Estimation of Factor Structured Covariance Mixed Logit Models

Jonathan James*

California Polytechnic State University

December 18, 2017

Abstract

Mixed logit models with normally distributed random coefficients are typically estimated under the extreme assumptions that either the random coefficients are completely independent or fully correlated. A factor structured covariance provides a middle ground between these two assumptions. However, because these models are more difficult to estimate, they are not frequently used to model preference heterogeneity. This paper develops a simple expectation-maximization algorithm for estimating mixed logit models when preferences are generated from a factor structured covariance. The algorithm is easy to implement for both exploratory and confirmatory factor models. The estimator is applied to stated-preference survey data from residential energy customers (Train, 2007). Comparing the fit across five different models, which differed in their assumptions on the covariance of preferences, the results show that all three factor specifications produced a better fit of the data than the fully correlated model measured by BIC and two out of three performed better in terms of AIC.

Keywords: Discrete Choice, Mixed Logit, EM Algorithm, Factor Models

JEL Classification Codes: C02, C13, C25, C35, C38

1 Introduction

The mixed logit with normally distributed random coefficients is one of the most widely used specifications of random utility models. Researchers typically estimate these models under the extreme

*Department of Economics, California Polytechnic State University. Email: jjames04@calpoly.edu. Homepage: http://www.calpoly.edu/~jjames04

assumptions that either the random coefficients are completely independent or fully correlated. While the fully correlated model nests the independence model, the number of parameters in the fully correlated model increases exponentially with the number of random coefficients, causing researchers in some situations to prefer the independence model because of computation time or because the model with fewer parameters performs better on model selection criteria like BIC. For example, Keane and Wasi (2013) estimate mixed logit models with normally distributed random coefficient on ten different stated-preference datasets and found that the uncorrelated specification performed better than the correlated specification using BIC in the majority of the datasets.

Although infrequently considered, a range of alternative models exists between the uncorrelated and fully correlated random coefficients logit. One intuitive alternative for modeling correlation in preferences is with a factor structured covariance. Factor models are used extensively in the social sciences to study correlated random variables and have been applied to modeling preference heterogeneity in discrete choice models in Elrod and Keane (1995), Keane and Wasi (2012) and Fiebig et al. (2010). In these models, individual preferences over product attributes are a function of a low-dimensional number of latent factors, which generates correlation in individual preferences. Factor models have a number of desirable benefits that extend naturally to modeling preferences. First, these models have substantially fewer parameters than models with a full covariance matrix. For example, with 10 product characteristics, a full covariance matrix contains 55 parameters while the covariance with a single factor is represented with only 20 parameters. Second, factor models offer a more parsimonious decomposition of preferences, where a large number of preferences over product attributes can be described by a handful of factors, which helps reveal the main features of the data. Finally, the factor structure is extremely flexible. Not only can researchers choose the number of factors to consider with exploratory factor analysis but also test hypothesis or consider model driven or theory based restrictions on the covariance through confirmatory factor analysis.

Despite the benefits, factor structured covariances are rarely used in mixed logit models in part due to their computational complexity. Like all mixed logit models, estimating a model with a factor structured covariance requires the maximization of an integrated likelihood function. However, maximization of the log-likelihood with a factor structured covariance is extremely slow

with quasi-Newton methods because the gradient of the log-likelihood is difficult to take analytically and optimization must rely on much slower numerical gradients.[1] The main issue is that in the fully correlated model there exists a mapping between the model parameters and the Cholesky components of the preference covariance matrix, facilitating a reformulation of the likelihood in terms of the Cholesky components where analytical gradients can be taken over these parameters. In the case of the factor structured covariance there is no mapping between the model parameters and the Cholesky components of the preference covariance matrix, so numerical gradients are typically the only option.

The contribution of this paper is to develop a simple estimator for factor structured covariance mixed logit models that can be used to estimate both exploratory and confirmatory factor models. The estimator is an extension of the expectation-maximization (EM) algorithms developed in Train (2007) and Train (2008). These papers show that mixed logit models can be easily estimated by taking draws from the mixing distribution and then updating the moments of the mixing distribution by computing weighted averages of the draws. Repeating this simple process yields maximum likelihood estimates of the model parameters. Given that the EM algorithm is widely used to estimate factor models (Rubin and Thayer, 1982), the algorithm in Train (2007) can be easily modified to allow for factor structured covariances. The algorithm easily generalizes to any factor structure without the need to derive or code problem specific gradients.

The algorithm is applied to the stated-preference survey on residential energy customers in Train (2007). In this data, customers are asked to make choices when faced with different energy pricing options and providers. While the fully correlated model performs better than the independence model, both a 1-factor and 2-factor model perform better in terms of BIC than the fully correlated model, providing evidence that correlation in preferences may be a function of a lower dimensional number of factors. To gain further insight into the structure of preferences, a confirmatory factor analysis is performed to test the hypothesis that consumer preference over the six attributes in the data set are a function of only two latent variables: a preference over the price components and

[1] In the influential paper, Fiebig et al. (2010) kindly provide MATLAB code for a number of mixed logit specifications with alternative covariance structures, which includes a factor analytic structure that relies on numerical gradients.

a preference over the supplier components. The results show that this theory based specification substantially outperforms all of the other models in part because it explains nearly the same amount of variation in the data with fewer parameters.

The remainder of the paper is organized as follows. Section 2 describes the mixed logit model with a factor structured covariance matrix. Section 3 discusses an example of a factor structured covariance using the model in Train (2007) and discusses the shortcomings of current estimation methods. Section 4 outlines the EM algorithm to estimate the model. Section 5 studies the performance of the algorithm with stated-preference data. Finally, Section 6 concludes.

2 Factor Structured Covariance Mixed Logit Model

In a mixed logit model, if consumer i chooses product j in choice situation t they obtain utility $U_{ijt} = x'_{ijt}\beta_i + \varepsilon_{ijt}$. Where x_{ijt} are the observed characteristics of product j, β_i are individual i's preferences over the product characteristics, and ε is an i.i.d random utility shock. Individual preferences are unobserved but are drawn from some known distribution $f(\cdot)$ that is parameterized by unknown Ψ, i.e., $\beta_i \sim f(\beta|\Psi)$. The goal of estimation is to recover maximum likelihood estimates of the parameters of the mixing distribution $\hat{\Psi}$. Since preferences are not observed, the likelihood of the observed data is written conditional on a given value of preferences and then integrated over the full distribution using $f(\cdot)$. Assuming there are J products to choose from and assuming ε is distributed type-I extreme value, the probability that i choose j in choice situation t given preferences β has the familiar logit expression, $P_{ijt}(\beta) = \exp(x'_{ijt}\beta)/(\sum_{j'=1}^{J}\exp(x'_{ij't}\beta))$. Letting $y_{ijt} = 1$ if i chooses j in choice situation t and zero otherwise and assuming T choice situations, the likelihood of i's choices conditional on preferences β is given as

$$L_i(\beta) = \prod_{t=1}^{T}\prod_{j=1}^{J}[P_{ijt}(\beta)]^{y_{ijt}} = \prod_{t=1}^{T}\prod_{j=1}^{J}\left[\frac{\exp(x'_{ijt}\beta)}{\sum_{j'=1}^{J}\exp(x'_{ij't}\beta)}\right]^{y_{ijt}} \tag{1}$$

Given n observations, the log-likelihood function is constructed by integrating the conditional likelihoods and summing over individuals: $LL(\Psi) = \sum_{i=1}^{n}\ln(\int L_i(\beta)f(\beta|\Psi)d\beta)$.

In the normally distributed random coefficients logit, $f(\cdot)$ is the probability density function

of a multivariate normal distribution with mean γ and covariance Σ, i.e. $\Psi = \{\gamma, \Sigma\}$. Letting K denote the dimension of x, if preferences are assumed to be uncorrelated, then Σ is diagonal and the model contains $2 \times K$ parameters. If preferences are assumed to be fully correlated then Σ is a symmetric positive definite matrix and the model contains $2 \times K + (K^2 - K)/2$ parameters.

A factor structured covariance offers an alternative way to model correlation in preferences and has been suggested by Fiebig et al. (2010) (footnote 23) as a compromise between the extreme assumptions of either uncorrelated or fully correlated preferences. Let θ_i denote an $M \times 1$ latent factor vector with $M \ll K$. The unobserved preferences for attribute $k \in \{1, 2, \ldots, K\}$ is assumed to be a linear function of the factors and a preference residual, $\beta_{ik} = \gamma_k + \lambda_k' \theta_i + \eta_{ik}$. Where λ_k are the factor loadings and η is the preference residual which is uncorrelated with the factors and the other preference residuals, i.e. $\text{Cov}(\theta, \eta_k) = 0$ for all k and $\text{Cov}(\eta_k, \eta_{k'}) = 0$ for all $k \neq k'$. The factors are assumed to be distributed multivariate normal with mean zero and covariance matrix Δ, $\theta_i \sim N(0, \Delta)$, and the preference residual for attribute k is normally distributed mean zero and attribute specific variance ω_k^2, $\eta_{ik} \sim N(0, \omega_k^2)$. Since θ and η are both mean zero, γ_k represents the mean value of preferences in the population, as was the case with the fully correlated model. Letting γ be the vector of preference means, Λ be the matrix of factor loadings, where the kth row contains λ_k', and $\eta_i = [\eta_{i1} \quad \eta_{i2} \quad \cdots \quad \eta_{iK}]'$ be the vector of preference residuals, then the vector of unobserved preferences for individual i is given by $\beta_i = \gamma + \Lambda\theta_i + \eta_i$, which is a draw from the multivariate distribution $\beta_i \sim N(\gamma, \Lambda\Delta\Lambda' + \Omega)$. Where Ω is a diagonal matrix with ω_k^2 in the kth row and kth column.

There are two approaches to factor models. In the first approach, exploratory factor models, the researcher places no *a priori* restrictions on the loadings Λ, which imposes no assumptions on the relationship between the factors and preferences. In this case, for identification, the factors are assumed to be uncorrelated and have unit variance, which normalizes Δ to the $M \times M$ identity matrix I, $\theta_i \sim N(0, I)$. In maximum likelihood estimation, the factor loadings are chosen to explain as much of the variation in the data as possible, under the restriction that preferences are only correlated through the low-dimensional factors. With an M-factor model, there are only $2 \times K + K \times M$ parameters, which requires drastically fewer parameters than a model with a full

covariance.[2] For example, with 20 product characteristics the fully correlated model contains 230 parameters, while a 2-factor model only contains 80.

The second approach to factor models is confirmatory factor analysis. In these models, the covariance of the factors is assumed to be unconstrained and the researcher places *a priori* restrictions on the factor loadings by fixing some to zero. These zeros control how the factors map to preferences. This class of factor models offer a tremendous amount of leeway to impose structure on the preference model, where the researcher not only chooses the number of factors but also which factors will load on which preferences. The normalizations required to identify these models are discussed in the next section.

3 Example

To illustrate the use of a factor structured covariance and motivate the estimator, we consider the residential energy customer data studied in Train (2007), which is also used in Huber and Train (2001) and Revelt and Train (1999), as an application. In this model, individual customers make choices among four possible residential energy suppliers where each supplier is represented by six attributes. The first three attributes describe the pricing arrangements, which are mutually exclusive, where suppliers offered either fixed pricing, time of day pricing, or seasonal pricing. The type of pricing arrangement as well as the prices charged are represented by the variables pf, a fixed kWh price, *tod*, stated prices that differed by time of day, and *seas*, stated prices that differed by season. Attributes four and five describe the status of the supplier, which included *loc*, an indicator if they where a local utility, and *wk*, an indicator if they were a well-known company other than the local utility. The final attribute was contract length, *cl*, which stated how long the supplier was obligated to provide service at the stated price and price arrangement.

A customer is described by a vector of preferences over these six attributes $\beta_i = [\beta_i^{pf} \quad \beta_i^{tod} \quad \beta_i^{seas} \quad \beta_i^{loc} \quad \beta_i^{wk} \quad \beta$ The objective of the factor structure is to reduce the relationship of these preferences to a lower-dimensional vector of factors. In a confirmatory factor model, the researcher brings to the model outside information to place more structure on the factors by setting some of the loadings to be

[2] K parameters each for γ and Ω and $K \times M$ parameters in the Λ matrix.

zero. This approach can be viewed as making the factors more interpretable, which address one of the main criticisms of exploratory factor models. In random utility models, one possibility is to stratify the factors along the same dimension as the natural stratification of the attributes. Let $\theta_i = [\theta_i^{pricing} \quad \theta_i^{supplier}]'$ be a vector of latent factors with the first factor representing individual i's responsiveness to attributes related to pricing and the second factor representing their responsiveness to attributes related to the supplier. The preferences relating to price (pf, tod, and $seas$) will only load on the first factor and the preferences relating to the supplier (loc and wk) will only load on the second factor, while the remaining attribute, cl, will load on both since it is not clear *a priori* to which stratification this attribute belongs. Under these assumptions, the preferences are a linear function of the factors plus an independent preference error, η, and represented by the system of equations:

$$\beta_i = \begin{bmatrix} \beta_i^{pf} \\ \beta_i^{tod} \\ \beta_i^{seas} \\ \beta_i^{loc} \\ \beta_i^{wk} \\ \beta_i^{cl} \end{bmatrix} = \gamma + \begin{bmatrix} \lambda_1^{pf} & 0 \\ \lambda_1^{tod} & 0 \\ \lambda_1^{seas} & 0 \\ 0 & \lambda_2^{loc} \\ 0 & \lambda_2^{wk} \\ \lambda_1^{cl} & \lambda_2^{cl} \end{bmatrix} \theta_i + \begin{bmatrix} \eta_i^{pf} \\ \eta_i^{tod} \\ \eta_i^{seas} \\ \eta_i^{loc} \\ \eta_i^{wk} \\ \eta_i^{cl} \end{bmatrix} = \gamma + \Lambda\theta_i + \eta$$

While it may appear that this specification precludes any relationship between β_i^{pf} and β_i^{loc} because these preferences load on different factors, this is not true because in the confirmatory factor model, the factors are allowed to be correlated, thus these preferences are related though the correlation of $\theta_i^{pricing}$ and $\theta_i^{supplier}$. We assume that $\theta_i \sim N(0, \Delta)$ where Δ is a full covariance matrix. Because we are not normalizing the variance of the factors to unity, this model is only identified if we set at least one of the loadings for each of the factors equal to one. In this example we could set $\lambda_1^{pf} = 1$ and $\lambda_2^{loc} = 1$. This model contains 20 parameters, 6 parameters in γ (the preference means), 6 parameters in Ω (the variance of the preference residuals), 5 parameters in Λ (the factor loadings), and 3 unique parameters in Δ (the covariance of the factors).

3.1 Difficulties of Gradient Based Maximum Likelihood Estimation

With a factor structured covariance, preferences are drawn from a $N(\gamma, \Lambda\Delta\Lambda' + \Omega)$. Letting $\Psi \in \{\gamma, \Lambda, \Omega, \Delta\}$ denote the model parameters, the log-likelihood for the factor structured mixed logit model is

$$LL(\Psi) = \sum_{i=1}^{n} \ln \left(\int_{\beta} L_i(\beta) f(\beta|\gamma, \Lambda\Delta\Lambda' + \Omega) d\beta \right)$$

Where again $f(\cdot)$ is the probability density function of a multivariate normal distribution. Similar to all mixed logit models, the integral in this equation does not have a closed form and must be simulated in estimation. The unique challenge to the factor structured mixed logit is that the likelihood cannot be re-written as a function of the Cholesky components of the variance of β as is common in the case of normally distributed random coefficients. This re-parameterization in terms of the Cholesky factors is done to make analytical gradients of the log-likelihood possible to avoid extremely slow numerical approximations to the gradient in estimation. Because there is no mapping between the Cholesky components of the variance of β and the parameters of the model, estimation of the model must rely on slower numerical gradients.

One possible workaround to the lack of analytical gradients is to re-write the log-likelihood to explicitly integrate over both the unobserved preferences as well as the unobserved factors as

$$LL(\Psi) = \sum_{i=1}^{n} \ln \left(\int_{\beta} \int_{\theta} L_i(\beta) f(\beta|\gamma + \Lambda\theta, \Omega) f(\theta|0, \Delta) d\theta d\beta \right)$$

Where $f(\cdot)$ again is the probability density function of a multivariate normal distribution with $\beta|\theta \sim N(\gamma + \Lambda\theta, \Omega)$ and $\theta \sim N(0, \Delta)$. The variance of both of these random variables can be written as their Cholseky components and all of the parameters of the model can be brought into the logit kernel. While this strategy makes possible analytical gradients, deriving and coding the gradient is much more difficult than conventional normally distributed random coefficients. Given the flexibility of the factor model, with this approach, if the researcher would like to compare different model specifications they may be forced to derive and code many different model specific gradients. Perhaps the greater issue with this approach is that it requires a higher dimensional

integral than the original problem as both β and θ must be simulated.

The next section proposes an estimator for factor covariance mixed logit models based on the EM algorithm that avoids numerical and analytical gradients of the original log-likelihood altogether, does not increase the dimension of integration, and can be easily applied to any factor specification with virtually no modification to the algorithm.

4 The Estimator

This section outlines an expectation-maximization (EM) algorithm for estimating factor structured covariance mixed logit models. Train (2007) demonstrates how the EM algorithm can be use to estimate mixed logit models with normally distributed random coefficients, while Train (2008) outlines how the EM algorithm can be used to estimate mixed logit models with a variety of non-parametric mixing distributions. Because the estimator developed in this paper is closely related to these other methods, this section begins with a brief review of current methods that rely on the EM algorithm and then subsequently outlines the proposed algorithm.

4.1 Review of EM Algorithm for ML Estimation of Mixed Logit Models

The EM algorithm is an iterative procedure that is initialized with starting values, $\Psi^{(0)}$. Rather than directly maximizing the log-likelihood function $LL(\Psi)$, the EM algorithm forms a surrogate function $Q(\Psi|\Psi^{(0)})$ that satisfies a tangency condition, $LL(\Psi^{(0)}) = Q(\Psi^{(0)}|\Psi^{(0)})$, and a bounding condition, $LL(\Psi) \geq Q(\Psi|\Psi^{(0)})$ for all values of Ψ. Because of these two conditions, choosing $\Psi^{(1)} = \text{argmax}_{\Psi}\, Q(\Psi|\Psi^{(0)})$ necessarily guarantees an improvement in the likelihood, i.e. $LL(\Psi^{(1)}) \geq LL(\Psi^{(0)})$. Iterating this process, with $\Psi^{(m+1)} = \text{argmax}_{\Psi}\, Q(\Psi|\Psi^{(m)})$ for $m = 0, 1, 2, \ldots$ leads to a local maximum of the log-likelihood function. The algorithm is stopped when the absolute or percentage change in the parameters or log-likelihood is within some pre-specified tolerance. The appeal of the EM algorithm is its simplicity. The surrogate function $Q(\Psi|\Psi^{(m)})$ is easy to form and more importantly easy to maximize.

For mixed logit models, Train (2008) shows that the EM surrogate function is

$$\begin{aligned} Q(\Psi|\Psi^{(m)}) &= \sum_{i=1}^{n} \int \ln\left[L_i(\beta) f(\beta|\Psi)\right] h_i(\beta|\Psi^{(m)}) d\beta \\ &= \sum_{i=1}^{n} \int \ln\left[L_i(\beta)\right] h_i(\beta|\Psi^{(m)}) d\beta + \sum_{i=1}^{n} \int \ln\left[f(\beta|\Psi)\right] h_i(\beta|\Psi^{(m)}) d\beta \end{aligned} \tag{2}$$

The first term in the surrogate function does not contain Ψ and drops out in maximization. The second term in the surrogate function is equivalent to the log-likelihood of the mixing distribution parameters, Ψ, in the case where β is observed, which is typically straightforward to maximize. The only difference from the complete data log-likelihood case is that the surrogate function weights the observed data points using the individual densities $h_i(\beta|\Psi^{(m)})$. These densities are defined as the posterior distribution of the unobserved β_i, given individual i's observed choices and the current iterations value of the parameters $\Psi^{(m)}$ and is defined as

$$h_i(\beta|\Psi^{(m)}) = \frac{L_i(\beta) f(\beta|\Psi^{(m)})}{\int L_i(\beta') f(\beta'|\Psi^{(m)}) d\beta'}$$

Note that the denominator in these densities is the integrated likelihood function. Just as maximum likelihood requires numerical simulation of this integral so does the EM algorithm. Train (2007) proposes approximating the integral in Eq. (2) at iteration m by taking R draws of β for each individual i from the mixing distribution $f(\beta|\Psi^{(m)})$. The rth draw for person i is labeled $\beta_{ir}^{(m)}$ and the weight for this observation is calculated as $w_{ir}^{(m)} = L_i(\beta_{ir}^{(m)}) / \left(\sum_{r'=1}^{R} L_i(\beta_{ir'}^{(m)})\right)$. Using these draw, the simulated surrogate function, dropping the first term, is defined as

$$Q(\Psi|\Psi^{(m)}) = \sum_{i=1}^{n} \sum_{r=1}^{R} w_{ir}^{(m)} \ln\left[f(\beta_{ir}^{(m)}|\Psi)\right] \tag{3}$$

As we did not specify the mixing distribution, this simulated surrogate function is applicable to any mixing distribution, which includes the factor structured covariance. If the maximum likelihood estimates of the parameters of the mixing distribution have a closed form solution in the complete data case, then the parameters that maximize Eq. (3) will as well. This is one of the key insights of the application of the EM algorithm to the normally distributed random coefficients in Train

(2007) discussed next.

4.2 EM Algorithm for Normally Distributed Random Coefficients: Train (2007)

Under the assumptions of normally distributed random coefficients and full covariance, the parameters Ψ include the mean γ and the covariance Σ, where $\beta \sim N(\gamma, \Sigma)$. The surrogate function in Eq. (3) takes the form

$$\begin{aligned} Q(\Psi|\Psi^{(m)}) =& \sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)} \ln\left[\frac{1}{\sqrt{(2\pi)^K|\Sigma|}}\exp\left(-\frac{1}{2}(\beta_{ir}^{(m)}-\gamma)'\Sigma^{-1}(\beta_{ir}^{(m)}-\gamma)\right)\right] \\ =& -\frac{1}{2}\sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)}\left[\ln(|\Sigma|) + (\beta_{ir}^{(m)}-\gamma)'\Sigma^{-1}(\beta_{ir}^{(m)}-\gamma) + K\ln(2\pi)\right] \end{aligned}$$

The main insight in Train (2007) is that this is the likelihood function of a multivariate normal distribution with weighted observations on β. This function is easily maximized by setting the parameters equal to their sample analogs, i.e. $\gamma^{(m+1)} = \overline{\beta}$ and $\Sigma^{(m+1)} = C_{\beta\beta}$ where

$$\text{Sample mean: } \overline{\beta} = (1/n)\textstyle\sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)}\beta_{ir}^{(m)} \tag{4}$$

$$\text{Sample covariance: } C_{\beta\beta} = (1/n)\textstyle\sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)}(\beta_{ir}^{(m)}-\overline{\beta})(\beta_{ir}^{(m)}-\overline{\beta})' \tag{5}$$

Forming a new surrogate function with these updated parameters and repeating this process leads to the maximum likelihood estimates of the normally distributed random coefficients model with full covariance (Train, 2007).

4.3 EM Algorithm for Factor Structured Random Coefficients

We now show how the EM algorithm can be extended to a factor structured covariance, which is the main contribution of this paper. For the factor structured covariance mixed logit in Section 2, the model parameters include $\Psi = \{\gamma, \Lambda, \Delta, \Omega\}$ and while β continues to be normally distributed its distribution is now $N(\gamma, \Lambda\Delta\Lambda' + \Omega)$. Given parameters $\Psi^{(m)} = \{\gamma^{(m)}, \Lambda^{(m)}, \Delta^{(m)}, \Omega^{(m)}\}$, as before we draw R values of β from $\beta_{ir}^{(m)} \sim N(\gamma^{(m)}, \Lambda^{(m)}\Delta^{(m)}\Lambda^{(m)\prime} + \Omega^{(m)})$ and compute the weights $w_{ir}^{(m)}$.

The surrogate function becomes

$$Q(\Psi|\Psi^{(m)}) = \sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)} \ln\left[f(\beta_{ir}^{(m)}|\gamma, \Lambda\Delta\Lambda' + \Omega)\right] \tag{6}$$

Eq. (6) is the log-likelihood function of a traditional factor model with observed values of β. As with the full covariance model, the maximum likelihood estimates of γ is the sample average using the weighted values of the β's, so $\gamma^{m+1} = \overline{\beta}$, where $\overline{\beta}$ is defined in Eq. (4). Unfortunately, no closed form solution exists for the other parameters of the model, so Eq. (6) needs to be numerically optimized with respect to Λ, Δ, and Ω. Maximum likelihood estimation of factor models is a well-studied problem dating back to Lawley (1940). These methods are relatively fast because they only operate on the observed sample covariance not the individual observed values of β, so their convergence time does not depend on the sample size or the number of simulated draws. Given that most statistical software provide built-in algorithms to estimate factor models via maximum likelihood, one possibility to maximizing Eq. (6) would be to use these built-in functions. The shortcomings of this approach are two-fold. First, most built-in functions typically do not allow user-defined starting values. This is unfortunate because good starting values ($\Psi^{(m)}$) are available, which would shorten the numerical optimization. Even worse, given that Eq. (6) possibly contains many local maxima, it is not guaranteed, without checking, that a built-in function would find a solution that actually improves the surrogate function, i.e. $Q(\Psi^{(m+1)}|\Psi^{(m)}) \geq Q(\Psi^{(m)}|\Psi^{(m)})$. If the built-in algorithm fails, then Eq. (6) may need to be maximized multiple times to insure the ascent property of the EM algorithm. The second shortcoming of relying on a built-in function is that they tend to focus on exploratory factor analysis and are difficult to customize, preventing researchers from exploiting the full power of these methods.

To update the factor parameters in a very simple and completely flexible way, we will take advantage of the fact that a popular method to estimate factor models is the EM algorithm. Rubin and Thayer (1982) outline a simple algorithm based on EM for both exploratory and confirmatory factor analysis that relies entirely on least squares operations. Using an inner EM algorithm to optimize Eq. (6) has a number of benefits. First, even a single iteration of the inner loop is

guaranteed to improve the surrogate function, which is all that is necessary to guarantee the ascent property of the outer algorithm. This gives the researcher more flexibility to either shorten the inner optimization to quickly move on to advancing the outer routine or to spend more resources precisely finding the maximum of the surrogate function. Second, the EM iterations are very easy to code and can generalize to any factor structure. This is an important point given many maximum likelihood routines for exploratory factor models do not extend to confirmatory factor models. Finally, the entire inner EM algorithm is made on computations of the observed sample covariance of preferences $C_{\beta\beta}$ defined in Eq. (5), so while only a single iteration of the algorithm is sufficient, in practice the computational cost of each iteration is so cheap that iterating to find the maximum of Eq. (6) leads to an overall faster converging algorithm.

The main principle with the EM algorithm for factor models (Rubin and Thayer, 1982) is that given that $\beta_{ik} = \gamma_k + \lambda_k'\theta_i + \eta_{ik}$ is linear in θ, if β and θ where observed, then λ_k could be estimated with ordinary least squares. However, in Eq. (6), although β is treated as observed data, θ remains unobserved, thus we need one intermediate step. The goal is to calculate from the data the empirical covariance of θ, labeled $C_{\theta\theta}$ and the empirical covariance of the factors and preferences, $C_{\theta\beta}$. Then we can apply the normal equations of least squares to get $\Lambda = C_{\theta\theta}^{-1} C_{\theta\beta}$. Choosing Λ in this way does not maximize the surrogate function in Eq. (6), however it is guaranteed to improve the function. As seen below $C_{\theta\theta}$ and $C_{\theta\beta}$ are both functions of the previous iterations value of Λ. Rubin and Thayer (1982) show that by repeating this process using the new value of Λ, the parameters will eventually converge to the values that maximize Eq. (6).

To derive the steps of the inner EM algorithm that optimizes Eq. (6), let $\psi^{(l)} = \{\Lambda^{(l)}, \Delta^{(l)}, \Omega^{(l)}\}$ denote the parameter values at the lth iteration of the inner optimization. Given the assumptions

of normality we can define the following conditional moments.[3]

$$\begin{aligned}
\mathrm{E}(\theta|\beta,\psi^{(l)}) =& d^{(l)}(\beta-\gamma) \\
\mathrm{Var}(\theta|\beta,\psi^{(l)}) =& D^{(l)} \\
\mathrm{E}(\theta\theta'|\beta,\psi^{(l)}) =& D^{(l)} + d^{(l)}(\beta-\gamma)(\beta-\gamma)'d^{(l)\prime}
\end{aligned} \tag{7}$$

Where

$$\begin{aligned}
d^{(l)} &= \Delta^{(l)}\Lambda^{(l)\prime}\left(\Lambda^{(l)}\Delta^{(l)}\Lambda^{(l)\prime}+\Omega^{(l)}\right)^{-1} \\
D^{(l)} &= \Delta^{(l)} - d^{(l)}\Lambda^{(l)}\Delta^{(l)}
\end{aligned}$$

Using these conditional moments and replacing γ with the sample average $\overline{\beta}$, we can compute the sample covariance of the unobserved factors conditional on $\psi^{(l)}$ as

$$\begin{aligned}
C_{\theta\theta} =& \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)}\,\mathrm{E}(\theta_{ir}\theta_{ir}'|\beta_{ir}^{(m)},\psi^{(l)}) \\
=& \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)}\left[D^{(l)} + d^{(l)}(\beta_{ir}^{(m)}-\overline{\beta})(\beta_{ir}^{(m)}-\overline{\beta})'d^{(l)\prime}\right] \\
=& D^{(l)} + d^{(l)}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)}(\beta_{ir}^{(m)}-\overline{\beta})(\beta_{ir}^{(m)}-\overline{\beta})'\right]d^{(l)\prime} \\
=& D^{(l)} + d^{(l)}C_{\beta\beta}d^{(l)\prime}
\end{aligned}$$

[3]These equations stem from the fact that the joint distribution of the factors and preferences is

$$\begin{bmatrix}\theta\\ \beta\end{bmatrix} \sim N\left(\begin{bmatrix}0\\ \gamma\end{bmatrix}, \begin{bmatrix}\Delta & \Delta\Lambda' \\ \Lambda\Delta & \Lambda\Delta\Lambda'+\Omega\end{bmatrix}\right)$$

Given the assumptions of normality, the moments of the distribution of the unobserved factors conditional on observed preferences are

$$\begin{aligned}
\mathrm{E}(\theta|\beta) =& \underbrace{\Delta\Lambda'\left(\Lambda\Delta\Lambda'+\Omega\right)^{-1}}_{d}(\beta-\gamma) \\
\mathrm{Var}(\theta|\beta) =& \Delta - \Delta\Lambda'\left(\Lambda\Delta\Lambda'+\Omega\right)^{-1}\Lambda\Delta \\
=& \underbrace{\Delta - d\Lambda\Delta}_{D}
\end{aligned}$$

$$
\begin{aligned}
C_{\theta\beta} =& \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)}\,\mathrm{E}(\theta_{ir}(\beta_{ir}^{(m)}-\overline{\beta})'|\beta_{ir}^{(m)},\psi^{(l)}) \\
=& \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)}\left[d^{(l)}(\beta_{ir}^{(m)}-\overline{\beta})(\beta_{ir}^{(m)}-\overline{\beta})'\right] \\
=& d^{(l)}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)}(\beta_{ir}^{(m)}-\overline{\beta})(\beta_{ir}^{(m)}-\overline{\beta})'\right] \\
=& d^{(l)}C_{\beta\beta}
\end{aligned}
$$

Using these covariances, if there are no constraints on the factor loading matrix, the update of the factor loadings is

$$\Lambda^{(l+1)} = C_{\theta\theta}^{-1}C_{\theta\beta}$$

If some of the factor loadings have constraints, then those rows of Λ need to be updated separately. Let $\mathtt{I}_k$ be an $M\times 1$ indicator vector for the elements of λ_k that need to be estimated. $!\mathtt{I}_k$ is an $M\times 1$ vector of indicators for the components that are normalized. We update the estimated parameters only with

$$\lambda_k^{(l+1)}\{\mathtt{I}_k\} = C_{\theta\theta}\{\mathtt{I}_k,\mathtt{I}_k\}^{-1}\left[C_{\theta\beta}\{\mathtt{I}_k,k\} - C_{\theta\theta}\{\mathtt{I}_k,!\mathtt{I}_k\}\lambda_k^{(l)}\{!\mathtt{I}_k\}\right] \tag{8}$$

Where the notation $\lambda_k^{l+1}\{\mathtt{I}_k\}$ signifies selecting only the estimated values of vector λ_k and $C_{\theta\theta}\{\mathtt{I}_k,\mathtt{I}_k\}$ signifies selecting the relevant rows and columns of the matrix $C_{\theta\theta}$.

Next, $\Omega^{(l+1)}$ is chosen as the sample variance of the factor residual given $\Lambda^{(l+1)}$

$$
\begin{aligned}
\Omega^{(l+1)} =& \mathrm{diag}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)}\,\mathrm{E}\left[(\beta_{ir}^{(m)}-\overline{\beta}-\Lambda^{(l+1)}\theta_{ir})(\beta_{ir}^{(m)}-\overline{\beta}-\Lambda^{(l+1)}\theta_{ir})'\middle|\beta_{ir}^{(m)},\psi^{(l)}\right]\right) \\
=& \mathrm{diag}\Bigg(\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)}\Big[(\beta_{ir}^{(m)}-\overline{\beta})(\beta_{ir}^{(m)}-\overline{\beta})' - \Lambda^{(l+1)}\,\mathrm{E}(\theta_{ir}(\beta_{ir}^{(m)}-\overline{\beta})'|\beta_{ir}^{(m)},\psi^{(l)}) \\
& - \left(\Lambda^{(l+1)}\,\mathrm{E}(\theta_{ir}(\beta_{ir}^{(m)}-\overline{\beta})'|\beta_{ir}^{(m)},\psi^{(l)})\right)' + \Lambda^{(l+1)}\,\mathrm{E}(\theta_{ir}\theta_{ir}'|\beta_{ir}^{(m)},\psi^{(l)})\Lambda^{(l+1)'}\Big]\Bigg) \\
=& \mathrm{diag}\left(C_{\beta\beta} - \Lambda^{(l+1)}C_{\theta\beta} - (\Lambda^{(l+1)}C_{\theta\beta})' + \Lambda^{(l+1)}C_{\theta\theta}\Lambda^{(l+1)'}\right)
\end{aligned}
$$

In an exploratory factor analysis, the covariance of the factors is normalized to the identity matrix so we set $\Delta^{(l+1)} = I$. If the covariance of the factors is not normalized, then we choose Δ to match the sample covariance, which has already been calculated, so $\Delta^{(l+1)} = C_{\theta\theta}$.

The results in Rubin and Thayer (1982) show that re-computing $d^{(l+1)}$ and $D^{(l+1)}$ with these updated parameters and repeating these calculations will eventually lead to a maximum of Eq. (6). Using this strategy to maximize the mixed logit EM surrogate function, the complete algorithm is outlined in Table 1. Standard errors can be computed either with bootstrapping or from the estimate of the information matrix suggested by Ruud (1991), which is the cross-product of the individual components of the gradient of the Q-function at the maximum likelihood solution. Decomposing the surrogate function, Eq. (6), as $Q(\Psi|\hat{\Psi}) = \sum_{i=1}^{n} Q_i(\Psi|\hat{\Psi})$, where $\hat{\Psi}$ denotes the maximum likelihood solution, Ruud (1991) suggests estimating the parameter variance-covariance matrix using $[\sum_{i=1}^{n}(\partial Q_i(\Psi|\hat{\Psi})/\partial\Psi)(\partial Q_i(\Psi|\hat{\Psi})/\partial\Psi)']^{-1}$. Although the gradient of Eq. (6) could be derived analytically, this function only contains the probability density function of a multivariate normal, so it is likely easier and faster to compute the derivative numerically to calculate the standard errors.

4.4 Discussion

In terms of the speed of the overall algorithm, it is not obvious whether it is better to optimize the surrogate function by iterating steps (4a)-(4d) until convergence or simply taking a pre-determined number of steps. In many other situations, performing numerical optimization at each iteration of the EM algorithm can be very problematic. For example, James (2017) shows that avoiding the numerical optimization necessary for the EM implementation of a mixed logit model with fixed coefficients can reduce computation times nearly 8-fold in certain situations. However, these concerns are not relevant in this situation because the computational cost in steps (4a)-(4d) are extremely low. As a general rule in the EM algorithm, the less precisely the surrogate function is maximized, the more iterations are involved in the overall algorithm. If the computation costs are very low in the inner optimization, it may make sense to spend more resources precisely maximizing the surrogate function to reduce the number of iterations of the overall algorithm. This is indeed

Table 1: EM Algorithm for Factor Structured Random Coefficients Logit

— Initialize with starting values $\Psi^{(0)} = \{\gamma^{(0)}, \Lambda^{(0)}, \Omega^{(0)}, \Delta^{(0)}\}$
— Repeat (1)-(4) until converged, e.g., $||\Psi^{(m+1)} - \Psi^{(m)}||_\infty < \kappa_1$

(1) Form surrogate function

(1a) For each i, take R draw of β labeled $\beta_{ir}^{(m)}$ with $\beta_{ir}^{(m)} \sim N(\gamma^{(m)}, \Lambda^{(m)}\Delta^{(m)}\Lambda^{(m)\prime} + \Omega^{(m)})$

(1b) Calculate conditional likelihood $L_i(\beta_{ir}^{(m)})$ for all i and r using Eq. (1)

(1c) Create weights $w_{ir}^{(m)} = L_i(\beta_{ir}^{(m)})/(\sum_{r'=1}^{R} L_i(\beta_{ir'}^{(m)}))$ for each i and r

(2) Sample moments of preferences

(2a) Sample mean $\overline{\beta} = (1/n)\sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)}\beta_{ir}^{(m)}$

(2b) Sample covariance $C_{\beta\beta} = (1/n)\sum_{i=1}^{n}\sum_{r=1}^{R} w_{ir}^{(m)}(\beta_{ir}^{(m)} - \overline{\beta})(\beta_{ir}^{(m)} - \overline{\beta})'$

(3) Update $\gamma^{(m+1)} = \overline{\beta}$

(4) Update $\Lambda^{(m+1)}$, $\Delta^{(m+1)}$, and $\Omega^{(m+1)}$ using inner EM algorithm

– Initialize with $\psi^{(0)} = \{\Lambda^{(m)}, \Omega^{(m)}, \Delta^{(m)}\}$
– Repeat (4a)-(4d) for fixed iterations or until converged, e.g., $||\psi^{(l+1)} - \psi^{(l)}||_\infty < \kappa_2$

(4a) Calculate samples moments

· $d^{(l)} = \Delta^{(l)}\Lambda^{(l)\prime}\left(\Lambda^{(l)}\Delta^{(l)}\Lambda^{(l)\prime} + \Omega^{(l)}\right)^{-1}$
· $D^{(l)} = \Delta^{(l)} - d^{(l)}\Lambda^{(l)}\Delta^{(l)}$
· $C_{\theta\theta} = D^{(l)} + d^{(l)}C_{\beta\beta}d^{(l)\prime}$
· $C_{\theta\beta} = d^{(l)}C_{\beta\beta}$

(4b) Update $\Lambda^{(l+1)}$

· If no constraints, $\Lambda^{(l+1)} = C_{\theta\theta}^{-1}C_{\theta\beta}$
· If λ_k has constraints, use Eq. (8)

(4c) Update $\Omega^{(l+1)} = \text{diag}\left(C_{\beta\beta} - \Lambda^{(l+1)}C_{\theta\beta} - (\Lambda^{(l+1)}C_{\theta\beta})' + \Lambda^{(l+1)}C_{\theta\theta}\Lambda^{(l+1)\prime}\right)$

(4d) Update $\Delta^{(l+1)}$

· If exploratory factor analysis $\Delta^{(l+1)} = I$ (identity matrix)
· If confirmatory factor analysis $\Delta^{(l+1)} = C_{\theta\theta}$

– $\Lambda^{(m+1)} = \Lambda^{(l^*)}$, $\Delta^{(m+1)} = \Delta^{(l^*)}$, and $\Omega^{(m+1)} = \Omega^{(l^*)}$, where l^* is the final iteration of the inner EM algorithm

Note: Fix the random number generator seed in step 1 at each iteration to guarantee convergence of the algorithm.

the case with the algorithm in Table 1. The most costly part of the algorithm is in step (1), taking simulated draws and computing the likelihood. The inner optimization is extremely fast because it only depends on the sample covariance $C_{\beta\beta}$ and does not depend on the sample size n or the number of simulation draws R. From experience, it appears that it is best to choose a stricter stopping criteria for the inner optimization than the one used for the overall algorithm, i.e. $\kappa_2 < \kappa_1$, but allow no more than 500 iterations of the inner algorithm. Capping the number of iterations makes sure that the algorithm does not spend too much time in the initial iterations making wasteful computations and allows the more stricter convergence criteria to take over as the algorithm gets closer to the solution. Although in some settings, performing roughly 10 iterations of the inner loop performed comparably well, this criteria did not have enough robust success to warrant recommendation. It was clear that a single iteration lead to an extremely slow overall time.

Step (4a) requires the inversion of a $K \times K$ matrix. This inversion could be quite expensive for mixed logit models with many product attributes. Rubin and Thayer (1982) suggest applying the Woodbury matrix identity so that this inversion can be carried out with $M \times M$ matrix inversion,

$$\left(\Lambda^{(l)}\Delta^{(l)}\Lambda^{(l)\prime} + \Omega^{(l)}\right)^{-1} = \Omega^{(l)^{-1}} - \Omega^{(l)^{-1}}\Lambda^{(l)}\left(\Delta^{(l)^{-1}} + \Lambda^{(l)\prime}\Omega^{(l)^{-1}}\Lambda^{(l)}\right)^{-1}\Lambda^{(l)\prime}\Omega^{(l)^{-1}}$$

Where Ω is a diagonal matrix so the inverse is simply the inverse of the diagonal elements.

The algorithm in Table 1 can be generalized to a broader set of models than the one outlined in Section 2. For example, since this algorithm is so closely related to the algorithm in Train (2007), extending the algorithm in Table 1 to allow for preferences that are distributed log-normal, censored normal, or from Johnson's S_B distribution can be achieved by making minor adjustments to the choice probabilities in step (1) (see Train (2007) for more details). Furthermore if some attributes have fixed parameters, the algorithm in James (2017) can be easily incorporated into this setting.

Finally, if the researcher is interested in stratifying customers for further analysis, factor scores can be computed from the model. Factor scores, which represent the predicted values of individual latent factors, are useful to study because the factor dimension is much smaller than the dimension of preferences. Factor scores can be calculated by the law of iterated expectations using the conditional factor moments in Eq. (7). Given the maximum likelihood estimates of the model $\hat{\Psi}$, first take

draws of preferences and compute weights according to step 1 of the algorithm. Then, defining $\hat{d} = \hat{\Delta}\hat{\Lambda}'(\hat{\Lambda}\hat{\Delta}\hat{\Lambda}' + \hat{\Omega})^{-1}$ and $\hat{D} = \hat{\Delta} - \hat{d}\hat{\Lambda}\hat{\Delta}$, the individual factor score and precision are given by

$$\begin{aligned}
\mathrm{E}(\theta_i) =& \frac{1}{R}\sum_{r=1}^{R} w_{ir}\,\mathrm{E}(\theta_i|\beta_{ir},\hat{\Psi}) \\
=& \hat{d}\left[\frac{1}{R}\sum_{r=1}^{R} w_{ir}\,(\beta_{ir}-\hat{\gamma})\right] \\
\mathrm{Var}(\theta_i) =& \left[\frac{1}{R}\sum_{r=1}^{R} w_{ir}\,\mathrm{E}(\theta_i\theta_i'|\beta_{ir},\hat{\Psi})\right] - \mathrm{E}(\theta_i)\,\mathrm{E}(\theta_i)' \\
=& \hat{D} + \hat{d}\left[\frac{1}{R}\sum_{r=1}^{R} w_{ir}\,(\beta_{ir}-\hat{\gamma})\,(\beta_{ir}-\hat{\gamma})'\right]\hat{d}' - \mathrm{E}(\theta_i)\,\mathrm{E}(\theta_i)'
\end{aligned}$$

5 Application

This section uses the residential energy customer data studied in Train (2007) and discussed in Section 3 to evaluate the performance of the algorithm and to see how well a factor structure covariance compares to other models. The dataset contains 361 individuals participating in a stated preference survey. The 13 customers that did not respond to all 12 choice situations were dropped, leaving 348 individuals for the analysis.[4] Five normally distributed random coefficients mixed logit models were estimated under different assumptions on the covariance of preferences: (1) uncorrelated, (2) fully correlated, (3) 1-factor exploratory, (4) 2-factor exploratory, and (5) the 2-factor confirmatory model outlined in Section 3.

All models were estimated with the EM algorithm, with the uncorrelated and correlated model using the algorithm in Train (2007) and all of the factor models by way of the algorithm in Table 1. For each model, $R = 6000$ draws were used to evaluate the integral. A large number of draws where used to reduce simulation error so that the log-likelihood values could be compared across models. Each model was considered converged when all of the parameters changed by less than one-tenth of one percent, i.e. $||(\Psi^{(m+1)} - \Psi^{(m)})/\Psi^{(m)}||_\infty < 1\text{e}{-3}$.[5] To get a fair comparison of computation time, each model used similar starting values, which were all computed from the estimates of the

[4] The data for the analysis was take from the `mlogit` package in `R` (Croissant, 2013).

[5] This convergence criteria is stricter than the one used in Train (2007), which was 5e−3.

full covariance model after 5 iterations.[6] For the factor models, the inner tolerance was set where the maximum absolute change in the parameters was less than 1e−7.

Table 2 shows the estimated mean, standard deviation and correlation of preferences for the six attributes across the five models. Both of the 2-factor models produced estimates that were very similar to the fully correlated model, while the 1-factor model yielded estimates that tended to lie between the uncorrelated and the correlated model. The 1-factor model fit the correlation among the pricing attributes very well, for example the correlation of pf and tod was estimated at 0.88 compared to 0.90 for the fully correlated model. However, the 1-factor model overstated the correlation of the pricing components and the supplier components. For example the 1-factor model estimated a correlation between pf and loc of 0.83, while the fully correlated model found that this correlation was only 0.56.

A comparison of the overall performance of each model is presented in Table 3. The different factor models provided a range of total parameters that fell between the extremes of the uncorrelated and fully correlated models. In general, the number of iterations each algorithm required to reach convergence was commensurate with the number of parameters. In terms of computation time, the fully correlated preference model containing the largest number of parameters took, nearly three times as long to estimate as the uncorrelated model. The factor models on the other hand only took 50% longer than the uncorrelated model. These results demonstrate a clear benefit of the factor structure covariance in that they allow for correlation in preferences but can be estimated in significantly less time. The magnitude of the computational savings is increasing in the number of product characteristics as the number of parameters increases exponentially with the fully correlated model, while it only increases linearly with the factor structured model.

Both the uncorrelated and correlated model have full solutions for the EM surrogate function, so we can investigate the computational cost of the repeated numerical optimization of the inner EM algorithm needed for estimating the factor structured covariance model by comparing the time

[6] Let $\gamma^{(0)}$ and $\Sigma^{(0)}$ denote the parameters estimates from the covariance model after 5 iterations. $\gamma^{(0)}$ was used as the starting values for the preference mean in all five models. Only the diagonal elements of $\Sigma^{(0)}$ where used as the starting values for the uncorrelated model. $\Sigma^{(0)}$ was used as the starting value for fully correlated model. For the exploratory factor models, $\Sigma^{(0)}$ was supplied to a built-in factor analysis function to produce starting values for the factor loadings and preference residuals. The starting values for the initial run of five iterations were the estimates from the standard logit.

Table 2: Comparison of Distribution of Preferences

	Uncorrelated	Correlated	1-Factor Exploratory	2-Factor Exploratory	2-Factor Confirmatory
	(1)	(2)	(3)	(4)	(5)
Panel A: Mean of Preferences					
γ_{pf}	-1.000	-1.048	-0.983	-1.052	-1.060
γ_{tod}	-9.595	-10.020	-9.443	-10.074	-10.120
γ_{seas}	-9.743	-10.112	-9.506	-10.160	-10.215
γ_{loc}	2.322	2.641	2.431	2.672	2.658
γ_{wk}	1.660	1.982	1.805	1.993	1.991
γ_{cl}	-0.226	-0.260	-0.250	-0.262	-0.262
Panel B: Standard Deviation of Preferences					
$sd(\beta_{pf})$	0.216	0.823	0.734	0.815	0.829
$sd(\beta_{tod})$	2.404	7.558	6.834	7.527	7.634
$sd(\beta_{seas})$	1.583	7.071	6.063	7.054	7.038
$sd(\beta_{loc})$	1.810	2.267	2.037	2.273	2.259
$sd(\beta_{wk})$	1.179	1.624	1.412	1.626	1.626
$sd(\beta_{cl})$	0.392	0.439	0.427	0.436	0.437
Panel C: Correlation of Preferences					
$corr(\beta_{pf}, \beta_{tod})$	–	0.905	0.892	0.902	0.908
$corr(\beta_{pf}, \beta_{seas})$	–	0.942	0.928	0.940	0.940
$corr(\beta_{pf}, \beta_{loc})$	–	0.544	0.823	0.558	0.531
$corr(\beta_{pf}, \beta_{wk})$	–	0.448	0.770	0.435	0.405
$corr(\beta_{pf}, \beta_{cl})$	–	0.138	0.187	0.086	0.091
$corr(\beta_{tod}, \beta_{seas})$	–	0.923	0.909	0.921	0.922
$corr(\beta_{tod}, \beta_{loc})$	–	0.542	0.805	0.547	0.521
$corr(\beta_{tod}, \beta_{wk})$	–	0.439	0.754	0.426	0.397
$corr(\beta_{tod}, \beta_{cl})$	–	0.111	0.183	0.084	0.089
$corr(\beta_{seas}, \beta_{loc})$	–	0.515	0.838	0.510	0.539
$corr(\beta_{seas}, \beta_{wk})$	–	0.401	0.784	0.390	0.411
$corr(\beta_{seas}, \beta_{cl})$	–	0.081	0.190	0.067	0.092
$corr(\beta_{loc}, \beta_{wk})$	–	0.758	0.695	0.764	0.761
$corr(\beta_{loc}, \beta_{cl})$	–	0.244	0.169	0.247	0.240
$corr(\beta_{wk}, \beta_{cl})$	–	0.145	0.158	0.217	0.183

Table 3: Comparison of Computation Time and Model Fit

	Uncorrelated	Correlated	1-Factor Exploratory	2-Factor Exploratory	2-Factor Confirmatory
	(1)	(2)	(3)	(4)	(5)
No. of Parameters	12	27	18	24	20
No. of Iterations	112	341	150	152	146
Time	272s	825s	364s	377s	364s
Time per Iteration	2.42s	2.42s	2.43s	2.48s	2.50s
Log-likelihood	-3739.8	-3530.6	-3554.5	-3533.3	-3535.8
AIC	7503.5	7115.1	7145.0	7114.5	7111.6
BIC	7579.6	7286.2	7259.1	7266.6	7238.4

Note: Each model used 6,000 pseudo-random draws to approximate the integral. Let p denote the number of parameter. AIC and BIC where computed as AIC $= 2p - 2LL$ and BIC $= \ln(N \times T)p - 2LL$ where $N = 348$, $T = 12$.

per iteration. On average, the factor models took about one-tenth of one second longer per iteration due to the numerical optimization. This implies that only around 5% of the total computation time is spent on the inner optimization. This low cost suggest that a stricter convergence criteria on the inner optimization is likely beneficial to the overall speed of the algorithm.

The last three rows in Table 3 show the log-likelihood at the maximized values as well as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) for the five models. AIC and BIC are used frequently to determine the number of classes in latent class mixed logit models. In a similar vein, AIC and BIC can be applied here for model selection on covariance matrix design. Because the fully correlated model nests the other four models, it is guaranteed to have the highest likelihood value, however, this does not always imply that this is the preferred model.[7] By applying a penalty to the unrestricted model because it contains more parameters, AIC and BIC allow us to determine if the improvements in the log-likelihood in the unrestricted model are the result of an improved understanding of the data or overfitting.[8] In terms of AIC, which applies a smaller penalty than BIC, both the two factor exploratory and confirmatory models

[7]This is only true if a sufficient number of draws are used to simulate the surrogate function in estimation. Otherwise simulation error could generated a simulated log-likelihood of the fully correlated model with a lower value.

[8]Fiebig et al. (2010) provide an interesting simulation to study the ability of AIC and BIC to recover the true preference model. They conclude that in general BIC is more reliable.

are slightly preferred to the fully correlated model.[9] However, in terms of BIC, which is most frequently used to compare latent class models (Bhat, 1997), all three of the factor specifications perform better than the fully correlated model, even the 1-factor model. Despite the fact that the 1-factor model was not able to precisely mimic the estimates from the fully correlated model in Table 2, judging by the model fit, it was able to capture a significant amount of the preference heterogeneity with substantially fewer parameters, thus out performing the fully correlated model as well as the 2-factor exploratory model.

In terms of the five estimated models, the model with the lowest BIC was the 2-factor confirmatory model. This is not surprising given that this model produced heterogeneity in preferences very similar to the fully correlated model in Table 2 and yet had 7 fewer parameters. The reasons the 2-factor confirmatory model performed so well is that we assumed that the factors could be split into a price dimension and a supplier dimension, which turned out to be a fairly accurate characterization of preferences. The results of the model are in Table 4. Panel A shows the covariance matrix of the factors. The price factor and the supplier factor have a strong positive correlation of 0.55, suggesting that consumers that are not very sensitive to price tend to have a stronger preference to their local utility or well-known companies. Panel B shows the factor loadings. Since we did not know which of the factors would be most important to preferences for contract length, these two loadings were estimated. These results indicate that although there is evidence of a positive correlation between price and contract length in Table 2, e.g. those that are less price sensitive tend to prefer longer contracts, this is a spurious correlation. While the price factor has no statistically significant effect on preferences for contract length, the loading on the supplier factor, 0.052, is statistically different from zero. This indicates that individuals that have strong preferences for local utilities also tend to like longer contracts. Thus the correlation of price and contract length occurs because they are both correlated through preference for local suppliers.

Finally, the last three columns in Panel B of Table 4 shows the estimated variance of the preference residual for each attribute and calculates the percent of the total variance of preferences for that attribute that is explained by the factors. Nearly all of the variation in preferences over

[9] Recall that for AIC and BIC, smaller values are preferred since they are functions of the negative log-likelihood.

Table 4: Estimates from 2-Factor Confirmatory Model

Panel A: Factor Covariance

	Factor 1	Factor 2
Factor 1	0.637 (0.123)	0.994 (0.203)
Factor 2	0.994 (0.203)	5.096 (1.164)

Panel B: Factor Loadings and Residual Variance

Attribute	Loadings Factor 1 (λ_{k1})	Loadings Factor 2 (λ_{k2})	Residual Variance (ω_k^2)	Total Variance	Percent of Total Variance Explained by Factors
pf	$1^\dagger$	$0^\dagger$	0.051 (0.010)	0.688	93%
tod	9.030 (0.397)	$0^\dagger$	6.347 (1.047)	58.277	89%
seas	8.617 (0.312)	$0^\dagger$	2.248 (0.730)	49.538	95%
loc	$0^\dagger$	$1^\dagger$	0.007 (0.803)	5.103	100%
wk	$0^\dagger$	0.549 (0.095)	1.109 (0.303)	2.643	58%
cl	−0.030 (0.072)	0.052 (0.025)	0.179 (0.023)	0.191	6%

Note: † indicates normalization. Standard errors in parenthesis computed using the cross-product of individual contributions of the gradient in Eq. (6).

the price attributes are captured by the factors, while only 6% of the preference heterogeneity over the contract length attribute is explained by the factors.

6 Conclusion

Current techniques for estimating mixed logit models with normally distributed random coefficients are only suitable under the two extreme assumptions of either independent preferences or fully correlated preferences, which leaves a vast space of alternative models for preference heterogeneity inaccessible to researchers. This paper develops an EM algorithm for estimating factor structured covariance mixed logit models. Factor structured covariances provide a more parsimonious repre-

sentation of correlation in preferences, resulting in fewer parameters and much faster computation times compared to models with a full covariance. An analysis of stated preference data for a six attribute choice model showed that a 1-factor covariance model outperformed the full covariance model in terms of BIC and took half the amount of time to estimate. These benefits would likely become even more substantial in choice models with 15 or 20 attributes given that the number of parameters in a full covariance model increases exponentially in the number of attributes while in the factor model the number of parameters only increases linearly in the number of attributes.

Finally, in addition to using factor models to reduce the number of parameters, confirmatory factor models allow the researcher to bring outside information to the model to investigate the underlying relationship among preference and directly test hypothesis about the sources of correlation. In an analysis of stated preference data from residential energy customers, a confirmatory factor model was estimated which stratified the factors in a similar dimension to the product attributes, i.e. pricing attributes and supplier attributes. This 2-factor confirmatory model greatly outperformed the exploratory factor models and the full covariance model in terms of BIC.

References

C. R. Bhat. An endogenous segmentation mode choice model with an application to intercity travel. *Transportation science*, 31(1):34–48, 1997.

Y. Croissant. *mlogit: multinomial logit model*, 2013. URL `https://CRAN.R-project.org/package=mlogit`. R package version 0.2-4.

T. Elrod and M. P. Keane. A factor-analytic probit model for representing the market structure in panel data. *Journal of Marketing Research*, pages 1–16, 1995.

D. G. Fiebig, M. P. Keane, J. Louviere, and N. Wasi. The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. *Marketing Science*, 29(3):393–421, 2010.

J. Huber and K. Train. On the similarity of classical and bayesian estimates of individual mean partworths. *Marketing Letters*, 12(3):259–269, 2001.

J. James. MM algorithm for general mixed multinomial logit models. *Journal of Applied Econometrics*, 32(4):841–857, 2017.

M. Keane and N. Wasi. Comparing alternative models of heterogeneity in consumer choice behavior. *Journal of Applied Econometrics*, 28(6):1018–1045, 2013.

M. P. Keane and N. Wasi. Estimation of discrete choice models with many alternatives using random subsets of the full choice set: With an application to demand for frozen pizza. 2012.

D. Lawley. The estimation of factor loadings by the method of maximum likelihood. proc. roy. soc. edinb. abo. 64-82. 1940.

D. Revelt and K. Train. Customer-specific taste parameters and mixed logit. *University of California, Berkeley*, 1999.

D. B. Rubin and D. T. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, 1982.

P. A. Ruud. Extensions of estimation methods using the em algorithm. *Journal of Econometrics*, 49(3):305–341, 1991.

K. Train. A recursive estimator for random coefficient models. *University of California, Berkeley*, 2007.

K. E. Train. EM algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*, 1(1):40–69, 2008.