Bayesian Analysis of Ecological Momentary Assessment (EMA) Data

Arne Leijon

January 3, 2022

1 Introduction

Ecological Momentary Assessment (EMA) has become a popular method, in particular, to investigate the benefit and subjective quality of hearing instruments in the user's real life (Holube et al., 2020). The main advantage is that EMA can achieve better ecological validity than conventional tests performed in a laboratory. In EMA, each study participant is requested to respond to a questionnaire during normal everyday life, typically several times per day. Some questions may address the current *real-life scenario*, i.e., the physical environment and the user's activity and intentions in that particular environment. The participant may also be also asked to rate, e.g., the pleasantness of aided sound or the difficulty of understanding speech in the current scenario, or any other *perceptual attribute* of interest in the study.

Most conveniently, the questions are presented and responses collected by a special-purpose app in the user's smartphone. The smartphone can also measure some physical characteristics of the current environment, such as the sound level or noise spectrum. The timing of assessments may be determined randomly by the app, and users may also be allowed to initiate an assessment at any time of their own choice.

1.1 EMA Challenges

The data set recorded in an EMA study presents several statistical challenges, some of which were discussed in depth by Oleson et al. (2021):

- There may be a large amount of data from each participant, but the number of assessments can vary greatly among respondents.

- Some of the responses are *nominal* (categorical). For example, the current real-life environment may be characterized as one of a fixed set of *Common Sound Scenarios* (CoSS) (Wolters et al., 2016), and one of a few available *hearing-aid programs* may be selected.

- Other responses are *ordinal*. For example, the user may rate the subjective speech understanding by selecting one response from a range of discrete alternatives like "nothing at all", "very difficult", "difficult", "easy", "perfect" (e.g., von Gablenz et al., 2021). The number of ordinal response alternatives may differ among questions.

- With many assessments by each participant, the responses collected in total from all respondents are *not statistically independent*. Therefore, a *multi-level* approach is needed, such that the responses are analyzed as nested within each individual, separate from the next level of variability across respondents.

- Typically, the ordinal ratings can not be encoded numerically to represent points on an *interval scale*: We cannot take it for granted that the steps between response categories are perceptually equal in magnitude. Therefore, it is questionable to aggregate responses within individuals by conventional measures such as mean and variance of ordinal ratings.

- Individual respondents might interpret and use the ordinal response scale in different ways, depending on their personality. For example, some people might tend to use the more extreme response alternatives, while others hesitate to do so (Rossi et al., 2001). Therefore, it may be questionable to encode ordinal responses by the same numerical value for all participants.

- The subjective benefit of a hearing instrument usually depends on the real-life environment. Therefore, the *ordinal* responses to performance-related questions should be analyzed as *conditional on the nominal* responses to the scenario questions.

- The nominal responses about the real-life scenario might also be used as a kind of outcome measure. For example, if the participant is more likely to visit a challenging sound environment when using a particular hearing-aid program, this might be interpreted as a benefit of the signal processing methods in that program.

- A complete experiment may be designed to include two or more *stages*, yielding separate series of EMA data. For example, participants may

be asked to record base-line responses first without hearing aids, and then to record a similar follow-up series a few months later, after acclimatization to their new hearing aids (e.g., von Gablenz et al., 2021).

It has been overwhelmingly common in the behavioral-science literature to apply conventional statistical measures and models such as mean and variance, t-test, ANOVA, linear or non-linear regression, using the raw subjective ratings as input, although all these analysis methods are *metric*, i.e., they presume that the input data have interval-scale properties. The main reason is that these conventional methods are readily available in statistical program packages. Oleson et al. (2021) argued that general-purpose statistical approaches such as linear or generalized-linear mixed regression models can adequately capture the fundamental characteristics of EMA data, in spite of the complex nature of the data.

However, analyzing ordinal data as if they were metric can cause errors (Liddell and Kruschke, 2018). The errors might even have severe scientific consequences, such as indicating a statistically significant effect in the *wrong direction*. These potential errors cannot be detected by calculations such as conventional normality tests. The only way to detect an error would be to compare the analysis results with those of another model that does not presume metric data.

A generalized linear mixed regression model can indeed account for some of the non-metric properties of ordinal ratings and nominal scenario categories, but it is not easy to formulate the model in the framework of a general regression program package. As noted by Oleson et al. (2021), the full complexity of EMA data could be handled properly by a hierarchical Bayesian model, but no current software is readily available that could perform the analysis automatically.

The purpose of the present work is to create such a model and make it freely and easily available for researchers in Audiology. An R package for advanced Bayesian multi-level generalized regression is available (Bürkner, 2018), but it is not quite equivalent to the present proposal. Although this advanced general program should give similar results, it may be easier in practice to use the present special-purpose package.

1.2 Requirements on the Statistical Analysis

The purpose of an EMA-based audiological study is usually to evaluate the benefit of a new hearing aid, or a new signal-processing feature, or the effect of some other kind of intervention, in comparison to a base-line reference. Then, the goal of a statistical analysis is to determine whether the intervention has

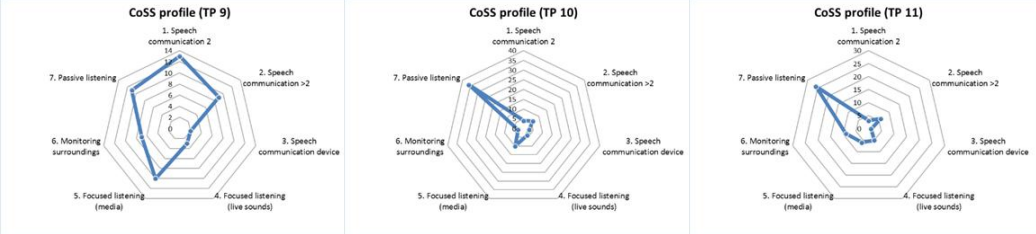


Figure 1: Example of scenario profiles collected from three hearing-aid users answering questions about *Common Sound Scenarios* with seven response alternatives. The profiles are illustrated by spider diagrams where the radial distance from the center indicates the response count and the angular directions show the response categories. Perhaps the second and third count profiles are so similar that the difference is just a random effect?

a *true benefit*, such that an apparent improvement is not simply a random effect caused by the limited number of study participants and the limited number of assessments by each respondent. The present report proposes a Bayesian probabilistic model that can answer the following types of research questions:

- Is there a *statistically credible* change in the *scenario profile* of users, i.e., the proportion of daily activities spent in various scenarios, as exemplified in Fig. 1? For example, does the user participate in more acoustically demanding daily activities when using hearing aid B than when using the reference hearing aid A?

- Is there a *statistically credible* improvement in users' subjective evaluation of some *perceptual attribute* of hearing-aid performance in a *specific real-life scenario*? For example, does hearing-aid program B make it easier to understand speech in the most demanding scenario, while program A performs better in other scenarios?

- Is there a *statistically credible* overall change in users' subjective impressions, when averaged across *scenarios*? For example, is program A generally better than program B, when weighted by the usage probabilities in various scenarios? Is some perceptual attribute generally better *after* an intervention than it was *before*?

- Are there *statistically credible* differences between the perceived hearing-aid performance, or the scenario profiles, in *separate sub-populations*, e.g., old versus young respondents?

Here, the *"credibility"* can mean three quite different quantities:

1. The predictive probability that the result is true for an unseen *random individual in the population* from which the study participants were recruited.

2. The predictive probability that the result is true for the *mean (median) in the population* from which the study participants were recruited.

3. The probability that the result is true for each and any *individual participant* in the study.

The first of these probability measures may be most important in a study designed by a hearing-aid manufacturer to predict the marketing success of some new hearing aid feature. The second probability measure is most closely related to the statistical significance as estimated by conventional hypothesis tests. The third credibility measure might be most important in a clinical study to quantify the benefit for individual clients.

2 Theory — Notation and Models

Let us assume a study involves N participants in total, all describing the *scenario* at each assessment by a choice among K *nominal* categories. The participant may also give an *ordinal rating* for each of I questions regarding the *perceptual attributes* evaluated in the study.

This section will show how all recorded EMA data from the nth participant can be encoded in a *count profile* vector $\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nJ})$ with J integer elements x_{nj}. The likelihood of the observed count profile is quantified by a probabilistic model specified by a parameter vector $\boldsymbol{\xi}_n$ with D elements ξ_{nd}, such that the log-likelihood of the observed count profile is just a scalar product

$$\ln p(\boldsymbol{x}_n \mid \boldsymbol{\xi}_n) = \boldsymbol{x}_n \cdot \boldsymbol{f}(\boldsymbol{\xi}_n) = \sum_{j=1}^{J} x_{nj} f_j(\boldsymbol{\xi}_n) \quad \forall n. \tag{1}$$

Here $\boldsymbol{f} = (f_1, \ldots, f_J)$ is a vector-valued function determined by the structure of the EMA responses and the desired analysis results.

The distribution of individual parameter vectors in the population, from which participants were recruited, is estimated in a separate *population model*. The number of EMA recordings may vary greatly among participants. If one subject provides unusually many EMA recordings, this will tend to improve

the precision of the individual parameter estimate, but this participant will still automatically have the same weight as every other participant for the estimation of population characteristics.

2.1 Individual Response Models

2.1.1 Nominal EMA Scenario Responses

The rth recorded EMA response by the nth participant in the tth test stage includes a one-of-K binary vector $\boldsymbol{z}_{nrt}$ with one element $z_{nrt,k} = 1$, indicating that the kth scenario category was selected, with all other $z_{nrt,j\neq k} = 0$.

The scenario response may be just a one-dimensional vector, with elements only identifying the main scenario (e.g., CoSS) category. If the study includes one or more scenario dimensions identifying, e.g., the participant's intention or the selected hearing-aid program, the scenario response from each subject might also be indexed as $z_{nrt,k_1,k_2\ldots}$ where k_1 is the category index in the main scenario dimension, and k_2 is the index in, e.g., the "HA program" sub-dimension, etc. The multi-dimensional array index $(k_1, k_2, \ldots)$ can always be uniquely represented by an equivalent linear index k, and vice versa. Therefore, we use mainly the one-dimensional index notation in the mathematical formulation.

The analysis model assumes that all responses from each participant are determined by a fixed but unknown *probability profile* $\boldsymbol{u}_{nt} = (u_{nt,1}, \ldots, u_{nt,K})$. Here, $u_{nt,k} \in [0, 1]$ is the probability that the nth participant reports from the kth scenario at any assessment in the tth test stage. The probability profile is one of the main desired result quantities to be estimated from the data.

The probability profile is assumed to be the same at all EMA replications but may vary between participants and between test stages. Thus, the responses follow a categorical distribution, conditional on the probability profile,

$$p(\boldsymbol{z}_{nrt} \mid \boldsymbol{u}_{nt}) = \prod_{k=1}^{K} u_{ntk}^{z_{nrtk}}; \quad \sum_{k=1}^{K} u_{ntk} = 1; \quad \forall n, r, t. \tag{2}$$

It is computationally convenient to express the probability profiles as an exponential function $\boldsymbol{u}_{nt}(\boldsymbol{\alpha}_{nt})$ with elements

$$u_{ntk}(\boldsymbol{\alpha}_{nt}) = \frac{e^{\alpha_{ntk}}}{\sum_{j=1}^{K} e^{\alpha_{ntj}}}, \tag{3}$$

given the mapped-parameter vectors $\boldsymbol{\alpha}_{nt} = (\alpha_{nt,1}, \ldots, \alpha_{nt,K})$, with elements $\alpha_{ntk} \in (-\infty, +\infty)$ for all $t = 1, \ldots, T$ and $k = 1, \ldots, K$. These parameters

for the nth subject are from now on denoted as the TK first elements in the individual parameter vector $\boldsymbol{\xi}_n$.

The total array of scenario responses is denoted $\underline{\boldsymbol{z}} = (\ldots, \boldsymbol{z}_{nrt}, \ldots)$, including recordings from all participants in all test stages, and the corresponding array of individual parameters is written $\underline{\boldsymbol{\xi}} = (\ldots, \boldsymbol{\xi}_n, \ldots)$. The model presumes that all responses are *conditionally independent* across subjects, assessments, and test stages, given the individual parameters. The probability mass of all observed nominal scenario data can then be expressed as

$$p(\underline{\boldsymbol{z}} \mid \underline{\boldsymbol{\xi}}) = \prod_{n=0}^{N-1} \prod_{r=1}^{R_n} \prod_{t=1}^{T} \prod_{k=1}^{K} u_{ntk}(\boldsymbol{\xi}_n)^{z_{nrt,k}}. \tag{4}$$

2.1.2 Ordinal EMA Ratings

The ordinal rating responses are analyzed with a variant of *Item Response Theory* (IRT). IRT is a family of probabilistic models designed to handle issues with ordinal ratings which are common to test instruments for any purpose in social, psychological, or educational research. The model includes individual parameters that can account for the possibility that different people use the ordinal response scale in different ways.

There is a rich literature on IRT, including several text books (e.g., Fox, 2010; Nering and Ostini, 2010) with good reviews of the literature. The *Graded Response* IRT model (Samejima, 1969), also called *Cumulative Model* (Bürkner and Vuorre, 2019), is mathematically very closely related to signal-detection theory and choice models that have a long history of use in psycho-acoustical research (e.g., Thurstone, 1927; Bradley and Terry, 1952; Luce, 1959; Durlach and Braida, 1969). These models are sometimes called "ordinal-probit" or "ordinal-logit".

The basic feature of ordinal rating models is that subjective responses are regarded as indicators ("symptoms") that are only probabilistically related to the individual trait or ability that is to be measured. The true individual trait cannot be directly observed but only indirectly estimated on the basis of test responses. The model treats each response as determined by an outcome of a latent random variable. The location (mean or median) of the probability distribution of that latent variable represents the individual characteristic to be estimated, whereas the response probabilities also depend on other parameters that may differ among individuals.

Similar to the nominal data, the ordinal rating response to the ith *attribute question* (with L_i ordinal response alternatives) is denoted by a one-of-L_i binary vector $\boldsymbol{y}_{nri}$ with one element $y_{nri,l} = 1$ indicating that the nth participant gave the lth ordinal response to the ith question in the rth assess-

ment, and all other $y_{nri,j\neq l} = 0$. The rth assessment is also characterized by the multi-dimensional *scenario* index vector $\boldsymbol{k}(r) = (k_0, k_1, \ldots)(r)$. The first index identifies the *test stage* $k_0 = t$ which is defined by the researcher, and the other indices $(k_1, k_2, \ldots)$ specify the scenario as recorded by the respondent, in the same way as in Sec. 2.1.1. The multi-dimensional array index $\boldsymbol{k}(r)$ can always be uniquely represented by an equivalent scalar linear index $k(r)$, so this notation is used where it is unambiguous.

In the model, illustrated in Fig. 2, each ordinal response is determined by an outcome of a continuous real-valued latent random variable $Y_{nik(r)}$. The lth ordinal response is given whenever the latent variable falls in an interval $\tau_{ni,l-1} < Y_{nik(r)} \leq \tau_{ni,l}$, where the thresholds separating the intervals form an increasing sequence $(-\infty = \tau_{ni,0} < \tau_{ni,1}, \ldots, < \tau_{ni,L_i} = +\infty)$. The thresholds may differ between respondents and between attribute questions, but are assumed to be identical in all assessments of the same attribute by each respondent. The latent variable is drawn from a logistic[1] probability distribution with location θ_{nik} and unity[2] scale. The location θ_{nik} is the desired outcome measure of the perceptual attribute that is to be estimated from the observed data. Although all responses are only discrete and ordinal, the estimated outcome measure θ_{nik} is continuous on an interval scale defined by the model.

The location parameter array for the nth participant can, in principle, include a very large number of free parameters, one for each combination of nominal categories from all scenario dimensions. However, just as in linear regression, the model may be simplified to consider only a linear combination of a smaller set of *scenario effects*, e.g., by expressing $\theta_{nik} = \beta_{nik_0} + \beta_{nik_1} + \beta_{nik_2}$, or $\theta_{nik} = \beta_{nik_0k_1} + \beta_{nik_2}$. Here, the value β_{nik_f} represents the *main effect* of the k_fth category in the fth scenario dimension, while $\beta_{nik_0k_1}$ also represents *interaction effects* between any combination (k_0, k_1) of categories in the zeroth and first scenario dimensions. Regardless of these indexing details, the location parameter can always be specified as a fixed linear function $\theta_{nik}(\boldsymbol{\beta}_{ni})$ of an *effect vector* $\boldsymbol{\beta}_{ni} = (\ldots, \beta_{nij}, \ldots)$ including all the regression effects for the ith perceptual attribute that are to be estimated in the statistical analysis.

Individual response thresholds: To ensure that the response thresholds form a strictly increasing sequence, and for numerical stability, it is convenient to map the response intervals to the range $[0, 1]$ using the logistic

[1]The Cumulative Model can also use the normal distribution.

[2]The unity scale is no restriction of generality, as the model scale is arbitrary anyway.

distribution function $F(\ )$ from (8), as

$$F(\tau_{nil}) = \frac{\sum_{j=1}^{l} e^{\eta_{nij}}}{\sum_{j=1}^{L_i} e^{\eta_{nij}}}. \tag{5}$$

Here the parameter vector $\boldsymbol{\eta}_{ni} = (\eta_{ni,1}, \ldots, \eta_{ni,l}, \ldots, \eta_{ni,L_i})$ includes the logarithms of the relative interval widths in this mapped range. The inverse mapping defines the thresholds as

$$\tau_{ni,l}(\boldsymbol{\eta}_{ni}) = \ln \frac{\sum_{j=1}^{l} e^{\eta_{nij}}}{\sum_{j=l+1}^{L_i} e^{\eta_{nij}}}. \tag{6}$$

Adding a constant to all elements in $\boldsymbol{\eta}_{ni}$ does not change the resulting thresholds, so the L_i values actually define only $L_i - 1$ free thresholds, as desired. From now on, we denote the individual ordinal-model parameters $\boldsymbol{\beta}_n, \boldsymbol{\eta}_n$ as parts of the total individual parameter vector $\boldsymbol{\xi}_n$, with an indexing scheme determined by the structure of the model and the desired analysis results.

Finally, using this simplified parameter notation, the conditional probability of any rating response, given the model parameters, is a known function of the parameters,

$$\begin{aligned} P(y_{nri,l} = 1 \mid \boldsymbol{\xi}_n) &= P(\tau_{ni,l-1} < Y_{ni\,k(r)} \leq \tau_{ni,l} \mid \boldsymbol{\beta}_{ni}) = \\ &= F\left(\tau_{ni,l} - \theta_{nik(r)}\right) - F\left(\tau_{ni,l-1} - \theta_{nik(r)}\right) = v_{ik(r),l}(\boldsymbol{\xi}_n), \quad \forall n, r, i \end{aligned} \tag{7}$$

where $F(\ )$ is the cumulative distribution function for a standard logistic-distributed random variable,

$$F(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

This part of the EMA model is very similar to a previous model for paired-comparison data (Leijon et al., 2019) with the Bradley-Terry-Luce (BTL) "ordinal-logit" choice model[3] (Bradley and Terry, 1952), i.e., assuming a logistic distribution for the latent sensory variables.

All ratings, gathered in an array $\underline{\boldsymbol{y}} = (\ldots, \boldsymbol{y}_{nri}, \ldots)$, are assumed conditionally independent across subjects, assessments, and questions, given the individual parameters $\underline{\boldsymbol{\xi}} = (\ldots, \boldsymbol{\xi}_n, \ldots)$. Thus the probability mass for the total set of observed ordinal response data can be written as

$$p(\underline{\boldsymbol{y}} \mid \underline{\boldsymbol{\xi}}) = \prod_{n=1}^{N} \prod_{r=1}^{R_n} \prod_{i=1}^{I} \prod_{l=1}^{L_i} v_{ik(r),l}(\boldsymbol{\xi}_n)^{y_{nri,l}} \tag{9}$$

[3]The implemented model can also use the "ordinal-probit" variant (Thurstone Case V) assuming normal distribution for the latent variables.

Since this regression model allows both the trait variables θ_{nik} and the threshold variables τ_{nil} to be freely variable for each respondent, the model is underdetermined: If a fixed constant value is added to all θ_{nik} and all τ_{nil}, the probability (7) of observed responses does not change. The weakly informative prior distributions slightly favors solutions with parameters near zero, so the learning always converges. However, the indeterminacy allows some artificial variability in the parameter values. To avoid this variance, the parameter values must be somehow restricted.

The current implementation[4] allows the researcher either (1, default) to force the median response threshold to zero, or (2) to force the average sensory-variable location to zero, for each respondent and each attribute.

2.1.3 Joint Nominal and Ordinal Response Data

With the data representation defined in Secs. 2.1.1 and 2.1.2, all data from the rth recording of the nth participant can be concatenated into a one-dimensional vector $\boldsymbol{x}_n = (\boldsymbol{z}_n, \boldsymbol{y}_n)$, just as all individual model parameters were gathered in a single vector $\boldsymbol{\xi}_n$. Using this notation, the total log-likelihood in (1) obviously follows from (4) and (9).

2.2 Individual vs. Population Models

In this model the joint nominal and ordinal response patterns are assumed probabilistically determined by individual parameters $\boldsymbol{\xi}_n$. Of course, these parameters cannot be directly observed. Their values must be estimated from the recorded data. In the Bayesian framework, all these parameters are regarded as random variables.

However, as all participants were recruited at random from the same *population*, as defined by the researcher, the model treats each individual parameter vector $\boldsymbol{\xi}_n$ as a sample drawn at random from a *population distribution* with density function $p(\boldsymbol{\xi} \mid \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}})$ specified by another set of parameters, as defined in Eq. A.1 in App. A. In this way the population model acts as a prior model for the individual parameter distributions.

The distribution of individual parameter vectors in the population might have been chosen as, e.g., a multivariate Gaussian (normal) distribution with full covariance matrix, i.e., with a very large number of parameters, or perhaps with a more restricted correlation matrix with fewer parameters as in (Bürkner, 2017). For the present purpose it is proposed instead to use a *Gaussian Mixture Model (GMM)*. Mixture models have the advantage that

[4] Alternatively, the sampling might be constrained to a manifold, where the restriction is satisfied. Using a more strongly informative prior can not solve this problem.

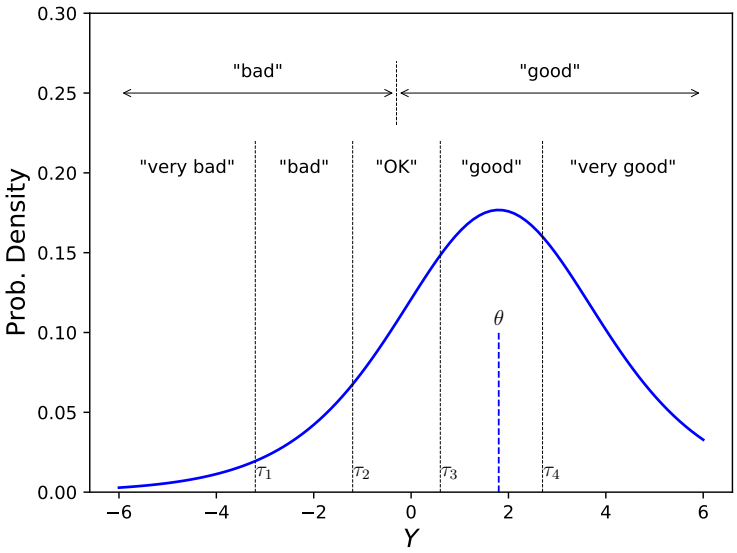


Figure 2: Example of a conditional probability density function of the sensory random variable Y that determines the response to a question about a perceptual attribute, e.g., speech understanding, given a true mean parameter $\theta = 1.8$ on the logit scale. Decision intervals are indicated for one experiment allowing only two responses "bad" or "good", with a response threshold $\tau_1 = -0.2$, and another experiment with the same attribute distribution, allowing five ordinal response alternatives, "very bad", "bad", ..., "very good" with response thresholds $\tau_1, \ldots, \tau_4$.

they can represent arbitrarily complex non-linear dependencies between vector elements, not only ordinary linear correlations[5], although each mixture component has independent elements. Another advantage is that the model complexity is automatically determined during learning, by the general *Occam's Razor* feature of Bayesian learning. Thus, after learning, the mixture model might automatically come out as including only a single Gaussian component with independent (uncorrelated) elements, if this is sufficient for the given data set, or as a multi-component mixture if the data indicate a more complex dependency structure.

The population density function belongs to the GMM family with a fixed mathematical form, but the parameter values $(\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}})$ are unknown a priori and must be estimated from the recorded data set. A weakly informative hyper-prior is assigned for these parameters, as discussed in App. A.2.

If the respondents were recruited from G separate *sub-populations*, e.g., younger vs. older people, the subject groups are distinguished by sets $\mathcal{S}_g$ of indices, with $n \in \mathcal{S}_g$ indicating that the nth respondent is a member of the group recruited from the gth sub-population. In this way, the complete population GMM represents individuals in all sub-groups. The separate sub-population models differ only in the mixture weights, using the same set of mixture components.

2.3 Variational Model Inference

As described in in App. B, variational learning (e.g., Bishop, 2006, Ch. 10) is used to derive approximate posterior density functions $q(\boldsymbol{\xi}_n)$, for the parameters of all participants, as well as a separate posterior density function $q(\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}})$ for the population parameters, given all observed data.

The individual results are estimated using the response data from each participant, but these individual estimates are also somewhat regularized by the population model: If the response pattern from one respondent deviates a lot from the data of most other participants, the hierarchical model will tend to "explain" any such extreme deviations as a random effect of the limited number of responses rather than an extreme deviation in the true individual characteristics.

The population model is simultaneously adapted to all the individual data together with a weakly informative hyper-prior density defined in App. A.2 for the population parameters. The combined hierarchical Bayesian model automatically includes measures of the uncertainty of each individual result as well as the inter-individual variability in the population.

[5]Random variables can be statistically dependent even if their correlation is zero.

2.4 Predictive Distributions

Both the individual and population models are generative, i.e., they can be used to calculate three desired predictive distributions as defined in App. C,

1. for an unseen *random individual in the population* from which the participants were recruited,
2. for the *population mean*,
3. for each individual in the *group of participants*.

The predictive distributions are used to evaluate the *joint credibility* for combinations of single hypotheses, as described in (Leijon et al., 2016, App. C).

3 Experimental Methods

In order to quantify the precision of the analysis results, it is necessary to evaluate the difference between estimated model parameters and the corresponding true parameter values. The only way to do this is to use simulated EMA experiments, because the true values are, of course, never known in real EMA studies.

3.1 Simulated EMA Data

To illustrate the performance of the analysis model, a synthetic EMA data set was constructed. An array of 2×7 nominal scenario categories was defined in two scenario dimensions: Two *hearing aid programs*, called A and B, were assumed to be freely selected by respondents in seven *Common Sound Scenarios (CoSS)*, labelled $1, 2, ..., 7$. The mean scenario probabilities were assigned with the highest values in CoSS-category 7 for program A, and in CoSS-category 1 for program B. Similarly, one perceptual attribute, called *Speech*, was simulated with the highest values in CoSS-category 7 for program A, and in in CoSS-category 1 for program B.

Thus, the simulation assumed that program A performed best in CoSS-category 7, while program B was best in CoSS-category 1. There was no difference between the hearing-aid programs if the performance is averaged across all CoSS categories.

**** Try different intra-individual theta slope by CoSS, i.e., unequal variability across CoSS categories, similar to the unequal SDs tested by Liddell and Kruschke..Compare with t-test of difference between HA A and B, using average rating across CoSS categories. ***

It was also assumed that these performance differences would cause the respondents to select each hearing-aid program with higher probability in CoSS situations where the program performed better, but still with some random variability in the choice. The simulated true scenario probabilities and inter-individual variations are shown together with the estimated results in Figs. 3a and 3b. The simulated true latent sensory variables and the inter-individual variations are shown together with the estimated results in Figs. 4a and 4b. The true response thresholds were chosen to yield a uniform distribution of ordinal response categories when the true sensory-variable location is zero. The thresholds were identical for all simulated subjects.

EMA data were generated for 20 simulated respondents, yielding 971 EMA records in total. The number of EMA records was randomly chosen for each respondent, with uniform distribution between 20 to 80 EMA records per subject.

3.2 Real EMA Evaluation of Hearing Aids

xxxx

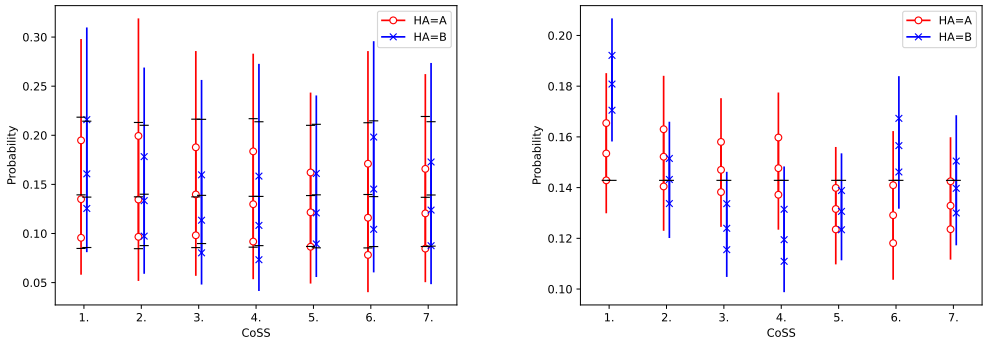
4 Results

4.1 Population Results for Simulated EMA Data

The estimated scenario probability profiles are shown in Figs. 3a and 3b. The regression results for the latent sensory variables for the simulated attribute *Speech*, and its dependence on scenarios, are shown in Figs. 4a and 4b. These results were obtained using the logistic (Bradley-Terry-Luce) distribution for the latent variables. The results were quite similar with the normal distribution (Thurstone model). The mixture model initially allowed 20 separate Gaussian components, but the learning converged on using only a single component for all subjects. This seems reasonable since the simulation generated random data with independent elements of the parameter vector.

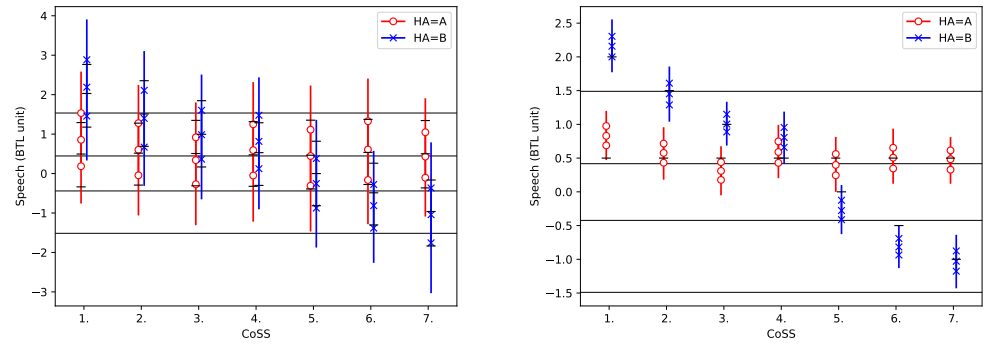
4.2 Individual Results for Simulated EMA Data

*** Compare with NAP analysis results ??? **********



(a) Random unseen individual (b) Population mean

Figure 3: Scenario probability profiles estimated for a *random unseen individual* in the population and for the *population mean*. Vertical lines show 90% credible intervals. Marker symbols show 25-, 50-, and 75-percentiles. Short black horizontal lines indicate the corresponding true 90% credible intervals and the population mean. This simulated experiment included 20 subjects with a total of 971 EMA records.



(a) Random unseen individual (b) Population mean

Figure 4: Sensory attribute *Speech* estimated by regression as a function of scenarios, for a *random unseen individual* in the population and for the *population mean*. Vertical lines show 90% credible intervals. Marker symbols show 25-, 50-, and 75-percentiles. Short black horizontal lines indicate the corresponding true 90% credible intervals and the population mean. This simulated experiment included 20 subjects with a total of 971 EMA records.

4.3 Real EMA Evaluation

5 Discussion

The estimated 90% credible intervals for the population mean in Figs. 3b and 4b include the corresponding true value in nearly all cases, as expected. The 90% credible intervals for random individuals in the population, in Figs. 3a and 4a, somewhat over-estimated the true 90% range of inter-individual variations in the population. Thus, the simulation results suggest that the estimation yields a rather conservative estimate for the inter-individual variance, probably because of the limited number of participants and limited number of EMA records. The simulation generated about 50 EMA records per subject. Thus, with 14 separate scenario categories, some subjects might show no response at all for some scenarios.

The proposed analysis method has been implemented as a python package `EmaCalc`, freely available at the Python Package Index (PyPi). The code package also includes simulation functions allowing the user to validate the performance of the method and to plan a practical experiment.

6 Conclusions

A new Bayesian parametric analysis method for EMA data was presented. The method was evaluated with simulated experimental data and exemplified using data from a real experiment. The results of the simulation study indicate that the model could estimate the true population values with good accuracy.

References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York, NY, USA.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 39:324–345.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395–411.

Bürkner, P.-C. and Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101.

Durlach, N. and Braida, L. (1969). Intensity perception. I. Preliminary theory of intensity resolution. *J Acoust Soc Am*, 46(2, pt. 2):372–383.

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. Statistics for Social and Behavioral Sciences. Springer.

Holube, I., von Gablenz, P., and Bitzer, J. (2020). Ecological momentary assessment (EMA) in audiology: Current state, challenges, and future directions. *Ear Hear*, 41(S1):79S–90S.

Leijon, A., Dahlquist, M., and Smeds, K. (2019). Bayesian analysis of paired-comparison sound quality ratings. *J Acoust Soc Am*, 146(5):3174–3183.

Leijon, A., Henter, G. E., and Dahlquist, M. (2016). Bayesian analysis of phoneme confusion matrices. *IEEE Trans Audio Speech Lang Proc*, 24(3):469–482.

Liddell, T. M. and Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79:328–348.

Luce, R. D. (1959). *Individual choice behavior: a theoretical analysis*. Wiley, New York, NY, USA.

Nering, M. L. and Ostini, R., editors (2010). *Handbook of polytomous item response theory models*. Routledge, New York, NY, USA.

Oleson, J. J., Jones, M. A., Jorgensen, E. J., and Wu, Y.-H. (2021). Statistical considerations for analyzing ecological momentary assessment data. *J Speech Lang Hear Res*, ePub ahead of issue:1–17.

Rossi, P. E., Gilula, Z., and Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association*, 96(453):20–31.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph Supplement No.17.

Singh, S. and Poczos, B. (2016). Analysis of k-nearest neighbor distances with application to entropy estimation. *arXiv:1603.08578 [math.ST]*.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4):273–286.

von Gablenz, P., Kowalk, U., Bitzer, J., Meis, M., and Holube, I. (2021). Individual hearing aid benefit in real life evaluated using ecological momentary assessment. *Trends in Hearing*, 25:1–18.

Wolters, F., Smeds, K., Schmidt, E., and Norup, C. (2016). Common sound scenarios: A context-driven categorization of everyday sound environments for application in hearing-device research. *J Am Acad Audiol*, 27(7):527–540.

A Population Model

The nominal and the ordinal response patterns from the nth participant were assumed determined by parameters $\boldsymbol{\xi}_n = (\ldots, \boldsymbol{\alpha}_{nt}, \ldots, \boldsymbol{\beta}_{ni}, \ldots, \boldsymbol{\eta}_{ni}, \ldots)^T$ as defined in (4) and (9). In the following, the elements in this vector are denoted either as ξ_{nd} or, equivalently, by the separate symbols α, β, η for the three subtypes of parameters. The total number of parameters and the indexing depends on the experimental structure and the types of regression effects to be estimated, but this notational equivalence is defined once and for all when the analysis model is set up.

A.1 Gaussian Mixture Model

The individual parameter vectors are assumed drawn from a population distribution in the form of a *Gaussian mixture model* (GMM),

$$p(\boldsymbol{\xi}_n \mid \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}) = \prod_{c=1}^{M} \left[\prod_{d=1}^{D} \sqrt{\frac{\lambda_{cd}}{2\pi}} e^{-(\xi_{nd}-\mu_{cd})^2 \lambda_{cd}/2} \right]^{\zeta_{nc}} \tag{A.1}$$

where $\boldsymbol{\mu}_c = (\mu_{c,1}, \ldots, \mu_{c,D})$ is the mean vector, and $\boldsymbol{\lambda}_c = (\lambda_{c,1}, \ldots, \lambda_{c,D})$ is the precision vector (inverse variance) of the cth mixture component, with $\underline{\boldsymbol{\mu}} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M)$ and $\underline{\boldsymbol{\lambda}} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_M)$ denoting the total set of parameters for the complete set of mixture components. It is no serious restriction to assume independent elements (diagonal covariance) in each mixture component, because the complete mixture model can still capture any statistical dependence between elements of $\boldsymbol{\xi}_n$.

The chosen mixture component is indicated by a latent 1-of-M binary array $\boldsymbol{\zeta}_n = (\zeta_{n1}, \ldots, \zeta_{nM})$, where $\zeta_{nc} = 1$ indicates that the nth respondent has a probability profile drawn from the cth mixture component, and all other elements are $\zeta_{n,j\neq c} = 0$.

Of course, it is never known exactly from which mixture component the individual parameter vector $\boldsymbol{\xi}_n$ is actually generated. Therefore, the resulting mixture density can be equivalently written as a weighted sum across all mixture components,

$$p(\boldsymbol{\xi}_n \mid \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}) = \sum_{c=1}^{M} \langle \zeta_{nc} \rangle \prod_{d=1}^{D} \sqrt{\frac{\lambda_{cd}}{2\pi}} e^{-\frac{1}{2}(\xi_{nd}-\mu_{cd})^2 \lambda_{cd}}. \tag{A.2}$$

with weights equal to the means $\langle \zeta_{nc} \rangle = E[\zeta_{nc}] \in (0, 1)$ of the estimated categorical distribution of $\boldsymbol{\zeta}_n$.

In the Bayesian framework, all these parameters are again modelled as random variables with weakly informative prior distributions reflecting the prior knowledge and assumptions about the model.

A.2 GMM Parameter Priors

Gauss-gamma priors are defined for all GMM components, as

$$p(\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}) = \prod_{c=1}^{M} p(\boldsymbol{\mu}_c, \boldsymbol{\lambda}_c); \tag{A.3}$$

$$p(\boldsymbol{\mu}_c, \boldsymbol{\lambda}_c) = \prod_{d=1}^{D} p(\mu_{cd} \mid \lambda_{cd}) p(\lambda_{cd}); \tag{A.4}$$

$$p(\mu_{cd} \mid \lambda_{cd}) = \sqrt{\frac{\nu' \lambda_{cd}}{2\pi}} e^{-\frac{1}{2}(\mu_{cd} - m'_{cd})^2 \nu' \lambda_{cd}}; \tag{A.5}$$

$$p(\lambda_{cd}) = \frac{{b'_d}^{a'}}{\Gamma(a')} \lambda_{cd}^{a'-1} e^{-b'_d \lambda_d}. \tag{A.6}$$

Here, $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$ is the gamma function. In the absence of prior information, we assign all $m'_{cd} = 0$. The Jeffreys prior for the mean and precision of a Gaussian distribution would suggest $\nu' \to 0, a' \to 0, b'_d \to 0$. However, to avoid computational indeterminacy causing numerical overflow in case of extreme response patterns, we must use a weakly informative prior.

The effective weight of the prior on the population mean, relative to the weight of one real test participant, is assigned as $\nu' = 0.2$. The gamma shape parameter is assigned $a' = \nu'/2$. For most parameters the inverse scale as $b'_d = \sigma_d^2/2$. These settings conform with the update equations (B.22) and (B.23). The prior probability density for the precision λ_d is then concentrated near zero. This choice prevents the Gaussian components from becoming centered on a single data point with infinite precision, and allows large inter-individual variance. With $a' \leq 1$ the prior expectation of the inter-individual variance is undefined, but the mode is

$$\text{mode}\,[1/\lambda_d \mid a', b'_d] = \frac{b'_d}{a' + 1} \approx \sigma_d^2/2. \tag{A.7}$$

This means that individual deviations from the population mean have a typical prior scale σ_d, i.e., with most deviations in the range $(-\sqrt{2}\sigma_d, +\sqrt{2}\sigma_d)$.

This prior scale is assigned as $\sigma_d = 1$ for all elements related to the logarithmic parameters α_{nt} for the nominal scenario patterns in (4). This prior typically allows a ratio $e^1.4/e^{-1.4} \approx 16$ between occurrence probabilities of the most and least likely scenario.

The same prior scale is also assigned for all elements related to the logarithmic parameters η_{nl} for the ordinal-rating thresholds in (6) and (9). This means we expect similar ratios between the largest and the smallest decision intervals when mapped to the $[0, 1]$ range. As the scale of a standardized logistic distribution is $\sigma = \pi/\sqrt{3} \approx 1.8$, the prior scale is assigned to this value for all elements related to the effect parameters β_{nij} in (9).

A.3 Mixture Weights — Sub-population Models

In case the experiment includes participants from separate sub-populations, e.g., young versus old people, the probability-mass (mixture weight) of profile components might be different for each of the G *sub-populations* from which the separate *subject groups* were recruited.

To allow for separate sub-populations, the prior mixture weights are denoted $\boldsymbol{v}_g = (v_{g1}, \ldots, v_{gc}, \ldots, v_{gM})$, where v_{gc} is the probability that any random individual in the gth sub-population has a probability profile drawn from the cth mixture component. The array of all weight vectors in the model is denoted $\underline{\boldsymbol{v}} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_G)$. Thus, the prior conditional probability that the nth respondent has a response profile from the cth component, given that this subject was recruited from the gth sub-population, is

$$p(\zeta_{nc} = 1 \mid \underline{\boldsymbol{v}}, n \in \mathcal{S}_g) = v_{gc}. \tag{A.8}$$

The prior distribution of mixture weights $\boldsymbol{v}_g$ for the gth sub-population is assigned as a Dirichlet distribution with concentration $\boldsymbol{\gamma}' = (\gamma'_1, \ldots, \gamma'_M)$,

$$p(\boldsymbol{v}_g) = \frac{1}{B(\boldsymbol{\gamma}')} \prod_{c=1}^{M} v_{gc}^{\gamma'_c - 1}, \quad \forall g. \tag{A.9}$$

(The normalizing $B(\ )$ is the multivariate Beta function.) We follow the conventional approach for training Bayesian mixture models and assign fixed small concentration parameters with equal small values for all c in order for the learning to favor sparse solutions. We choose the non-informative Jeffreys prior $\gamma'_c = 0.5$ for all c, but this value is not critical. This prior value plays the role of *pseudo-counts* in the estimation of mixture weights $\boldsymbol{v}_g$ for each group.

Even if none of the respondents in gth group is related to, say, the cth mixture component, the predictive distribution for a random individual in the gth sub-population will in effect be calculated as if 0.5 subject had actually had a response profile related to this component. The use of the Jeffreys prior as pseudo-count is a theoretically well-defined way to account for the

very real possibility that a subject recruited from this sub-population in a future study might actually have a response profile in the cth component, although no such subjects were found in the given data set.

B Model Learning

B.1 Total Log-likelihood

Using (1), (A.1), (A.5), (A.6) and (A.9), the total log-likelihood of all observed data and all model parameters is[6] (omitting irrelevant constants)

$$\begin{aligned}\ln p(\underline{\boldsymbol{x}}, \underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}, \underline{\boldsymbol{v}}) = \text{const.} &+ \sum_{n=0}^{N-1}\sum_{j=1}^{J} x_{nj} f_j(\boldsymbol{\xi}_n) \\ &+ \sum_{n=0}^{N-1}\sum_{c=1}^{M}\zeta_{nc}\sum_{d=1}^{D}\frac{1}{2}\ln\lambda_{cd} - \frac{1}{2}(\xi_{nd}-\mu_{cd})^2\lambda_{cd} \\ &+ \sum_{c=1}^{M}\sum_{d=1}^{D}\frac{1}{2}\ln\lambda_{cd} - \frac{1}{2}(\mu_{cd}-m'_{cd})^2\nu'\lambda_{cd} + (a'-1)\ln\lambda_{cd} - b'_d\lambda_{cd} \\ &+ \sum_{g=1}^{G}\sum_{c=1}^{M}(\gamma'_c-1)\ln v_{gc} + \sum_{n\in\mathcal{S}_g}\zeta_{nc}\ln v_{gc} \quad \text{(B.1)}\end{aligned}$$

Here the first line represent the log-likelihood of observed nominal and ordinal EMA data, given the individual models, the secocnd line represents the log-likelihood of individual model parameters as samples from the population model, and the remaining two lines specify the population model, possibly including sub-populations. The functions $f_j(\ )$ are just a shorthand notation for the log-likelihood of observed data given parameters $\boldsymbol{\xi}_n$, as specified in (1).

B.2 Variational Inference

The model is trained from data with a variant of the standard Variational Inference procedure (Bishop, 2006, Ch.10). A partially factorized density function $q(\ )$ is adapted to be a good approximation of the exact posterior density $p(\ )$ of all model parameters, given the observed data, as

$$p(\underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}, \underline{\boldsymbol{v}} \mid \underline{\boldsymbol{x}}) \approx q(\underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}, \underline{\boldsymbol{v}}) = q(\underline{\boldsymbol{\xi}})q(\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}})q(\underline{\boldsymbol{\zeta}})q(\underline{\boldsymbol{v}}). \quad \text{(B.2)}$$

[6]The summation signs on each line are meant to include all terms on the same line, but not those on other lines.

This variational approximation factorizes the density for the total set of model parameters between the individual parameter distributions $q(\underline{\boldsymbol{\xi}})$, the set of population mixture components $q(\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}})$, the individual component selector variables $q(\underline{\boldsymbol{\zeta}})$, and the component probabilities $q(\underline{\boldsymbol{v}})$.

The VI procedure maximizes a lower bound to the data log-likelihood[7],

$$\mathcal{L}(q) = \left\langle \ln \frac{p(\underline{\boldsymbol{x}}, \underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}, \underline{\boldsymbol{v}})}{q(\underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}, \underline{\boldsymbol{v}})} \right\rangle_{q()} \leq \ln p(\underline{\boldsymbol{x}}) \tag{B.3}$$

and minimizes the Kullback-Leibler divergence

$$\mathrm{KL}\left(q \,\|\, p\right) = \left\langle \ln \frac{q(\underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}, \underline{\boldsymbol{v}})}{p(\underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}, \underline{\boldsymbol{v}} \mid \underline{\boldsymbol{x}})} \right\rangle_{q(\ )} \tag{B.4}$$

between the approximate and exact posterior parameter distributions. The procedure is iterative and theoretically guaranteed to converge.

Since (B.1) includes only a sum of terms across n for the individual parameters, and only a sum across c for the mixture components, while the individual and the population parameters are linked only by the selector parameters $\underline{\boldsymbol{\zeta}}$, the variational distributions are naturally factorized without any further approximation, as

$$q(\underline{\boldsymbol{\xi}}) = \prod_{n=0}^{N-1} q(\boldsymbol{\xi}_n) \tag{B.5}$$

$$q(\underline{\boldsymbol{\zeta}}) = \prod_{n=0}^{N-1} q(\boldsymbol{\zeta}_n) \tag{B.6}$$

$$q(\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}) = \prod_{c=1}^{M} q(\boldsymbol{\mu}_c, \boldsymbol{\lambda}_c) = \prod_{c=1}^{M} \prod_{d=1}^{D} q(\mu_{cd}, \lambda_{cd}) \tag{B.7}$$

$$q(\underline{\boldsymbol{v}}) = \prod_{g=1}^{G} q(\boldsymbol{v}_g) \tag{B.8}$$

Since the prior Gauss-gamma are conjugate distributions for the Gaussian mixture components, the variational $q(\mu_{cd}, \lambda_{cd})$ will naturally get the same Gauss-gamma form as the priors, without any further approximation. Similarly, the variational $q(\boldsymbol{v}_g)$ will also naturally become Dirichlet densities. However, the individual densities $q(\boldsymbol{\xi}_n)$ cannot be expressed as a member of a known distribution family and are therefore approximated by a large number of equally probable sample vectors $\boldsymbol{\xi}_{ns}$, generated by Hamiltonian sampling.

[7]The notation $\langle\ \rangle$ means expectation calculated using the current $q(\ )$ distribution

B.2.1 Individual Component Selector Variables

The joint variational distribution $q(\underline{\boldsymbol{\zeta}})$ is obtained by averaging (B.1) across the distributions of other parameters, as

$$\ln q(\underline{\boldsymbol{\zeta}}) = \left\langle \ln p(\underline{\boldsymbol{x}}, \underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}, \underline{\boldsymbol{v}}) \right\rangle_{q(\underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{v}})} + \text{const.} = \sum_{n=0}^{N-1} \ln q(\boldsymbol{\zeta}_n) \tag{B.9}$$

yielding

$$\begin{aligned} &\ln q(\boldsymbol{\zeta}_n) = \text{const.} + \\ &+ \sum_{c=1}^{M} \zeta_{nc} \underbrace{\left[\langle \ln v_{gc} \rangle + \frac{1}{2} \sum_{d=1}^{D} \langle \ln \lambda_{cd} \rangle - \left\langle (\xi_{nd} - \mu_{cd})^2 \lambda_{cd} \right\rangle \right]}_{\ln \tilde{r}_{nc}}, \quad n \in \mathcal{S}_g \end{aligned} \tag{B.10}$$

This is just the logarithm of a new categorical distribution for $\boldsymbol{\zeta}_n$ with normalized probabilities ("responsibilities") $r_{nc} = p(\zeta_{nc} = 1)$,

$$q(\boldsymbol{\zeta}_n) = \prod_{c=1}^{M} r_{nc}^{\zeta_{nc}}, \quad \text{with } r_{nc} = \langle \zeta_{nc} \rangle = \frac{\tilde{r}_{nc}}{\sum_{i=1}^{M} \tilde{r}_{ni}} \tag{B.11}$$

B.2.2 Mixture Weights in Sub-populations

As (B.1) is a sum of terms involving $\ln v_{gc}$ for each sub-population, the variational mixture-weight distribution is defined by

$$\ln q(\boldsymbol{v}_g) = \text{const.} + \sum_{c=1}^{M} \left(\gamma'_c - 1 + \sum_{n \in \mathcal{S}_g} \langle \zeta_{nm} \rangle \right) \ln v_{gc}. \tag{B.12}$$

Thus, the variational distribution again has the Dirichlet form,

$$q(\boldsymbol{v}_g) \propto \prod_{m=1}^{M} v_{gc}^{\gamma_{gc}-1} \tag{B.13}$$

with concentration parameters $\boldsymbol{\gamma}_g = (\gamma_{g1}, \ldots, \gamma_{gM})$ updated as

$$\gamma_{gc} = \gamma'_c + \sum_{n \in \mathcal{S}_g} \langle \zeta_{nc} \rangle \tag{B.14}$$

B.2.3 Gauss-gamma Mixture Components

As (B.1) is a sum across terms for every element in every mixture component, the variational distributions are defined by

$$\begin{aligned}
&\ln q(\mu_{cd}, \lambda_{cd}) + \text{const.} = \\
&\quad = \frac{1}{2}\ln\lambda_{cd} - \frac{1}{2}(\mu_{cd} - m'_{cd})^2\nu'\lambda_{cd} - \frac{1}{2}\sum_{n=0}^{N-1}\langle\zeta_{nc}\rangle\left\langle(\xi_{nd} - \mu_{cd})^2\right\rangle_\xi \lambda_{cd} \\
&\qquad + (a' - 1)\ln\lambda_{cd} - b'_d\lambda_{cd} + \frac{1}{2}\sum_{n=0}^{N-1}\langle\zeta_{nc}\rangle\ln\lambda_{cd} = \\
&= \frac{1}{2}\ln\lambda_{cd} - \frac{1}{2}\underbrace{\left(\nu' + \textstyle\sum_n \langle\zeta_{nc}\rangle\right)}_{\nu_c}\mu_{cd}^2\lambda_{cd} + \mu_{cd}\underbrace{\left(\nu' m'_{cd} + \textstyle\sum_n \langle\zeta_{nc}\rangle\langle\xi_{nd}\rangle\right)}_{\nu_c m_{cd}}\lambda_{cd} \\
&\qquad - \frac{1}{2}{m'_{cd}}^2\nu'\lambda_{cd} - \frac{1}{2}\sum_{n=0}^{N-1}\langle\zeta_{nc}\rangle\left\langle\xi_{nd}^2\right\rangle\lambda_{cd} \\
&\qquad + (a' - 1)\ln\lambda_{cd} - b'_d\lambda_{cd} + \frac{1}{2}\sum_{n=0}^{N-1}\langle\zeta_{nc}\rangle\ln\lambda_{cd} \qquad \text{(B.15)}
\end{aligned}$$

As this is a second-degree polynomial in μ_{cd}, this part can be written as the logarithm of a Gaussian density

$$\ln q(\mu_{cd} \mid \lambda_{cd}) + \text{const.} = \frac{1}{2}\ln\lambda_{cd} - \frac{1}{2}(\mu_{cd} - m_{cd})^2\nu_c\lambda_{cd} \qquad \text{(B.16)}$$

Thus, the variational density is again a conditional Gaussian

$$q(\mu_{cd} \mid \lambda_{cd}) = \sqrt{\frac{\nu_c\lambda_{cd}}{2\pi}}e^{-\frac{1}{2}(\mu_{cd} - m_{cd})^2\nu_c\lambda_{cd}} \qquad \text{(B.17)}$$

with parameters updated as

$$\nu_c = \nu' + \sum_{n=0}^{N-1}\langle\zeta_{nc}\rangle \qquad \text{(B.18)}$$

$$m_{cd} = \frac{1}{\nu_c}\left(\nu' m'_{cd} + \sum_{n=0}^{N-1}\langle\zeta_{nc}\rangle\langle\xi_{nd}\rangle\right) \qquad \text{(B.19)}$$

Similarly, the variational density for the precision parameters is defined by

$$\begin{aligned}\ln q(\lambda_{cd}) &= \ln q(\mu_{cd}, \lambda_{cd}) - \ln q(\mu_{cd} \mid \lambda_{cd}) = \\ &= \text{const.} + \left(a' - 1 + \frac{1}{2}\textstyle\sum_{n=0}^{N-1} \langle \zeta_{nc} \rangle\right) \ln \lambda_{cd} \\ &- \left(b'_d + \frac{1}{2} {m'_{cd}}^2 \nu' - \frac{1}{2}\nu_c m_{cd}^2 + \frac{1}{2}\sum_{n=0}^{N-1} \langle \zeta_{nc} \rangle \left\langle \xi_{nd}^2 \right\rangle\right) \lambda_{cd} \end{aligned} \tag{B.20}$$

This is again the logarithm of a gamma density

$$q(\lambda_{cd}) \propto \lambda_{cd}^{a_c - 1} e^{-b_{cd}\lambda_{cd}} \tag{B.21}$$

with parameters updated as

$$a_c = a' + \frac{1}{2}\sum_{n=0}^{N-1} \langle \zeta_{nc} \rangle \tag{B.22}$$

$$b_{cd} = b'_d + \frac{1}{2}\nu' {m'_{cd}}^2 - \frac{1}{2}\nu_c m_{cd}^2 + \frac{1}{2}\sum_{n=0}^{N-1} \langle \zeta_{nc} \rangle \left\langle \xi_{nd}^2 \right\rangle \tag{B.23}$$

B.2.4 Individual Parameters

As (B.1) is a sum of terms for the individual parameters $\boldsymbol{\xi}_n$, the variational distributions are defined by the log-likelihood

$$\begin{aligned}\ln q(\boldsymbol{\xi}_n) &= \text{const.} + \left\langle \ln p(\underline{\boldsymbol{x}}, \underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}, \underline{\boldsymbol{v}}) \right\rangle_{q(\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}, \underline{\boldsymbol{v}})} = \\ &= \sum_{j=1}^{J} x_{nj} f_j(\boldsymbol{\xi}_n) \\ &+ \sum_{c=1}^{M} \langle \zeta_{nc} \rangle \sum_{d=1}^{D} \frac{1}{2} \langle \ln \lambda_{cd} \rangle - \frac{1}{2} \left\langle (\xi_{nd} - \mu_{cd})^2 \lambda_{cd} \right\rangle_{\mu_{cd}, \lambda_{cd}} \end{aligned} \tag{B.24}$$

Because of the non-linear functions f_j, this log-likelihood can not be associated with a known distribution family. The variational distribution is instead approximated by a large number of samples drawn from $q(\boldsymbol{\xi}_n)$ by Hamiltonian sampling. This sampling procedure uses only the known log-likelihood function (B.24) and its gradient with regard to $\boldsymbol{\xi}_n$ which is defined by the known gradients of $f_j(\boldsymbol{\xi}_n)$. The sampling is done separately for each participant model, and must be repeated for each iteration of the variational procedure to account for the current estimate of the population model.

B.3 Log-likelihood Lower Bound

To monitor the progress of variational learning the variational lower bound (B.3) is most conveniently calculated at each iteration as

$$\begin{aligned}\mathcal{L}(q) = \left\langle \ln \frac{p(\underline{\boldsymbol{x}}, \underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}, \underline{\boldsymbol{v}})}{q(\underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}, \underline{\boldsymbol{v}})} \right\rangle_{q()} = \\ = \sum_n \langle \ln p(\boldsymbol{x}_n \mid \boldsymbol{\xi}_n) \rangle_\xi + \left\langle \ln p(\boldsymbol{\xi}_n \mid \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}}, \underline{\boldsymbol{\zeta}}) \right\rangle_q - \langle \ln q(\boldsymbol{\xi}_n) \rangle_q \\ - \left\langle \ln \frac{q(\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}})}{p(\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\lambda}})} \right\rangle_q - \left\langle \ln \frac{q(\underline{\boldsymbol{\zeta}})}{p(\underline{\boldsymbol{\zeta}} \mid \underline{\boldsymbol{v}})} \right\rangle_q - \left\langle \ln \frac{q(\underline{\boldsymbol{v}})}{p(\underline{\boldsymbol{v}})} \right\rangle_q \quad \text{(B.25)}\end{aligned}$$

This lower bound is theoretically guaranteed to be non-decreasing for each step of the learning procedure. Here the first two terms were already calculated during sampling by (B.24). The third term is the sum of entropy for each $\boldsymbol{\xi}_n$, which is calculated from the samples using a nearest-neighbour ("Kozachenko-Leonenko") estimator (Singh and Poczos, 2016). The last three terms subtract the Kullback-Leibler divergence $\mathrm{KL}(q \,\|\, p)$ between posterior and prior distributions for the three types of population parameters.

The subtraction of Kullback-Leibler divergences represent the cost of model complexity which is the basis of the *Occam's Razor* effect. The variational learning tends to push parameter distributions toward the priors for any mixture component that is not really needed to model the observed data, because this reduces the Kullback-Leibler divergence and increases $\mathcal{L}(q)$.

C Predictive Results

The learned individual and population models are used to calculate two predictive distributions:

C.1 (Sub-)Population Mean

The predictive distribution of the *mean* vector $\boldsymbol{\mu}_g = (\mu_{g1}, \ldots, \mu_{gD})$ (equal to the median) for the gth *sub-population* is a mixture density, based on (B.17) and (B.13), integrated over the variational distributions (B.21) of the precision parameters.

$$p(\boldsymbol{\mu}_g) = \sum_{c=1}^{M} \langle v_{gc} \rangle \prod_{d=1}^{D} \int q(\mu_{cd} \mid \lambda_{cd}) q(\lambda_{cd}) \, d\lambda_{cd}$$
$$\propto \sum_{c=1}^{M} \langle v_{gc} \rangle \prod_{d=1}^{D} \left(1 + \frac{(\xi_{Nd} - m_{cd})^2 \nu_c}{2 b_{cd}} \right)^{-\frac{2a_c+1}{2}} \quad \text{(C.1)}$$

The resulting marginal density for the mean is a mixture including a univariate Student-t distribution for each element of each mixture component, with location m_{cd}, scale $\sqrt{b_{cd}/a_c\nu_c}$, and degrees-of-freedom $2a_c$.

C.2 Random Individual

The predictive distribution of parameters $\boldsymbol{\xi}_{Ng} = (\ldots, \xi_{Ngd}, \ldots)$ for a future (Nth) unknown individual randomly drawn from the gth *sub-population*, from which the gth test group was recruited, is a mixture density, based on (A.2) and (B.13) for the gth sub-population, integrated over the learned variational distributions for the Gaussian mean and precision parameters:

$$p(\boldsymbol{\xi}_{Ng}) = \sum_{c=1}^{M} \langle v_{gc} \rangle \prod_{d=1}^{D} \int \int \sqrt{\frac{\lambda_{cd}}{2\pi}} e^{-\frac{1}{2}(\xi_{nd} - \mu_{cd})^2 \lambda_{cd}} q(\mu_{cd} \mid \lambda_{cd}) q(\lambda_{cd}) \, d\mu_{cd} \, d\lambda_{cd}$$
$$\propto \sum_{c=1}^{M} \langle v_{gc} \rangle \prod_{d=1}^{D} \left(1 + \frac{(\xi_{Nd} - m_{cd})^2 \nu_c}{2 b_{cd}(\nu_c + 1)} \right)^{-\frac{2a_c+1}{2}} \quad \text{(C.2)}$$

Using the learned variational distributions (B.17) and (B.21) for the Gaussian component mean and precision, the result is a univariate Student-t distribution for each element of each mixture component, with location m_{cd}, scale $\sqrt{b_{cd}(\nu_c + 1)/a_c\nu_c}$, and degrees-of-freedom $df = 2a_c$.

Proof: For any c, d the inner integral over μ_{cd} can be evaluated as (indices omitted)

$$\int p(\xi \mid \mu, \lambda) q(\mu \mid \lambda) \, d\mu = \sqrt{\frac{\lambda}{2\pi}} \sqrt{\frac{\nu\lambda}{2\pi}} \int e^{-\frac{1}{2}\left[(\xi-\mu)^2\lambda + (\mu-m)^2\nu\lambda\right]} \, d\mu =$$
$$= \sqrt{\frac{\lambda}{2\pi}} \sqrt{\frac{\nu\lambda}{2\pi}} \int e^{-\frac{1}{2}\left[(\xi-m)^2\lambda - 2(\xi-m)(\mu-m)\lambda + (\mu-m)^2\lambda + (\mu-m)^2\nu\lambda\right]} \, d\mu = [\mu - m \to z]$$
$$= \sqrt{\frac{\lambda}{2\pi}} \sqrt{\frac{\nu\lambda}{2\pi}} \int e^{-\frac{1}{2}\left[(\xi-m)^2\lambda - 2z(\xi-m)\lambda + z^2(\nu+1)\lambda\right]} \, dz \quad \text{(C.3)}$$

Noting that the exponent is just a second-degree polynomial in μ, we complete the square using the central location $\bar{z} = (\xi - m)/(\nu + 1)$ defined by the mixed term $2z(\xi - m)\lambda = 2\bar{z}(\nu + 1)\lambda$. Then the integral can be simplified as

$$\begin{aligned}
\int p(\xi \mid \mu, \lambda) q(\mu \mid \lambda)\, d\mu &= \\
&= \sqrt{\frac{\lambda}{2\pi}} \sqrt{\frac{\nu\lambda}{2\pi}} e^{-\frac{1}{2}\left[(\xi-m)^2\lambda - \bar{z}^2(\nu+1)\lambda\right]} \int e^{-\frac{1}{2}(z-\bar{z})^2(\nu+1)\lambda}\, dz = \\
&= \sqrt{\frac{\lambda}{2\pi}} \sqrt{\frac{\nu\lambda}{2\pi}} e^{-\frac{1}{2}\left[(\xi-m)^2\lambda - (\xi-m)^2\lambda/(\nu+1)\right]} \sqrt{\frac{2\pi}{(\nu+1)\lambda}} = \\
&= \sqrt{\frac{\lambda}{2\pi}} \sqrt{\frac{\nu}{\nu+1}} e^{-\frac{1}{2}\frac{(\xi-m)^2\nu}{\nu+1}\lambda} \qquad \text{(C.4)}
\end{aligned}$$

Finally, the outer integral over λ is evaluated using the gamma density $q(\lambda)$ as

$$\begin{aligned}
\int \sqrt{\frac{\lambda}{2\pi}} \sqrt{\frac{\nu}{\nu+1}} e^{-\frac{1}{2}\frac{(\xi-m)^2\nu}{\nu+1}\lambda} q(\lambda)\, d\lambda &= \\
&= \int \sqrt{\frac{\lambda}{2\pi}} \sqrt{\frac{\nu}{\nu+1}} e^{-\frac{1}{2}\frac{(\xi-m)^2\nu}{\nu+1}\lambda} \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda)}\, d\lambda \\
&\propto \int \lambda^{a+\frac{1}{2}-1} e^{-\left[b + \frac{1}{2}\frac{(\xi-m)^2\nu}{\nu+1}\right]\lambda}\, d\lambda \\
&\propto \left(b + \frac{1}{2}\frac{(\xi-m)^2\nu}{\nu+1}\right)^{-\frac{2a+1}{2}} \qquad \text{(C.5)}
\end{aligned}$$

which concludes the derivation of the Student-t factors in (C.2).