Bayesian non-parametric inference with monotonicity restrictions for discrete choice experiments

Remi Daviet [*] and William McCausland[†]

May 15, 2017

Abstract

Insert abstract here

Keywords: Big Data, Discrete Choice experiments

[*]Corresponding author. Department of economics, University of Toronto; remi.daviet@mail.utoronto.ca
[†]Department of economics, Université de Montréal

1 Products attributes and products classes

Every product considered is represented by a point in the chosen attributes space. For instance, a mobile phone can be represented by a set of two attributes, the screen size and the color. Each attribute has a level. In our example, the screen size can have 3 levels ($Small, Medium$ or $Large$) and the color can have 4 levels ($Black, White, Blue$ or $Other$). Considering this attributes space, two possible phones are phone 1 ($Large, Blue$) and phone 2 ($Medium, White$).

Products can be regrouped into classes. Classes are set of attributes commons to every product belonging to the class. A class can be very simple, for instance the class ($\{Small\}, any$) which regroups all the phones with a small screen size, no matter the color. Classes can also be more complicated such as ($\{Small, Large\}, \{White, Other\}$) which regroups phones of small or large screen size having a white or "other" color. Note that the ($\{Small\}, any$) class can be written ($\{Small\}, \{Black, White, Blue, Other\}$) and that the keyword "any" is just used for convenience. Finally, a class can specify a unique product in the attribute space, such as the class ($\{Large\}, \{Blue\}$).

More formally, considering a product space with L attributes, a product x_m is represented by the vector of attributes $(x_{m1}, x_{m2}, \ldots, x_{mL})$. Each attribute i can have L_i possible levels. For instance, if the attribute (1) "screen size" has 3 possible levels ($Small, Medium$ or $Large$), we write $L_1 = 3$. For the second attribute "Color" which has 4 possible levels, we write $L_2 = 4$. The attributes levels attributed to a product x_m are indicated by their corresponding number. For instance, if $Small$ and $Black$ are the first levels of their respective attributes, the phone will be represented by the vector $(x_{m1} = 1, x_{m2} = 1)$.

A class of product Λ_k is represented by a Cartesian product of L sets $(\Lambda_{k1}, \Lambda_{k2}, \ldots, \Lambda_{kL})$. Each Λ_{ki} represents the set of possible levels for the attribute i for a product in the class k. The levels belonging to Λ_{ki} are called active levels. For instance, a class regrouping $Black$ phones with either $Medium$ or $Large$ screen size will be represented by $(\Lambda_{k1} = \{2, 3\}, \Lambda_{k2} = \{1\})$. When all the levels of an attribute are active, it means that this particular attribute does not matter when defining the class, and the attribute is defined as inactive. On the contrary, when only some levels are active, it means that the attribute is used as a criterion to select if a product is a member of the class, and the attribute is defined as active.

A product x_m belongs to a class Λ_k if and only if $x_{m1} \in \Lambda_{k1}, x_{m2} \in \Lambda_{k2}, \ldots$ and $x_{mL} \in \Lambda_{kL}$. For convenience, we will use the notation $x_m \in \Lambda_k$ to indicate that the product x_m is a member of the class Λ_k.

2 Individual preferences

We have a population of N individuals sequentially choosing T times one product within a set. The choice of an individual n at a time t is done by maximizing a function $u_t(x, \psi_n)$ representing the preferences this individual over the set of available choices X_{nt}, where ψ_n is some set of parameters. The choice maximizing the individual preferences is denoted x^*. The preferences for each individual are built by attributing some value to various classes of products. The value of a class can be positive or negative. For instance, at a time t an individual n attributes the value θ_{nkt} to the class Λ_k. The value of a given product for an individual is obtained by summing the values of all the classes it belongs to. If the value of a class k is zero for an individual, we say that the individual does not possess the feature k. It means

that whether a product belongs to the class k or not does not change the value of the product for the given individual.

Within the population, when two individuals have a non-zero value θ_{nkt} for a class of products k, we say that they share the common feature k. It is important to note however that the value of the class does not need to be the same for both individuals sharing the feature. One individual with feature k may attribute a positive value to the class k while the other attributes a negative value. Two individuals sharing a feature k means that they are both sensitive to the corresponding class of products. Some features can be shared by many individuals, while some other features can be rare. Note that there is some redundancy: a same class can appear in two different features. The two classes of products Λ_k and $\Lambda_{k'}$ can be identical ($\Lambda_k = \Lambda_{k'}$), but to be respectively associated with a θ_{nk} and a $\theta_{nk'}$ following different distributions.

The total number of features in the population is potentially infinite. However, in a given sample only a finite number of features are realized. The features present in our sample are ordered and numbered from 1 to K_+ for convenience.

Formally, the value of an object x for and individual n at time t can be written:

$$u_t(x, \psi_n) = \sum_{k=1}^{K_+} \theta_{nkt} \cdot \mathbb{1}(x \in \Lambda_k), \tag{2.1}$$

where ψ_n contains the set of θ_{nkt} and the set of Λ_k:

$$\psi_n = \left\{ \{\theta_{nkt}\}_{t=1}^{T}, \Lambda_k \right\}_{k=1}^{K_+}$$

For convenience, the value θ_{nk} may be decomposed into four parts as follow:

$$\theta_{nkt} = z_{nk} \cdot (-1)^{s_{nk}} \cdot (\theta_{nk} + \epsilon_{nkt})$$

where z_{nk} is a binary variable indicating whether individual n possesses feature k, s_{nk} is a binary variable that indicates whether the value of the class k for individual n is positive or negative, θ_{nk} is the stable part of the value in the consumer preferences, and ϵ_{nkt} is a tremble part. We can rewrite ψ_n as:

$$\psi_n = \left\{ z_{nk}, s_{nk}, \theta_{nk}, \{\epsilon_{nkt}\}_{t=1}^{T}, \Lambda_k \right\}_{k=1}^{K_+}.$$

It is possible that we want to restrict preferences such that they only increase in some attributes. For such attribute, the active levels are restricted to be of the form $\{l, l+1, \ldots, L\}$ if the the class Λ_k is associated with a positive value ($s_{nk} = 0$), and of the form $\{1, \ldots, l-1, l\}$ if the class is associated with a negative value ($s_{nk} = 1$). We denote the first case as a upper active levels, and the second as lower active levels. If a class Λ_k has at least one active monotone attribute, we call it a monotone class and denote it with the binary variable $m_k = 1$. Note that if several monotone attributes are active in the class, they should all be of the same type. If all the active monotone attribute are of the upper active levels type, we say that the class has a positive polarity and denote it $\rho_k = 0$. conversely, if the monotone attributes are of the lower active levels type, we say that the class has a negative polarity and denote it $\rho_k = 1$.

??????????????????????QUESTION??????????????????

Shall I write $u_t(x, \psi_n)$ with $\psi_n = \left\{ z_{nk}, s_{nk}, \theta_{nk}, \{\epsilon_{nkt}\}_{t=1}^{T}, \Lambda_k \right\}_{k=1}^{K_+}$,

Or $u(x, \psi_{nt})$, with $\psi_{nt} = \{z_{nk}, s_{nk}, \theta_{nk}, \epsilon_{nkt}, \Lambda_k\}_{k=1}^{K_+}$?

in the second case, the various ψ_{nt} have common parameters indexed by nk, which may look weird. I used the first approach where the index t is on u_t because the t is relevant both for the component ϵ_{nkt} of ψ_n to retain, and for the choice set / observation considered.

3 Hierarchichal model

The Bayesian framework gives us a simple and intuitive way to model our uncertainty about the various parameters. We have to define how the various variables and parameters are distributed in order to be able to obtain a posterior distribution.

Model parameters:

$$\{\{\psi_n\}_{n=1}^N, \{\Lambda_k\}_{k=1}^{K_+}\} = \left\{\left\{\{\epsilon_{nkt}\}_{t=1}^T, z_{nk}, s_{nk}, \theta_{nk}\right\}_{n=1}^N, \Lambda_k\right\}_{k=1}^{K_+}$$

We add a hierarchical level requiring the addition of the following hyper-parameters:

$$\left\{\{\pi_k^s\}_{k=1}^{K_+}, \alpha^\theta, \sigma^\theta\right\}$$

The joint distribution of the data and parameters satisfies the conditional independence relationships implied by the following decomposition:

$$P\left(\{\{x_{nt}^*\}_{t=1}^T\}_{n=1}^N, \left\{\left\{\{\epsilon_{nkt}\}_{t=1}^T, z_{nk}, s_{nk}, \theta_{nk}\right\}_{n=1}^N, \Lambda_k, \pi_k^s\right\}_{k=1}^{K_+}, \alpha^\theta, \sigma^\theta\right) \tag{3.2}$$

$$= \prod_t \prod_n P\left(x_{nt}^* | \{\epsilon_{nkt}, z_{nk}, s_{nk}, \theta_{nk}, \Lambda_k\}_{k=1}^{K_+}\right) \tag{3.3}$$

$$\times P(Z) \prod_k \left[\prod_n \left(\prod_t P(\epsilon_{nkt}|\alpha^\theta, \sigma^\theta, z_{nk})\right) P(s_{nk}|\pi_k^s, z_{nk}) P(\theta_{nk}|\alpha^\theta, \sigma^\theta, z_{nk}) P(\Lambda_k)\right] \tag{3.4}$$

$$\times \left(\prod_k P(\pi_k^z) P(\pi_k^s)\right) P(\alpha^\theta) P(\sigma^\theta), \tag{3.5}$$

with $Z = \{z_{nk}\}$ being the binary matrix of features allocation where each row represents an individual and each column represents a feature. In this equation, line (3.3) is the likelihood, line (3.4) is the prior and line (3.5) is the hyper-prior. The various distributions are described in the following sections.

3.1 Features allocation

Our first object of interest is the binary matrix of features allocation $Z = \{z_{nk}\}$. While the Z matrix has an infinite number of columns, we are only interested in the features (columns) that at least one person in our sample possesses. We use the two-parameter infinite feature model defined by Griffiths and Ghahramani (2011) in their paper about the Indian Buffet Process (IBP). The IBP approach presents

a convenient way to perform inference by simulation for models using latent features. We use the two-parameter exchangeable Indian Buffet Process as a prior for our binary Matrix Z:

$$Z \sim eIBP(\alpha^z, \beta^z)$$

Using this process, the average number of feature per individual will be α^z. The β^z parameter lets us set the average overall number of features observed in a sample. The expected overall number of features present in a sample will be $\alpha^z \sum_{n=1}^{N} \frac{\beta^z}{\beta^z+n-1}$. Depending on β^z, this number can range from α^z where everybody shares the same features, to $N\alpha^z$ where no features are shared.

3.2 Attributes with monotone preferences and Λ_k

The class Λ_k can be denoted as a set of active attributes C_k with cardinality c_k, and for each active attribute $i \in C_k$, a set of active levels Λ_{ki} with cardinality λ_{ki}.

The prior distribution of Λ_k can be decomposed as follow:

$$P(\Lambda_k) = P(c_k) \cdot P(\rho_k) \cdot P(C_k|c_k) \cdot \prod_{i \in C_k} P(\Lambda_{ki}|\rho_k) \tag{3.6}$$

The corresponding distributions are detailed below:

- The number of active attributes is uniformly distributed on $\{1, ..., L\}$:

$$c_k \sim U(1, ..., L)$$

- The polarity has equal probability of being positive or negative:

$$\rho_k \sim Bernoulli(0.5)$$

- The attributes share the same probability of being active. The probability of observing a set of active attributes C_k is consequently:

$$P(C_k|c_k) = \binom{L}{c_k}^{-1}$$

- The probability of observing the set of active levels Λ_{ki} depends on several factors:
  - If Λ_{ki} is monotone ($m_k = 1$):

$$P(\Lambda_{ki}|\rho_k) = \begin{cases} 1/L_i, & \text{if } \Lambda_{ki} \text{ has upper active levels and } \rho_k = 0 \\ 1/L_i, & \text{if } \Lambda_{ki} \text{ has lower active levels and } \rho_k = 1 \\ 0, & \text{otherwise} \end{cases}$$

– If Λ_{ki} is not monotone ($m_k = 0$), it does not depend on the polarity ρ_k. The number of active levels λ_{ki} is uniformly distributed on $\{1, ..., L_i\}$ and each level has an equal probability of being active.

$$\begin{aligned} P(\Lambda_{ki}|\rho_k) &= P(\Lambda_{ki}|\lambda_{ki}) \cdot P(\lambda_{ki}) \\ &= \binom{L_i}{\lambda_{ki}}^{-1} \cdot \frac{1}{\lambda_{ki}} \end{aligned}$$

3.3 Other parameters

The other parameters $(s_{nk}, \theta_{nk}, \epsilon_{nkt})$ are defined conditional on z_{nk}. When $z_{nk} = 0$, the values of all these parameters are arbitrarily set to 0. We have to define the prior distributions of $(s_{nk}, \theta_{nk}, \epsilon_{nkt})$ conditional on $z_{nk} = 1$:

$$s_{nk}|\pi^s_k, z_{ik} = 1 \sim \begin{cases} 0 & if\ m_k = 1 \text{ and } \rho_k = 0 \\ 1 & if\ m_k = 1 \text{ and } \rho_k = 1 \\ Bernoulli(\pi^s_k) & \text{otherwise} \end{cases}$$

$$\theta_{nk}|\sigma^\theta, \alpha_\theta, z_{ik} = 1 \sim Gamma(\sigma^\theta \alpha^\theta, 1)$$

$$\epsilon_{nkt}|\sigma^\theta, \alpha^\theta, z_{ik} = 1 \sim Gamma((1 - \sigma^\theta)\alpha^\theta, 1)$$

where (s_{nk}, θ_{nk}) are independent across n and k; and ϵ_{nkt} are independent across n,k and t.

Note that the distribution of the value size $(\theta_{nk} + \epsilon_{nkt})$ is also Gamma with parameters $Gamma(\alpha^\theta, 1)$ from the properties of a sum of Gamma distributed variables. The parameter α^θ is the shape parameter of the distribution of the value size, while σ^θ can be interpreted as a stability parameter. For σ^θ close to one, the stable component θ_{nk} is favored and the utility functions have high serial dependence; for σ^θ close to zero, the variable component ϵ_{nkt} is favored and the utility function has low serial dependence.

Finally, we have to define hyperpriors on some of the hyperparameters:

$$\alpha^\theta \sim Gamma(1, 1)$$

$$\sigma^\theta \sim Gamma(1, 1)$$

$$\pi^s_k \sim Beta(1, 1)$$

4 Inference

4.1 Likelihood and Posetrior

Conditional on the set of all parameters, denoted ψ, there is no random term in the utility function. The likelihood becomes a degenerate function that is equal to 1 if all the observed choices are predicted correctly and 0 otherwise. The probability of observing the set of choices $X^* = \{x^*_{nt}\}_{n,t}$ is:

$$P(X^*|\psi) = \mathbb{1}\{\arg\max_{x \in X_{nt}} u(x, \psi) = x^*_{nt}, \forall n, t\}$$

Likelihood-based inference consists in recovering the identified set of parameters such that:

$$P(X^*|\psi) = 1$$

We can see that with the current likelihood function, if at least one choice is not maximizing utility according to the current value of ψ, the likelihood takes the value of 0.

In the Bayesian framework, the target distribution is the corresponding posterior:

$$P(\psi|X^*) \propto P(X^*|\psi) \cdot P(\psi),$$

where $P(\psi)$ denotes the prior density of our parameters.

4.2 Tempering and Sequence of Distributions

Simulating the posterior can be quite difficult, and we simulate instead a sequence of target distributions converging to this Posterior. The functions used in the sequence are increasingly difficult to simulate. This process is known as tempering, and we achieve it by combining two methods. First, we use Data Tempering and introduce the observations one by one in the likelihood. We start with the first observation of the first individual $x^*_{1,1}$, then add the subsequent observations for this individual until all the observations are included in the likelihood. We then add the observations of the following individual, and continue until every observation of every individual are included. For the second tempering approach, we use an instrumental function that does not take a value of zero if the last observed choice is not utility maximizing with the current value of ψ. We call this approach ζ-tempering and substitute the likelihood with the following function:

$$Q(X^*|\psi;\eta,\tau,\zeta) = P(\{\{x^*_{nt}\}_{t=1}^{T}\}_{n=1}^{\eta-1}|\psi) \cdot P(\{x^*_{\eta t}\}_{t=1}^{\tau-1}|\psi) \cdot \zeta^{u(x^*_{\eta\tau},\psi)-\max_{x\in X_{\eta\tau}} u(x,\psi)},$$

where $\zeta > 1$ is a constant chosen by the researcher. We can see that this instrumental function is the product of 3 terms. The first one is the likelihood of observing the choices of individuals 1 to $\eta - 1$. The second is the likelihood of observing the $\tau - 1$ first choices of individual η. The last term requires further attention. If the last observed choice is utility maximizing, this term is equal to 1 as for the likelihood. However, it is equal to a value between 0 and 1 if another choice would be utility maximizing. The smaller the difference between the observed choice and the utility maximizing one, the closer the term gets to 1. It is important to note that $Q(X^*|\psi;\eta,\tau,\zeta) \to P(\{x^*_{\eta t}\}_{t=1}^{\tau} \cup \{\{x^*_{nt}\}_{t=1}^{T}\}_{n=1}^{\eta-1}|\psi)$ as $\zeta \to \infty$.

The corresponding quasi-posterior distribution is defined as

$$Q(\psi|X^*;\eta,\tau,\zeta) \propto Q(X^*|\psi;\eta,\tau,\zeta) \cdot P(\psi).$$

Using these tempering approaches, we can define a sequence of distributions by specifying the values (η,τ,ζ) for each distribution in the sequence. We start by the distribution

$$Q(\psi|X^*;\eta=1,\tau=1,\zeta=2) \propto 2^{u(x^*_{1,1},\psi)-\max_{x\in X_{1,1}} u(x,\psi)} \cdot P(\psi)$$

and progressively increase ζ by multiplying it by 2 in each subsequent distribution until $\zeta \to \infty$ and the target distribution becomes

$$Q(\psi|X^*;\eta=1,\tau=1,\zeta\to\infty) \propto P(\psi|\{x^*_{1,1}\}).$$

The initial distributions in our sequence follow the following pattern for the values of (η, τ, ζ):

$$(1,1,2),(1,1,4),(1,1,8),...,(1,1,\infty)$$

We then increase the value of τ by 1 and progressively increase ζ from $(1,2,2)$ to $(1,2,\infty)$. We repeat this step with until we reach the last observation for the current individual $(1,T,\infty)$. We then increase η by one and start again alternatively increasing the value of τ by 1 and progressively increasing ζ from $(2,1,2)$ to $(2,1,\infty)$. We repeat this process until we reach the last observation of the last individual (N,T,∞), where $Q(\psi|X^*;\eta=N,\tau=T,\zeta=\infty)$ is equal to the posterior using the full likelihood.

Note that the ratio of two consecutive target distributions when increasing ζ by a factor of a is equal to

$$\frac{P(\{\{x^*_{nt}\}_{t=1}^{T}\}_{n=1}^{\eta-1}|\psi)\cdot P(\{x^*_{\eta t}\}_{t=1}^{\tau-1}|\psi)\cdot(a\zeta)^{u(x^*_{\eta\tau},\psi)-\max_{x\in X_{\eta\tau}}u(x,\psi)}}{P(\{\{x^*_{nt}\}_{t=1}^{T}\}_{n=1}^{\eta-1}|\psi)\cdot P(\{x^*_{\eta t}\}_{t=1}^{\tau-1}|\psi)\cdot\zeta^{u(x^*_{\eta\tau},\psi)-\max_{x\in X_{\eta\tau}}u(x,\psi)}}=a^{u(x^*_{\eta\tau},\psi)-\max_{x\in X_{\eta\tau}}u(x,\psi)}.$$

The ratio of two consecutive target distributions when increasing τ by 1 and reseting ζ to 2 is equal to:

$$\frac{P(\{\{x^*_{nt}\}_{t=1}^{T}\}_{n=1}^{\eta-1}|\psi)\cdot P(\{x^*_{\eta t}\}_{t=1}^{\tau-1}|\psi)\cdot 2^{u(x^*_{\eta\tau},\psi)-\max_{x\in X_{\eta\tau}}u(x,\psi)}}{P(\{\{x^*_{nt}\}_{t=1}^{T}\}_{n=1}^{\eta-1}|\psi)\cdot P(\{x^*_{\eta t}\}_{t=1}^{\tau-1}|\psi)}=2^{u(x^*_{\eta\tau},\psi)-\max_{x\in X_{\eta\tau}}u(x,\psi)}$$

4.3 Quasi-Posterior Simulation

We propose the adaptive Sequential Monte Carlo (SMC) approach (Durham and Geweke, 2014) as a posterior simulation strategy, for two main reasons. First, many utility functions of the form in (2.1) give identical predictions. We cannot expect these functions to be close, in the sense of a practical MCMC chain passing from one to the other in a small number of steps with reasonable probability. Having many particles exploring the parameter space solves that problem. Second, a sequential inference method is required for our tempering approach.

4.3.1 Algorithm

The algorithm uses a population of J groups of P particles, each of them containing a set of realizations of all our simulated parameters:

$$\psi^{(jp)}=\left\{\left\{\{\epsilon^{(jp)}_{nkt}\}_{t=1}^{T},z^{(jp)}_{nk},s^{(jp)}_{nk},\theta^{(jp)}_{nk}\}_{n=1}^{N},\Lambda^{(jp)}_{k}\right\}_{k=1}^{K_+},\alpha^{\theta(jp)},\sigma^{\theta(jp)}\right\}.$$

Note that π^s_k is not simulated since we can use conjugacy rules between the Binomial and the Beta distributions.

The algorithm starts with an Initialization (I) phase and then iterates over Correction (C), Selection (S) and Mutation (M) phases until the last distribution in the sequence is reached.

The I phase draws the $J\times P$ particles from the prior.

The C phase assigns a weight to particles following an importance sampling approach. The particles are distributed proportionally to an original quasi-posterior $Q(\psi|X^*;\eta_o,\tau_o,\zeta_o)$, and we want to approximate

the distribution coming further in the sequence $Q(\psi|X^*; \eta_f, \tau_f, \zeta_f)$. In this context, importance sampling relies on weighting each particles according to the following function:

$$w^{(j,p)} = \frac{Q(\psi|X^*; \eta_f, \tau_f, \zeta_f)}{Q(\psi|X^*; \eta_o, \tau_o, \zeta_o)}.$$

The transition from the C phase to the S phase is done adaptively as a function of particles' Relative Sample Size (RSS). The RSS is computed by dividing the Effective Sample Size (ESS) by the total number of particles:

$$RSS(\{w^{(j,p)}\}) = \frac{[\sum_j \sum_p w^{(j,p)}]^2}{J \cdot P \cdot \sum_j \sum_p w^{(j,p)^2}}$$

We want to find the position $(\eta_f, \tau_f, \zeta_f)$ in the sequence such that the RSS is below 0.5, but above 0.2. We propose here a strategy to find $(\eta_f, \tau_f, \zeta_f)$ satisfying this condition without having to compute the weights for every distribution in the sequence.

Starting with the distribution at $(\eta_o, \tau_o, \zeta_o)$, we compute the weights and RSS for a subsequent distribution at (η_s, τ_s, ∞). If the RSS is above 0.5 we directly move to the next distribution having $\zeta = \infty$. In other words, we add an observation x_{nt} while maintaining $\zeta = \infty$. We repeat this process by adding observations until the RSS falls below 0.5. At this point we have reached the distribution (η_f, τ_f, ∞). If the RSS is above 0.2, we directly move to the S phase. If the RSS is below 0.2 we compute the weights and RSS for each distribution in the sequence starting at $(\eta_f, \tau_f, 2)$. Once we find a value ζ^* for which the RSS is below 0.5, we set $\zeta_f = \zeta^*$ if $RSS > 0.2$, and $\zeta_f = \zeta^*/2$ otherwise. Note that in the special case where $\zeta_f = 1$, the distribution $(\eta_f, \tau_f, 1)$ is in fact the same as the distribution preceding $(\eta_f, \tau_f, 2)$ where $\zeta = \infty$ (with one less observation).

The S phase resamples P new particles within each group j, using a multinomial resampling scheme with weights proportional to $\{w^{(j,p)}\}$. The weights need to be normalized within each group j such that:

$$\sum_p w^{(j,p)} = 1$$

???????????????? Note: should we use residual/stratified/systematic resampling instead ?????????????

The M phase applies a series of Markov Chain Monte-Carlo (MCMC) steps to each particle such that the current distribution $Q(\psi|X^*; \eta_f, \tau_f, \zeta_f)$ is the stationary distribution of the chain. The M step is detailed in the following section.

4.3.2 M phase

In this phase, 20 MCMC steps are applied to each particles. Each step consists in an application of the Metropolis within Gibbs approach. The parameter space being of high dimension, we partition the parameter dimensions into subspaces called Gibbs blocks. We then apply a MCMC step withing each block such that the target distribution $Q(\psi|X^*; \eta_f, \tau_f, \zeta_f)$ is preserved. In this section, when using the Metropolis-Hastings approach, we denote a proposal by a *. The step being done within each particle independently, the index notation $^{(jp)}$ is not needed and will be dropped for clarity.

The list of blocks and the MCMC method used is detailed in the following table.

Table 1: List of Gibbs blocks

Subspace of	Number of blocks	MCMC strategy
$z_{nk},\ (k \leq K_+)$	$\eta \times K_+$	Gibbs
$\{\epsilon_{nkt}\}_{t=1}^{T}, s_{nk}, \theta_{nk},\ (k \leq K_+)$	$\eta \times K_+$	Metropolis within Gibbs
$\left\{\{\epsilon_{nkt}\}_{t=1}^{T}, s_{nk}, \theta_{nk}, z_{nk}\right\}_{K_+}^{\infty}$	η	Metropolis within Gibbs [1]
$\alpha^{\theta}, \sigma^{\theta}$	1	Metropolis within Gibbs

We suggest to proceed in the following order. We first resample alternatively z_{nk} and $\left(\{\epsilon_{nkt}\}_{t=1}^{T}, s_{nk}, \theta_{nk}\right)$ starting at $(n = 1, k = 1)$ and finishing at $(n = 1, k = K_+)$. We then sample jointly additional z_{nk} and $\left(\{\epsilon_{nkt}\}_{t=1}^{T}, s_{nk}, \theta_{nk}\right)$ that would create new non-zero columns in Z, raising K_+. We then move to the next individual and repeat the process until $(n = \eta, k = K_+)$. We finally resample $\left(\alpha^{\theta}, \sigma^{\theta}\right)$.

Note that the space of Λ_k is explored through two machanisms. A class Λ_k might be deleted during the move in the subspace of $z_{nk},\ (k \leq K_+)$ if no individual possesses the corresponding feature. New classes Λ_k might be created during the move in the subspace of $\left\{\{\epsilon_{nkt}\}_{t=1}^{T}, s_{nk}, \theta_{nk}, z_{nk}\right\}_{K_+}^{\infty}$ if a new non-zero column is created in Z. More details are given in the following sections.

1. **Subspace of** $z_{nk},\ (k \leq K_+)$

   Our goal is to sample from the following conditional distribution:

$$
\begin{aligned}
& Q\left(z_{nk} | X^*; \left\{\{\epsilon_{nkt}\}_{t=1}^{T_n}, s_{nk}, \theta_{nk}, \Lambda_k\right\}_{k=1}^{K_+}, \{z_{n'k}\}_{n' \neq n, n' \leq \eta}; \eta, \tau, \zeta\right) \\
\propto\ & Q(X^* | \psi; \eta, \tau, \zeta) P\left(z_{nk} | \{z_{n'k}\}_{n' \neq n, n' \leq \eta}\right) \\
\propto\ & Q\left(\{x_{nt}^*\}_{t=1}^{T_n} | \left\{\{\epsilon_{nkt}\}_{t=1}^{T_n}, s_{nk}, \theta_{nk}, z_{nk}, \Lambda_k\right\}_{k=1}^{K_+}; \eta, \tau, \zeta\right) P\left(z_{nk} | \{z_{n'k}\}_{n' \neq n, n' \leq \eta}\right),
\end{aligned}
$$

   where $T_n = \tau$ if $n = \eta$, and $T_n = T$ otherwise. The distribution $P\left(z_{nk} | \{z_{n'k}\}_{n' \neq n, n' \leq \eta}\right)$ has a simple closed form (Griffiths and Ghahramani, 2011):

$$
P\left(z_{nk} = 1 | \{z_{n'k}\}_{n' \neq n, n' \leq \eta}\right) = \frac{\sum_{n' \neq n, n' \leq \eta} z_{n'k}}{\beta^z + \eta - 1}.
$$

   Note that if $\sum_{n' \neq n, n' \leq \eta} z_{n'k} = 0$, the class k is simply deleted and the corresponding $\{\epsilon_{nkt}\}_{t=1}^{T}, s_{nk}, \theta_{nk}$ are set to 0.

[1] K_+ may change during the move as new non-zero columns in $Z^{(jp)}$ appear. If new non-zero columns are created, new $\Lambda_k^{(jp)}$ will also be drawn.

We do not need an accept-reject strategy and can resample directly from our target distribution by sampling from a $Bernoulli(\frac{A_1}{A_0+A_1})$ where:

$$A_0 = Q\left(\{x^*_{nt}\}_{t=1}^{T_n} | \left\{\{\epsilon_{nkt}\}_{t=1}^{T_n}, s_{nk}, \theta_{nk}, \Lambda_k\right\}_{k=1}^{K_+}, \{z_{n'k}\}_{n'\neq n, n'\leq\eta}, z_{nk}=0; \eta, \tau, \zeta\right) P\left(z_{nk}=0 | \{z_{n'k}\}_{n'\neq n, n'\leq\eta}\right)$$

$$A_1 = Q\left(\{x^*_{nt}\}_{t=1}^{T_n} | \left\{\{\epsilon_{nkt}\}_{t=1}^{T_n}, s_{nk}, \theta_{nk}, \Lambda_k\right\}_{k=1}^{K_+}, \{z_{n'k}\}_{n'\neq n, n'\leq\eta}, z_{nk}=1; \eta, \tau, \zeta\right) P\left(z_{nk}=1 | \{z_{n'k}\}_{n'\neq n, n'\leq\eta}\right)$$

2. **Subspace of** $\{\epsilon_{nkt}\}_{t=1}^{T}, s_{nk}, \theta_{nk},\ (k \leq K_+)$

Our goal is to sample from the following conditional distribution:

$$\begin{aligned}
&Q\left(\{\epsilon_{nkt}\}_{t=1}^{T}, s_{nk}, \theta_{nk} | X^*; \alpha^\theta, \sigma^\theta, \{s_{n'k}\}_{n'\neq n, n'\leq\eta}, \{z_{n'k}\}_{n'=1}^{\eta}, \Lambda_k; \eta, \tau, \zeta\right) \\
\propto\ & Q(X^*|\psi; \eta, \tau, \zeta) P\left(\{\epsilon_{nkt}\}_{t=1}^{T}, s_{nk}, \theta_{nk} | \alpha^\theta, \sigma^\theta, \{s_{n'k}\}_{n'\neq n, n'\leq\eta}, z_{nk}, \Lambda_k\right) \\
\propto\ & Q\left(\{x^*_{nt}\}_{t=1}^{T_n} | \left\{\{\epsilon_{nkt}\}_{t=1}^{T_n}, s_{nk}, \theta_{nk}, z_{nk}, \Lambda_k\right\}_{k=1}^{K_+}; \eta, \tau, \zeta\right) \\
&\times \prod_{t}^{T_n} P(\epsilon_{nkt} | \alpha^\theta, \sigma^\theta, z_{nk}) \\
&\times P(\theta_{nk} | \alpha^\theta, \sigma^\theta, z_{nk}) \\
&\times P(s_{nk} | \{s_{n'k}\}_{n'\neq n, n'\leq\eta}, \{z_{n'k}\}_{n'=1}^{\eta}, \Lambda_k).
\end{aligned}$$

Note that we can use conjugacy rules to derive $P(s_{nk}|\{s_{n'k}\}_{n'\neq n, n'\leq\eta}, \{z_{n'k}\}_{n'=1}^{\eta}, \Lambda_k)$:

$$s_{nk} | \{s_{n'k}, z_{n'k}\}_{n'\neq n, n'\leq\eta}, z_{nk}=1, \Lambda_k \sim \begin{cases} 0 & if\ m_k = 1 \text{ and } \rho_k = 1 \\ 1 & if\ m_k = 1 \text{ and } \rho_k = 0 \\ Bernoulli\left(\frac{1+\sum_{n'\neq n, n'\leq\eta}(s_{n'k})}{2+\sum_{n'\neq n, n'\leq\eta}(z_{n'k})}\right) & \text{otherwise} \end{cases}$$

We must consider two cases here. In the first case, $z_{nk} = 0$ and $\{\epsilon_{nkt}\}_{t=1}^{T}, s_{nk}, \theta_{nk}$ are all left to 0 with probability 1. In the second case, where $z_{nk} = 1$, we propose a Metropolis within Gibbs approach where the proposal distribution is:

$$\begin{aligned}
&P\left(\{\epsilon_{nkt}\}_{t=1}^{T}, s_{nk}, \theta_{nk} | \alpha^\theta, \sigma^\theta, \{s_{n'k}\}_{n'\neq n, n'\leq\eta}, z_{nk}, \Lambda_k\right) \\
=&\prod_{t}^{T_n} P(\epsilon_{nkt} | \alpha^\theta, \sigma^\theta, z_{nk}) \\
&\times P(\theta_{nk} | \alpha^\theta, \sigma^\theta, z_{nk}) \\
&\times P(s_{nk} | \{s_{n'k}\}_{n'\neq n, n'\leq\eta}, \{z_{n'k}\}_{n'=1}^{\eta}, \Lambda_k)
\end{aligned}$$

This distribution is conveniently chosen to cancel out with the second part of our target distribution in the Metropolis-Hastings acceptance ratio, which becomes the ratio of the Quasi-likelihoods:

$$\min\left\{1, \frac{Q(X^*|\psi^*; \eta, \tau, \zeta)}{Q(X^*|\psi; \eta, \tau, \zeta)}\right\}$$

3. **Subspace of** $\left\{\{\epsilon_{nkt}\}_{t=1}^{T}, s_{nk}, \theta_{nk}, z_{nk}\right\}_{K_+}^{\infty}$

Similarly, we propose a Metropolis within Gibbs approach where the proposal distribution is derived from the prior and can be decomposed as follow:

$$
\begin{aligned}
&P\left(\left\{\{\epsilon_{nkt}\}_{t=1}^{T_n}, s_{nk}, \theta_{nk}, z_{nk}\right\}_{K_+}^{\infty} | \alpha^{\theta}, \sigma^{\theta}, \{\{z_{n'k}=0\}_{n'\neq n, n'\leq\eta}\}_{K_+}^{\infty}\right) \\
=&P\left(\{z_{nk}\}_{K_+}^{\infty} | \{z_{n'k}=0\}_{n'\neq n, n'\leq\eta}\right) \prod_{K_+}^{\infty} P\left(\left\{\{\epsilon_{nkt}\}_{t=1}^{T_n}, s_{nk}, \theta_{nk}\right\}_{K_+}^{\infty} | \alpha^{\theta}, \sigma^{\theta}, \{z_{nk}\}_{K_+}^{\infty}\right) \\
=&P\left(\{z_{nk}\}_{K_+}^{\infty} | \{z_{n'k}=0\}_{n'\neq n, n'\leq\eta}\right) \\
&\times \prod_{K_+}^{\infty} \prod_{t}^{T_n} P(\epsilon_{nkt} | \alpha^{\theta}, \sigma^{\theta}, z_{nk}) \\
&\times \prod_{K_+}^{\infty} P\left(\theta_{nk} | \alpha^{\theta}, \sigma^{\theta}, z_{nk}\right) \\
&\times \prod_{K_+}^{\infty} P\left(s_{nk} | z_{nk}\right)
\end{aligned}
$$

Note that according to our prior, if $z_{nk}=0$, then $\{\epsilon_{nkt}\}_{t=1}^{T}, s_{nk}, \theta_{nk}$ are all 0 with certainty.

The distribution $P\left(\{z_{nk}\}_{K_+}^{\infty} | \{z_{n'k}=0\}_{n'\neq n, n'\leq\eta}\right)$ requires more attention. It can be seen as two separate elements. First a distribution over the number k^* of classes for which $z_{nk}=1$. Second, assuming some ordering on the Λ_k, a distribution to determine which of these classes have $z_{nk}=1$. Once the classes Λ_k have been determined, a reordering of the labels k can be done to assign the indexes between $K_+ + 1$ and $K_+ + k^*$ to these classes. We will denote the new proposed number of classes with non zero columns as K_+^*. Consequently, we can rewrite this distribution as:

$$
P\left(\{z_{nk}\}_{K_+}^{\infty} | \{z_{n'k}=0\}_{n'\neq n, n'\leq\eta}\right) = P(k^*) \prod_{k=K_+ +1}^{K_+^*} P(\Lambda_k).
$$

The distribution $P(k^*)$ can be derived from our prior and is $Poisson(\frac{\alpha^z \beta^z}{\beta^z + \eta - 1})$ (Griffiths and Ghahramani, 2011).

The proposal $\left\{\{\epsilon^*_{nkt}\}_{t=1}^{T}, s^*_{nk}, \theta^*_{nk}, z^*_{nk}\right\}_{K_+}^{\infty}$ being sampled from the prior, the Metropolis-Hastings acceptance ratio is the ratio of the Quasi-likelihoods:

$$
\min\left\{1, \frac{Q(X^*|\psi^*; \eta, \tau, \zeta)}{Q(X^*|\psi; \eta, \tau, \zeta)}\right\}
$$

4. **Subspace of** $\left(\alpha^{\theta}, \sigma^{\theta}\right)$

Our goal is to sample from the distribution $P\left(\alpha^\theta, \sigma^\theta | \left\{\{\{\epsilon_{nkt}\}_{t=1}^{T_n}, \theta_{nk}, z_{nk}\}_k^{K_+}\right\}_{n=1}^{\eta}\right)$, where $T_n = \tau$ if $n = \eta$, and $T_n = T$ otherwise. This distribution is independent of ζ.

We use a Metropolis-Hastings step with proposal distributions:

$$\alpha^* \sim Gamma(1,1)$$

$$\sigma^* \sim Gamma(1,1).$$

Consequently, the acceptance probability is as follow :

$$\min\left\{1, \prod_k^{K_+}\prod_n^{\eta}\left(\frac{P_\Gamma(\theta_{nk}|\alpha^*\sigma^*,1)\prod_t^{T_n}P_\Gamma(\epsilon_{nkt}|\alpha^*(1-\sigma^*),1)}{P_\Gamma(\theta_{nk}|\alpha^\theta\sigma^\theta,1)\prod_t^{T_n}P_\Gamma(\epsilon_{nkt}|\alpha^\theta(1-\sigma^\theta),1)}\right)^{z_{nk}}\right\}$$

where $P_\Gamma(\cdot|\alpha,\beta)$ is the PDF of the $Gamma(\alpha,\beta)$ distribution.

?????????—IGNORE EVERYTHING BELOW THIS LINE—?????????

- Initialization phase: Initialize $J \times P$ particles $\{\{\psi_{jp}\}_{p=1}^P\}_{j=1}^J$ by drawing from the prior
- Repeat until
  - Draw $\{z_{\eta k}, s_{\eta k}, \theta_{\eta k}\}$ for the newly added individual (*see section 4.3.4*)
  - For $\tau = 1$ to T *(Iterate over observations for each individual)*
    - Draw $\epsilon_{\eta k\tau}$ for the newly added observation (*see section 4.3.4*)
    - Set $\zeta = 1$
    - Repeat until $\zeta = \infty$:
      - Correction phase: increase ζ until $ESS_1/P < 0.5$ or $\zeta = \infty$ (*see section 4.3.5*)
      - Selection phase: re-sample particles within each group based on weight
      - Mutation phase: move each particle with an Metropolis-Hastings within Gibbs step using the target density $Q(\psi_{jp}|X^*;\eta,\tau,\zeta)$

The various parts of the algorithm are detailed in the following sections.

4.3.3 Initialization phase

Applied to every particle:

- Draw $\alpha^\theta \sim Gamma(1,1)$
- Draw $\sigma^\theta \sim Gamma(1,1)$
- Set $K_+ = 0$

4.3.4 Drawing $\{z_{\eta k}, s_{\eta k}, \theta_{\eta k}\}$ and $\{\epsilon_{\eta k \tau}\}$

To draw the $z_{\eta k}$ we follow the IBP approach. As detailed in section 3.2, when Λ_k has an active monotone attribute, we set $m_k = 1$. When the monotone active attributes have upper active levels (all levels above a specified level are active), we set $\rho_k = 1$. When the monotone active attributes have lower active levels (all levels below a specified level are active), we set $\rho_k = 0$.

- For $k = 1$ to K_+
  - If $\sum_{n=1}^{\eta-1} z_{nk} > 0$:
    - Draw $z_{nk} \sim Bernoulli(\frac{\sum_{n=1}^{\eta-1} z_{nk}}{\beta^z + \eta - 1})$
    - If $z_{nk} = 1$:
      - Draw $s_{\eta k} \sim \begin{cases} 0 & if\ m_k = 1 \text{ and } \rho_k = 1 \\ 1 & if\ m_k = 1 \text{ and } \rho_k = 0 \\ Bernoulli(\frac{1+\sum_{n=1}^{\eta-1} s_{nk}}{2+\eta-1}) & \text{otherwise} \end{cases}$
      - Draw $\theta_{\eta k} \sim Gamma(\sigma^\theta \alpha^\theta, 1)$
  - Else :
    - Set $z_{nk} = 0$
- If $K_+ = 0$:
  - While $k^* < 1$: Draw $k^* \sim Poisson(\frac{\alpha^z \beta^z}{\beta^z + \eta - 1})$
- Else:
  - Draw $k^* \sim Poisson(\frac{\alpha^z \beta^z}{\beta^z + \eta - 1})$
- Set $K_+ = K_+ + k^*$
- For $k = K_+ - k^*$ to K_+
  - Det $z_{\eta k} = 1$
  - Draw Λ_k
  - Draw $s_{\eta k} \sim \begin{cases} 0 & if\ m_k = 1 \text{ and } \rho_k = 1 \\ 1 & if\ m_k = 1 \text{ and } \rho_k = 0 \\ Bernoulli(\frac{1+\sum_{n=1}^{\eta-1} s_{nk}}{2+\eta-1}) & \text{otherwise} \end{cases}$
  - Draw $\theta_{\eta k} \sim Gamma(\sigma^\theta \alpha^\theta, 1)$

In order to draw the $\{\epsilon_{\eta k \tau}\}$ when a new observation is added, we follow the following simple algorithm:

- For $k = 1$ to K_+
  - If $z_{\eta k} = 1$: Draw $\epsilon_{\eta k \tau} \sim Gamma((1 - \sigma^\theta)\alpha^\theta, 1)$

4.3.5 Correction phase

Applied to every particle:

- Repeat until $ESS_1/P < 0.5$ or $\zeta = \infty$:
  - compute utility differences $du_{jp} = u(x^*_{\eta\tau}, \psi_{jp}) - \max_{x \in X_{\eta\tau}} u(x, \psi_{jp})$
  - If $\sum_p du_{1p} = 0$:
    - Set $\zeta = \infty$
    - compute weight $w_{jp} = \mathbb{1}\{du_{jp} = 0\}$
  - Else:
    - Set $\zeta = 2 \cdot \zeta$
    - compute weight $w_{jp} = w_{jp} \cdot 2^{du_{jp}}$
- compute $ESS_1/P = \frac{(\sum_p w_{1p})^2}{\sum_p w_{1p}^2}$

4.3.6 Selection phase

The selection phase draws a new set of P within group j particles with replacement from the set of old particles using weights w_{jp}. For each group j of particle:

- For each group j of particle:
  - For $p' = 1$ to P
    - Draw a particle ψ^* from set $\{\psi_{jp}\}_{p=1}^P$ using weights $\{w_{jp}\}_{p=1}^P$
    - Set $\psi_{jp'} = \psi^*$

4.3.7 Mutation phase

The mutation phase do an MCMC step using Gibbs sampling and Metropolis-Hastings (MH) accept-reject algorithm.

- For $n = 1$ to η
  - For $k = 1$ to K_+
    - If $\sum_{n' \neq n, n' \leq \eta} z_{n'k} > 0$:
      - Set $\psi^* = \psi_{jp}$
      - Draw $z^*_{nk} \sim Bernoulli(\frac{\sum_{n' \neq n, n' \leq \eta} z_{n'k}}{\beta^z + \eta - 1})$
      - If $z_{nk} = 1$: Draw $\{s^*_{nk}, \theta^*_{nk}, \{\epsilon^*_{nkt}\}\}$ (*see below*)
      - Store new values into ψ^*

· Set $\psi_{jp} = \psi^*$ with probability $\min\{1, \frac{Q(X^*|\psi^*;\eta,\tau,\zeta)}{Q(X^*|\psi_{jp};\eta,\tau,\zeta)}\}$

* Else :
  · Set $z_{nk} = 0$

– Draw $k^* \sim Poisson(\frac{\alpha^z \beta^z}{\beta^z + \eta - 1})$

– Set $K_+ = K_+ + k^*$

– For $k = K_+ - k^*$ to K_+

* Set $\psi^* = \psi_{jp}$
* Set $z^*_{nk} = 1$
* Draw Λ^*_k
* Draw $s^*_{nk} \sim \begin{cases} 0 & if\ \Lambda^*_k \text{ has active upper monotone atributes} \\ 1 & if\ \Lambda^*_k \text{ has active lower monotone atributes} \\ Bernoulli(\frac{1+\sum_{n=1}^{\eta-1} s_{nk}}{2+\eta-1}) & \text{otherwise} \end{cases}$
* Draw $\theta^*_{nk} \sim Gamma(\sigma^\theta \alpha^\theta, 1)$
* Store new values into ψ^*
* Set $\psi_{jp} = \psi^*$ with probability $\min\{1, \frac{Q(X^*|\psi^*;\eta,\tau,\zeta)}{Q(X^*|\psi_{jp};\eta,\tau,\zeta)}\}$

- Draw $\alpha^* \sim Gamma(1, 1)$
- Draw $\sigma^* \sim Gamma(1, 1)$
- Set $(\alpha^\theta, \sigma^\theta) = (\alpha^*, \sigma^*)$ with probability

$$\min\left\{1, \prod_k^{K_+} \prod_n^{\eta} \left(\frac{P_\Gamma(\theta_{nk}|\alpha^*\sigma^*, 1) \prod_t^{T_n} P_\Gamma(\epsilon_{nkt}|\alpha^*(1-\sigma^*), 1)}{P_\Gamma(\theta_{nk}|\alpha^\theta\sigma^\theta, 1) \prod_t^{T_n} P_\Gamma(\epsilon_{nkt}|\alpha^\theta(1-\sigma^\theta), 1)} \right)^{z_{nk}} \right\}$$

where $P_\Gamma(\cdot|\alpha, \beta)$ is the PDF of the $Gamma(\alpha, \beta)$ distribution, and $T_n = \tau$ if $n = \eta$ and $T_n = T$ otherwise.

For this phase, in order to draw $\{s^*_{nk}, \theta^*_{nk}, \{\epsilon^*_{nkt}\}\}$, we propose the following process:

- Draw $s^*_{nk} \sim \begin{cases} 0 & if\ m^*_k = 1 \text{ and } \rho^*_k = 1 \\ 1 & if\ m^*_k = 1 \text{ and } \rho^*_k = 0 \\ Bernoulli(\frac{1+\sum_{n' \neq n, n' \leq \eta}^{\eta} s_{nk}}{2+\eta-1}) & \text{otherwise} \end{cases}$
- Draw $\theta^*_{nk} \sim Gamma(\sigma^\theta \alpha^\theta, 1)$
- If $n < \eta$:
  – For $t = 1$ to T:
    * $\epsilon^*_{nkt} \sim Gamma((1 - \sigma^\theta)\alpha^\theta, 1)$
- Else:
  – For $t = 1$ to τ:
    * $\epsilon^*_{nkt} \sim Gamma((1 - \sigma^\theta)\alpha^\theta, 1)$

References

G. Durham and J. Geweke. Adaptive sequential posterior simulators for massively parallel computing environments. In *Bayesian Model Comparison*, pages 1–44. Emerald Group Publishing Limited, 2014.

T. L. Griffiths and Z. Ghahramani. The indian buffet process: An introduction and review. *The Journal of Machine Learning Research*, 12:1185–1224, 2011.