierarchy into the models and thus evaluating the statistical significance of such parameters (even if applicable) may not necessarily be indicative of the presence/absence of unobserved heterogeneity in data.

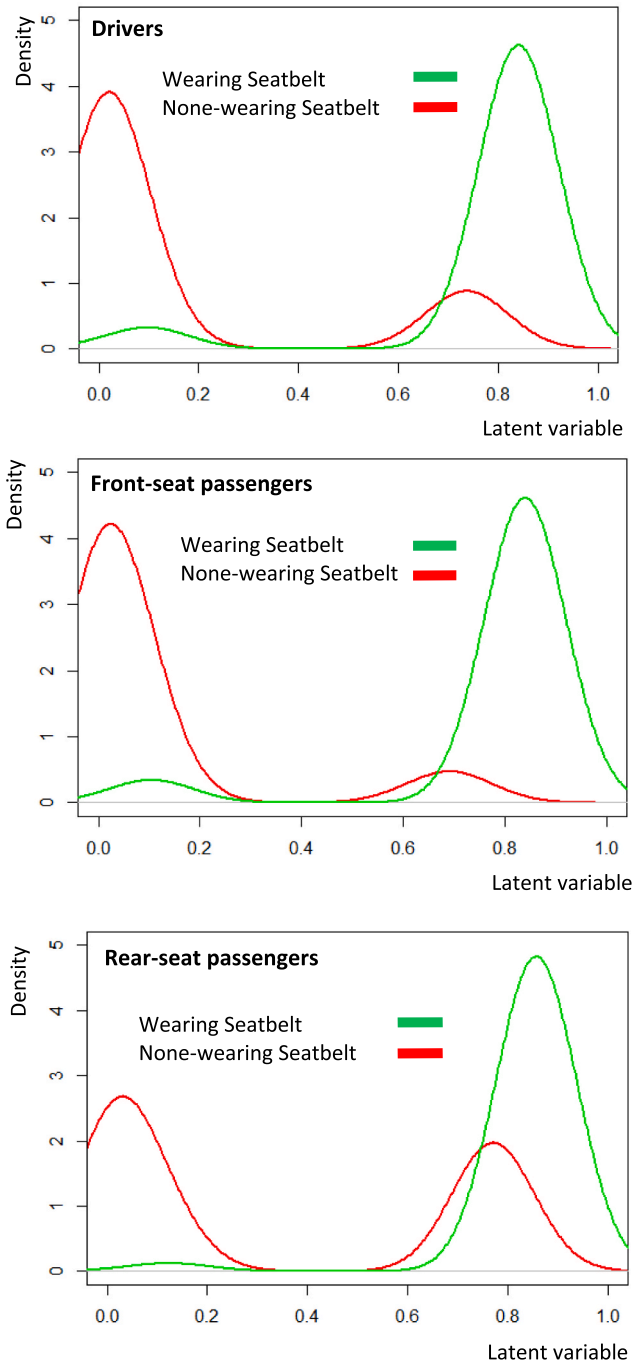

Fig. 2. Density plots of latent variables for drivers, front-seat and rear-seat passengers in the sample.

4.2. Model results

The multivariate model with unrestricted latent variables is selected for making inferences about the effects of explanatory variables on seatbelt use choices of drivers and passengers. The result of this model is presented in Table 4. According to the 95% credible intervals, drivers of larger vehicles (i.e. trucks, vans, and buses) with multiple occupants are more likely to use seatbelt than the drivers of smaller vehicles with multiple occupants (i.e. passenger cars). This finding is intuitive and indicates that drivers of larger vehicles (buses/trucks) are more likely to wear seatbelt compared to drivers of passenger cars. While this finding is in contrast with the seatbelt use rates of professional versus regular drivers in the literature (Afghari et al., 2020; Kim and Yamashita, 2007), it might reflect the difference in the seatbelt use behaviour of professional drivers in the presence and absence of other passengers (i.e. large vehicles with multiple versus single occupants). In addition, drivers are more likely to wear seatbelt during the day than during the night. This finding is also consistent with the previous findings in the literature (Chaudhary et al., 2005; Chaudhary and Preusser, 2006; Solomon et al., 2007; Tison et al., 2010; Vivoda et al., 2007) and might reflect the behaviour of high-risk individuals in not wearing seatbelt during the night (Noordzij et al., 1988).

The parameters of latent variables representing the atmosphere of the vehicle are positive for all vehicle occupants indicating that higher values of the latent variable are associated with increased likelihood of seatbelt use. This finding implies that there is indeed an unobserved factor in the vehicle that jointly influences collective seatbelt use behaviour of vehicle occupants. To better illustrate the effects of this unobserved factor on seatbelt use, the densities of the latent variables are plotted for seatbelt wearing and non-wearing occupants (Fig. 2). These density plots show that the mean of latent variables are substantially higher for those occupants who wear seatbelt (green curves) compared to those occupants who do not wear seatbelt (red curves). This finding implies that the higher proportion of occupants wear seatbelt in the vehicle, the more likely another occupant is to wear a seatbelt.

The above finding supports our hypothesis about the effects of the collective unobserved character in the vehicle (which we have named it 'the atmosphere of the vehicle') on seatbelt use choices. Another interesting finding is that the magnitude of the parameter of the latent variable for the rear-seat passenger (71.420) is larger than its counterparts for the driver and the front-seat passenger (62.880 and 38.750, respectively) implying that the rear-seat passenger is influenced to a higher extent by the atmosphere of the vehicle compared to the driver and the front-seat passenger. This finding reflects the varied effects of unobserved factors on vehicle occupants, depending on their seating position and is consistent with the findings of previous studies suggesting that there is a difference between seatbelt wearing behaviour of vehicle occupants in single-occupant versus multi-occupant vehicles (Drury and Drake, 2002; Hong, 1998).

The parameters of explanatory variables within the structural equations for the latent variables show that the first principal components of vehicle occupants are statistically significant in these equations and predict the latent variables. The parameters of these principal components are negative indicating that they have decreasing effects on the probability of seatbelt use choices. Since the actual values of the principal components do not provide direct interpretation about the determinants of the atmosphere, a correlation analysis is now conducted between the first principal components and the census tract data in order to shed more light on their effects on the vehicle atmosphere (Table 5).

The Pearson correlation coefficients reported in Table 5 show that higher percentage of old population (older than 60 years old) and white racial mix, higher percentage of commuting trips by car, higher average travel time, higher percentage of the population with college and bachelors education levels, higher income (average household income and income per capita), and higher vehicle

Table 5
Pearson correlation coefficients between census tract data and their first principal component.

Variable	First PC of the driver	First PC of the front-seat passenger	First PC of the rear-seat passenger
Total population density (1 000 person/km2)	0.405	0.436	0.474
Percentage population under 16 years old	0.118	0.083	0.131
Percentage population above 60	−0.103	−0.094	−0.132
Percentage of white race	−0.599	−0.633	−0.658
Percentage of commuting to work by car	−0.413	−0.452	−0.472
Percentage of commuting to work by car pool	0.371	0.369	0.344
Percentage of commuting to work by bus	0.528	0.546	0.553
Percentage of commuting to work by bike	0.092	0.086	0.119
Percentage of commuting to work by walk	0.376	0.361	0.392
Average travel time (hours)	−0.206	−0.223	−0.219
Average household size	0.057	−0.007	0.046
Percentage of high school education	0.625	0.631	0.600
Percentage of college education	−0.132	−0.159	−0.127
Percentage of bachelors education	−0.645	−0.627	−0.616
Average household income ($100,000)	−0.813	−0.825	−0.812
Income per capita	−0.756	−0.765	−0.752
Number of vacant houses	0.440	0.482	0.440
Proportion of households with 0 vehicle	0.748	0.747	0.756
Proportion of households with 1 vehicle	0.634	0.599	0.609
Proportion of households with 2 vehicles	−0.730	−0.745	−0.757
Proportion of households with 3 or more vehicles	−0.590	−0.620	−0.626

PC: Principal component.

ownership are all associated with lower values of the principal components and thus higher values of the latent variable in the vehicle (as a result of the negative association between principal components and atmosphere in the structural equations). Bearing in mind that the parameters of the latent variables are positive, these findings imply that the above census tract characteristics are ultimately associated with higher likelihood of seatbelt use choices. On the contrary, higher population density, higher percentage of young population (under 16 years old), larger average household size, and higher number of vacant houses are associated with higher values of the principal components, lower values of the latent variable, and lower likelihood of seatbelt use choices.

In addition to the principal components, vehicle occupants' age, alcohol and drug consumption are also statistically significant in the structural equations. The parameters of age is mostly positive, indicating that older occupants have a positive influence on the latent variable in general and increase the likelihood of using seatbelt. This finding is intuitive and in line with the previous findings in the literature (Calisir and Lehto, 2002; Glassbrenner et al., 2004) indicating that older people are more likely to wear seatbelt. On the contrary, the parameters of alcohol and drug consumption are negative indicating that they have decreasing effect on the latent variable, and thus result in lower likelihood of seatbelt use. These findings are also intuitive and consistent with the previous findings in the literature (Foss et al., 1994).

Overall, the above correlation coefficients and parameter estimates indicate the contribution of census tract variables, age, alcohol and drug consumption to a latent construct and consequently to the vehicle occupants' decisions to wear a seatbelt. We argue that because it is very difficult to measure or disentangle the components of this latent construct from crash reports or even ask them in surveys, the census tract attributes, age, alcohol and drug consumption are used as proxies in order to capture a part of this latent construct. While perhaps each of the correlation coefficients or parameter estimates does not have a particular interesting meaning, they are rather showing a broader effect –the effect of an unobserved factor that is latent to the analyst and yet is present in the vehicle.

As a final note, it is worth mentioning that we did try to include these variables directly in the utilities of seatbelt use choices but they were not statistically significant. The lack of their statistical significance in the utility functions and their statistical significance in the structural equations of the latent variables illuminates the main advantage of the proposed latent variable modelling methodology in this study.

5. Conclusions

The complexities of seatbelt use choice behaviour of individuals in multi-occupant vehicles have largely prevented a comprehensive understanding of these choices and their contributing factors. On the one hand, psychological and behavioural determinants of seatbelt use are difficult to measure and are not usually available. On the other hand, these factors may have shared influences on the seatbelt use choice behaviour of driver and passengers causing their choices to be interrelated. This study investigated these complexities by developing a comprehensive econometric model that explains seatbelt use behaviour of vehicle occupants and testing that model using data from Tennessee, United States.

Empirical testing of the proposed econometric model showed that it has substantially improved statistical fit compared to the alternative models and indicated that using a latent variable measured by observed seatbelt use choices and socioeconomic attributes can better explain the complexities of seatbelt use choices compared to directly using the observed seatbelt use choices. The results of this model showed that drivers of larger vehicles are more likely to use seatbelt in comparison with smaller vehicles. In addition, older front-seat passengers are more likely to use seatbelt, raising the flag for applying behavioural policies and countermeasures to the younger generation.

More importantly, we theoretically hypothesized and then empirically showed that there is a common underlying unobserved factor in a vehicle –we named it 'vehicle atmosphere'– that affects vehicle occupants' decisions to wear seatbelt. This unobserved factor measured by the proportion of self-excluded seatbelt use choice of vehicle occupants jointly influences seatbelt use of the driver and seatbelt use of other passengers. The impact of this unobserved factor is larger for the rear-seat passenger in comparison with the driver and the front-seat passenger. In addition, we found that higher proportion of old population and white racial mix, higher income per capita, higher education levels and higher vehicle ownership are highly correlated with this unobserved factor.

This study is not without limitations. From the methodological perspective, we did not examine the temporal variation in the seatbelt use choices of vehicle occupants. While these choices were assumed to be static in our model specifications, it is possible that vehicle occupants change their seatbelt use choice behaviour dynamically and with respect to the real-time behaviour of each other. Investigating the temporal variation in seatbelt use choices using proper methodological approaches such as dynamic discrete choice models is a worthy research direction given proper data exist (e.g., from naturalistic driving experiments).

From the empirical perspective, an important limitation of this study is that vehicle occupants' data have been extracted from crash reports and thus may not be a proper representative of all vehicle occupants in Tennessee. Although seatbelt use data have been validated with roadside observations, the data may still be subject to selectivity bias because risky individuals are more likely to be involved in crashes and thus are over-represented in crash data (Mannering et al., 2020). In other words, vehicle occupants who are involved in crashes may have different risk profiles than the ones whose data are obtained from roadside observations, despite the percentage of seatbelt use may be similar in these two samples. This selectivity bias may have affected the causal inference obtained from the data analysis in this study. As such, the findings of this study should be interpreted with caution. In particular, the effect of the atmosphere on seatbelt use might be more subtle for less-risky individuals who may have been under-represented in the crash data. Future research should repeat this study using roadside observations and a more random sample of vehicle occupants.

In addition and due to the lack of data, we only included residential location characteristics as proxies of individuals' behaviour in the models. Future research should collect merely behavioural data, test alternative constructs for the unobserved factor hypothesized in this study and validate our findings. As previous studies have shown that the overall atmosphere of a place might be related to social

influence (Lopez-Pintado and Watts, 2008), additional empirical data should be collected to further investigate what portion of this unobserved factor could be related to social influence. In addition, we did not examine the effects of transport mode on seatbelt use in multi-occupant vehicles. The seatbelt use choices of drivers and passengers may be significantly different in ridesharing modes (e.g., taxi) compared to the private vehicles. Future research should explore these differences. Finally, we only considered three-occupant vehicles in this study due to the small number of records of vehicles with more than three occupants. Further research is needed to compare the results with data consisting additional occupants which might reveal differential effects of variables and strengthen the atmosphere component in the analysis. In addition, the three-occupant vehicles obtained from crash data might not be a proper representative of all three-occupant vehicles. As a result, it is important to validate the findings of this study using data from observational studies that do not rely on crash occurrence because multi-occupant vehicles are more likely to be reported in crash databases (Chang and Mannering, 1998).

Author statement

A.P. Afghari, A.F. Imani, E. Papadimitriou: study conception and design; **A.M. Hezaveh:** data collection; **A.P. Afghari, A.F. Imani, E. Papadimitriou, P. van Gelder:** analysis and interpretation of results; **A.P. Afghari:** draft manuscript preparation. All authors reviewed the results and approved the final version of the manuscript.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgment

This study was funded by Delft University of Technology, Imperial College London and the University of Tennessee, Knoxville. The authors would like to acknowledge the anonymous reviewers for their insightful comments that helped improve the quality of this manuscript.

Appendix. Result of the principal component analysis

Table A1
Eigenvalues, proportion of explained variance and cumulative proportion of explained variance for the first five principal components

	Driver	Front-seat passenger	Rear-seat passenger
First principal component			
Eigenvalue	5.423	5.606	5.638
Proportion of explained variance	0.258	0.267	0.269
Cumulative proportion of explained variance	0.258	0.267	0.269
Second principal component			
Eigenvalue	2.818	2.694	2.721
Proportion of explained variance	0.134	0.128	0.130
Cumulative proportion of explained variance	0.392	0.395	0.398
Third principal component			
Eigenvalue	2.254	2.135	2.130
Proportion of explained variance	0.107	0.102	0.101
Cumulative proportion of explained variance	0.500	0.497	0.500
Fourth principal component			
Eigenvalue	1.766	1.778	1.763
Proportion of explained variance	0.084	0.085	0.084
Cumulative proportion of explained variance	0.584	0.581	0.583
Fifth principal component			
Eigenvalue	1.089	1.123	1.084
Proportion of explained variance	0.052	0.053	0.052
Cumulative proportion of explained variance	0.636	0.635	0.635

References

Abay, K.A., Paleti, R., Bhat, C.R., 2013. The joint analysis of injury severity of drivers in two-vehicle crashes accommodating seat belt use endogeneity. Transp. Res. Part B Methodol. 50, 74–89.
Afghari, A.P., Haque, M.M., Washington, S., Smyth, T., 2016. Bayesian latent class safety performance function for identifying motor vehicle crash black spots. Transport. Res. Rec. 2601, 90–98.

Afghari, A.P., Haque, M.M., Washington, S., Smyth, T., 2019a. Effects of globally obtained informative priors on bayesian safety performance functions developed for Australian crash data. Accid. Anal. Prev. 129, 55–65.
Afghari, A.P., Hezaveh, A.M., Haque, M.M., Cherry, C., 2020. A home-based approach to understanding seatbelt use in single-occupant vehicles in Tennessee: application of a latent class binary logit model. Accid. Anal. Prev. 146, 105743.
Afghari, A.P., Papadimitriou, E., Li, X., Kaye, S., Oviedo-Trespalacios, O., 2021. How much should a pedestrian be fined for intentionally blocking a fully automated vehicle? A random parameters Beta hurdle model with heterogeneity in the variance of the Beta distribution. Analytic Methods in Accident Research 100186.
Afghari, A.P., Washington, S., Haque, M.M., Li, Z., 2018. A comprehensive joint econometric model of motor vehicle crashes arising from multiple sources of risk. Analytic Methods in Accident Research 18, 1–14.
Afghari, A.P., Washington, S., Prato, C., Haque, M.M., 2019b. Contrasting case-wise deletion with multiple imputation and latent variable approaches to dealing with missing observations in count regression models. Analytic Methods in Accident Research 24, 100104.
Ali, M., Haidar, N., Ali, M.M., Maryam, A., 2011. Determinants of seat belt use among drivers in Sabzevar, Iran: a comparison of theory of planned behavior and health belief model. Traffic Inj. Prev. 12, 104–109.
Bhat, C.R., Srinivasan, S., 2005. A multidimensional mixed ordered-response model for analyzing weekend activity participation. Transp. Res. Part B Methodol. 39, 255–278.
Blincoe, L., Miller, T.R., Zaloshnja, E., Lawrence, B.A., 2015. The Economic and Societal Impact of Motor Vehicle Crashes, 2010. (Revised) (Report No. DOT HS 812 013). National Highway Traffic Safety Administration, Washington, DC.
Bolduc, D., Ben-Akiva, M., Walker, J., Michaud, A., 2005. Hybrid Choice Models with Logit Kernel: Applicability to Large Scale Models. Integrated Land-Use and Transportation Models: Behavioural Foundations. Elsevier, Oxford, pp. 275–302.
Byrne, D., Nelson, D., 1965. Attraction as a linear function of proportion of positive reinforcements. J. Pers. Soc. Psychol. 1 (6), 659. https://doi.org/10.1037/h0022073.
Calisir, F., Lehto, M.R., 2002. Young drivers' decision making and safety belt use. Accid. Anal. Prev. 34, 793–805.
Chang, L.-Y., Mannering, F.L., 1998. Predicting vehicle occupancies from accident data: an accident severity approach. Transport. Res. Rec. 1635, 93–104.
Chaudhary, N.K., Alonge, M., Preusser, D.F., 2005. Evaluation of the Reading, PA nighttime safety belt enforcement campaign: september 2004. J. Saf. Res. 36, 321–326.
Chaudhary, N.K., Preusser, D.F., 2006. Connecticut nighttime safety belt use. J. Saf. Res. 37, 353–358.
Cherry, C., Hezaveh, A.M., Noltenius, M., Khattak, A., Merlin, L., Dumbaugh, E., Ragland, D., Sandt, L., 2018. Completing the Picture of Traffic Injuries: Understanding Data Needs and Opportunities for Road Safety.
Cunill, M., Gras, M.E., Planes, M., Oliveras, C., Sullman, M.J., 2004. An investigation of factors reducing seat belt use amongst Spanish drivers and passengers on urban roads. Accid. Anal. Prev. 36, 439–445.
Dee, T.S., 1998. Reconsidering the effects of seat belt laws and their enforcement status. Accid. Anal. Prev. 30, 1–10.
Drury, C.G., Drake, M., 2002. Demographic effects of behavior modeling in seat belt use: analysis of 15,000 observations. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 46 (19), 1730–1733.
Eby, D.W., Molnar, L.J., Olk, M.L., 2000. Trends in driver and front-right passenger safety belt use in Michigan: 1984–1998. Accid. Anal. Prev. 32, 837–843.
Eluru, N., Bhat, C.R., 2007. A joint econometric analysis of seat belt use and crash-related injury severity. Accid. Anal. Prev. 39, 1037–1049.
Foss, R.D., Beirness, D.J., Sprattler, K., 1994. Seat belt use among drinking drivers in Minnesota. Am. J. Publ. Health 84, 1732–1737.
Geedipally, S.R., Lord, D., Dhavala, S.S., 2014. A caution about using deviance information criterion while modeling traffic crashes. Saf. Sci. 62, 495–498.
Gkritza, K., Mannering, F.L., 2008. Mixed logit analysis of safety-belt use in single-and multi-occupant vehicles. Accid. Anal. Prev. 40, 443–451.
Glassbrenner, D., Carra, J.S., Nichols, J., 2004. Recent estimates of safety belt use. J. Saf. Res. 35, 237–244.
Golob, T.F., Regan, A.C., 2002. Trucking industry adoption of information technology: a multivariate discrete choice model. Transport. Res. C Emerg. Technol. 10, 205–228.
Guevara, C.A., 2015. Critical assessment of five methods to correct for endogeneity in discrete-choice models. Transport. Res. Pol. Pract. 82, 240–254.
Guevara, C.A., Ben-Akiva, M., 2010. Addressing Endogeneity in Discrete Choice Models: Assessing Control-Function and Latent-Variable Methods. *Choice Modelling: the State-Of-The-Art and the State-Of-Practice*. Emerald Group Publishing Limited.
Guevara, C.A., Hess, S., 2019. A control-function approach to correct for endogeneity in discrete choice models estimated on SP-off-RP data and contrasts with an earlier FIML approach by Train & Wilson. Transp. Res. Part B Methodol. 123, 224–239.
Henry, R., Hidy, G., 1979. Multivariate analysis of particulate sulfate and other air quality variables by principal components-Part I: annual data from Los Angeles and New York. Atmos. Environ. 13, 1581–1596.
Hezaveh, A.M., Arvin, R., Cherry, C.R., 2019a. A geographically weighted regression to estimate the comprehensive cost of traffic crashes at a zonal level. Accid. Anal. Prev. 131, 15–24.
Hezaveh, A.M., Cherry, C.R., 2019a. Neighborhood-level factors affecting seat belt use. Accid. Anal. Prev. 122, 153–161.
Hezaveh, A.M., Cherry, C.R., 2019b. Neighborhood-level factors affecting seat belt use. Accid. Anal. Prev. 122, 153–161.
Hezaveh, A.M., Nordfjærn, T., Everett, J., Cherry, C.R., 2019b. The correlation between education, engineering, enforcement, and self-reported seat belt use in Tennessee: incorporating heterogeneity and time of day effects. Transport. Res. F Traffic Psychol. Behav. 66, 379–392.
Hodson-Walker, N., 1970. The value of safety belts: a review. Can. Med. Assoc. J. 102, 391.
Hong, S., et al., 1998. Effects of imitative behavior on seat belt usage: Three field observational studies. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 42 (15), 1093–1097.
Huitema, B.E., 1986. Autocorrelation in Behavioral Research. *Research Methods in Applied Behavior Analysis*. Springer.
Huitema, B.E., McKean, J.W., 1991. Autocorrelation estimation and inference with small samples. Psychol. Bull. 110, 291.
Jonah, B.A., Grant, B.A., 1985. Long-term effectiveness of selective traffic enforcement programs for increasing seat belt use. J. Appl. Psychol. 70, 257.
Kim, K., Yamashita, E.Y., 2007. Attitudes of commercial motor vehicle drivers towards safety belts. Accid. Anal. Prev. 39, 1097–1106.
Klair, A.A., Arfan, M., 2014. Use of seat belt and enforcement of seat belt laws in Pakistan. Traffic Inj. Prev. 15, 706–710.
Knapper, C., Cropley, A.J., Moore, R., 1976. Attitudinal factors in the non-use of seat belts. Accid. Anal. Prev. 8, 241–246.
Kouvonen, A., Oksanen, T., Vahtera, J., Stafford, M., Wilkinson, R., Schneider, J., Väänänen, A., Virtanen, M., Cox, S.J., Pentti, J., 2008. Low workplace social capital as a predictor of depression: the Finnish Public Sector Study. Am. J. Epidemiol. 167, 1143–1151.
Kulanthayan, S., Law, T., Raha, A., Radin, U.R., 2004. Seat belt use among car users in Malaysia. IATSS Res. 28, 19–25.
Lopez-Pintado, D., Watts, D.J., 2008. Social influence, binary decisions and collective dynamics. Ration. Soc. 20, 399–443.
Lord, D., Washington, S., 2018. Safe Mobility: Challenges, Methodology and Solutions. Emerald Publishing.
Lund, A.K., 1986. Voluntary seat belt use among US drivers: geographic, socioeconomic and demographic variation. Accid. Anal. Prev. 18, 43–50.
Mannering, F., Bhat, C.R., Shankar, V., Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Analytic Methods in Accident Research 25, 100113.
Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic methods in accident research 11, 1–16.
McFadden, D., 1973. Conditional Logit Analysis of Qualitative Choice Behavior. University of California Berkeley, US.
McFadden, D., 1981. Econometric models of probabilistic choice. In: Manski, D., McFadden (Eds.), A Structural Analysis of Discrete Data with Econometric Applications. The MIT Press, Cambridge, MA.
Merlin, L.A., Cherry, C.R., Mohamadi-Hezaveh, A., Dumbaugh, E., 2020. Residential accessibility's relationships with crash rates per capita. Journal of Transport and Land Use 13, 113–128.
Noordzij, P., Meester, A., Verschuur, W.L., 1988. Night-time driving: the use of seat-belts and alcohol. Ergonomics 31, 663–668.

Okamura, K., Fujita, G., Kihira, M., Kosuge, R., Mitsui, T., 2012. Predicting motivational determinants of seatbelt non-use in the front seat: a field study. Transport. Res. F Traffic Psychol. Behav. 15, 502–513.
Oviedo-Trespalacios, O., Afghari, A.P., Haque, M.M., 2020. A hierarchical Bayesian multivariate ordered model of distracted drivers' decision to initiate risk-compensating behaviour. Analytic Methods in Accident Research 26, 100121.
Passmore, J., Ozanne-Smith, J., 2006. Seatbelt use amongst taxi drivers in Beijing, China. Int. J. Inj. Contr. Saf. Promot. 13, 187–189.
Peltzman, S., 1975. The effects of automobile safety regulation. J. Polit. Econ. 83, 677–725.
Pickrell, T., Ye, T., 2009. Traffic Safety Facts (Research Note): Seat Belt Use in 2008–Demographic Results. National Highway Traffic Safety Administration/ Department of Transportation, Washington DC.
Ravulaparthy, S., Yoon, S.Y., Goulias, K.G., 2013. Linking elderly transport mobility and subjective well-being: a multivariate latent modeling approach. Transport. Res. Rec. 2382, 28–36.
Rentfrow, P.J., 2010. Statewide differences in personality: toward a psychological geography of the United States. Am. Psychol. 65, 548.
Russo, B.J., Savolainen, P.T., Schneider IV, W.H., Anastasopoulos, P.C., 2014. Comparison of factors affecting injury severity in angle collisions by fault status using a random parameters bivariate ordered probit model. Analytic methods in accident research 2, 21–29.
Sanko, N., Hess, S., Dumont, J., Daly, A., 2014. Contrasting imputation with a latent variable approach to dealing with missing income in choice models. Journal of choice modelling 12, 47–57.
Shakya, I., Shults, R.A., Stevens, M.R., Beck, L.F., Sleet, D.A., 2020. State-level seat belt use in the United States, 2011–2016: comparison of self-reported with observed use and use by fatally injured occupants. J. Saf. Res. 73, 103–109.
Solomon, M.G., Chaudhary, N.K., Preusser, D.F., 2007. Daytime and Nighttime Seat Belt Use at Selected Sites in New Mexico.
Suzuki, E., Yamamoto, E., Takao, S., Kawachi, I., Subramanian, S., 2012. Clarifying the use of aggregated exposures in multilevel models: self-included vs. self-excluded measures. PloS One 7, e51717.
Tipping, M.E., Bishop, C.M., 1999. Probabilistic principal component analysis. J. Roy. Stat. Soc. B 61, 611–622.
Tison, J., Williams, A.F., Chaudhary, N.K., 2010. Daytime and Nighttime Seat Belt Use by Fatally Injured Passenger Vehicle Occupants. Preusser Research Group, Inc.
Vivoda, J.M., Eby, D.W., St Louis, R.M., Kostyniuk, L.P., 2007. A direct observation study of nighttime safety belt use in Indiana. J. Saf. Res. 38, 423–429.
Washington, S., Karlaftis, M.G., Mannering, F., Anastasopoulos, P., 2020. Statistical and Econometric Methods for Transportation Data Analysis. CRC press.