TI 2016-075/III
Tinbergen Institute Discussion Paper

Flexible Mixture-Amount Models for Business and Industry using Gaussian Processes

Aiste Ruseckaite[1]
Dennis Fok[1]
Peter Goos[2]

[1] *Erasmus School of Economics, Erasmus University Rotterdam, and Tinbergen Institute, the Netherlands;*
[2] *KU Leuven, Belgium.*

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Flexible Mixture-Amount Models for Business and Industry Using Gaussian Processes

Aiste Ruseckaite[1,4], Dennis Fok[1,4] and Peter Goos[1,2,3]

[1]Erasmus School of Economics, Erasmus University Rotterdam, the Netherlands
[2]Faculty of Bioscience Engineering, KU Leuven, Belgium
[3]Faculty of Applied Economics & StatUa Center for Statistics, Universiteit Antwerpen, Belgium
[4]Tinbergen Institute, the Netherlands

September 9, 2016

Abstract

Many products and services can be described as mixtures of ingredients whose proportions sum to one. Specialized models have been developed for linking the mixture proportions to outcome variables, such as preference, quality and liking. In many scenarios, only the mixture proportions matter for the outcome variable. In such cases, mixture models suffice. In other scenarios, the total amount of the mixture matters as well. In these cases, one needs mixture-*amount* models. As an example, consider advertisers who have to decide on the advertising media mix (e.g. 30% of the expenditures on TV advertising, 10% on radio and 60% on online advertising) as well as on the total budget of the entire campaign. To model mixture-amount data, the current strategy is to express the response in terms of the mixture proportions and specify mixture parameters as parametric functions of the amount. However, specifying the functional form for these parameters may not be straightforward, and using a flexible functional form usually comes at the cost of a large number of parameters.

In this paper, we present a new modeling approach which is flexible but parsimonious in the number of parameters. The model is based on so-called Gaussian processes and avoids the necessity to a-priori specify the shape of the dependence of the mixture parameters on the amount. We show that our model encompasses two commonly used model specifications as extreme cases. Finally, we demonstrate the model's added value when compared to standard models for mixture-amount data. We consider two applications. The first one deals with the reaction of mice to mixtures of hormones applied in different amounts. The second one concerns the recognition of advertising campaigns. The mixture here is the particular media mix (TV and magazine advertising) used for a campaign. As the total amount variable, we consider the total advertising campaign exposure.

Keywords: Gaussian process prior; Nonparametric Bayes; Advertising mix; Ingredient proportions; Mixtures of ingredients

1 Introduction

Many products and services can be described as mixtures of ingredients. Examples are mixtures of different fruits composing a fruit salad (e.g. 50% of apples, 30% of wild berries and 20% of grapes) or the mixture of different transportation modes used by an individual on a particular trip (e.g. 70% of travel time by metro and 30% by bike). In marketing, advertisers have to decide on the advertising media mix (e.g. 30% of the expenditures on TV advertising, 10% on radio and 60% on the Internet). As another example, hormone mixture treatments are of interest in biological research. In general, the response to such a product, service, media mix or treatment depends on the proportions of the individual ingredients. To explain such responses, specialized models are necessary to account for the fact that the proportions sum to one (Cornell, 2002).

In many cases, some other quantitative variable describing each mixture may also be relevant, both to the effect of individual ingredients on the response and to the response itself. In the marketing example, advertisers decide on the advertising media mix as well as on the *total budget* of the entire campaign. The total advertising budget will of course affect the impact of the campaign. Additionally, it is likely that the total budget also affects the impact of a particular advertising medium. In a transportation setting, the attractiveness of a trip depends on the mix of transportation modes but also on the *total travel time*. However, the total travel time can affect the sensitivity of the attractiveness to particular transportation modes as well. Finally, the choice of a salad is affected by both its ingredients and the price. At the same time, the price may have an impact on how important different ingredients composing the salad are.

In general, a quantitative variable often impacts not only the response but also the effect of each ingredient in a mixture. Although this quantitative variable does not always correspond to a true amount, for simplicity, we will refer to this variable as the amount variable. Models that simultaneously link mixture proportions and amount variables to response variables are called mixture-amount models (Cornell, 2002; Piepel and Cornell, 1985).

If the total amount of a mixture affects the impact of mixture proportions, the parameters corresponding to the mixture ingredients in a model need to vary with the amount. For this reason, mixture-amount models typically express the mixture parameters as a parametric function of the amount. The effect of the amount on the response is then captured through its effect on the mixture parameters (Piepel and Cornell, 1985). However, such models require the specification of a functional form relating the mixture parameters to the amount variable a priori. Correctly specifying such a function may not be straightforward. Some flexible functional forms are available, see Piepel and Cornell (1985). However, the number of parameters in these specifications is usually very large. This prevents the use of the resulting models in practice, as these models are usually fitted to experimental data, the sample size of which tends to be small.

In this paper, we introduce an alternative approach which is parsimonious in the number of parameters as well as flexible. Our approach is based on so-called Gaussian processes (Rasmussen and Williams, 2005) and avoids the necessity to specify the shape of the functional form of the relationship between the amount variable and the mixture parameters. We only use a smoothness assumption, meaning that, for similar values of the amount, we expect the mixture parameters to be similar as well. The degree of smoothness is captured by a parameter that can also be estimated if sufficient data are available. Another way to interpret our model is that we treat mixture parameters as functions of the amount and that we specify a prior distribution directly over these functions.

In technical terms, we specify a separate parameter vector for every unique observed amount value. One such parameter vector describes the impact of the mixture components on the response at a specific amount. These parameter vectors are, however, not independent across amounts. As explained above, the model incorporates the idea that, for amount values that are close to each other, the model parameters are expected to be rather similar. The Gaussian process formalizes this by specifying the correlation between all parameter vectors. The correlation structure itself is governed by the so-called Gaussian kernel, which is described by a single parameter. This parameter specifies the dependence of the correlations on the amount differences and, therefore, controls for the smoothness of the mixture parameters as a function of the amount. If one sets this parameter to zero or to a large positive value, one can obtain existing models as special cases. When the parameter equals zero, the correlations approach zero and one obtains different and independent mixture parameters for each unique amount value. Such a model has been considered by Piepel and Cornell (1985). When the parameter approaches infinity, the correlations tend to one and one obtains a single vector for the mixture parameters such that the amount variable does not play a role. In this case, we are left with a standard mixture model, as, for example, used in Sahrmann et al. (1987).

Finally, apart from the correlations across amounts, the mixture parameters of the model at a given amount value might also be correlated. For instance, the impacts of radio advertising and TV advertising may move up and down together as one considers different advertising amounts. In our model, we also allow for this type of correlation. As a result, the overall variance-covariance structure of the mixture parameters depends on a parameter that controls the correlation across amounts and a parameter that controls the correlation across individual parameters at a given amount. If the latter correlation approaches one, we obtain another special case of our model in which the amount has a separate, additive impact on the response variable.

We demonstrate that our approach naturally leads to a model specification in which the mixture parameters follow a matrix normal distribution with the variance-covariance matrix consisting

of two parts. The parameters of the resulting model can be estimated using Bayesian techniques. In this paper, we also provide the details of the required sampling procedures.

To illustrate our approach, we present two examples. The first example concerns the reaction of mice to different mixtures of hormones administered at different amount levels. The second illustration considers the recognition of advertising campaigns for skin and hair care products. The mixture here is a particular media mix used for a campaign. The amount variable is the total advertising campaign exposure. We introduce both examples in more detail in the subsequent sections.

The remainder of this paper is organized as follows. In the next section, we review the literature on mixture-amount models and Gaussian processes. Section 3 introduces our new approach to model mixture-amount data. Section 4 presents our Bayesian estimation procedure. In Section 5, we illustrate the new modeling approach. We end the paper with a discussion in Section 6.

2 Literature

In this section, we first review the existing literature on mixture-amount models. Next, we discuss Gaussian processes which we use to develop our new models for mixture-amount data.

2.1 Mixture-amount models

When a response variable is modeled as a function of proportions of ingredients in a mixture, the *mixture constraint*, defined by $\sum_{i=1}^{q} x_i = 1$, has a significant impact on the models that can be fitted. Here, x_i is the proportion of ingredient i and q is the number of ingredients in the mixture. The first consequence is that a linear regression model for mixture data cannot contain an intercept. Furthermore, cross-products $x_i x_j$ and squares x_i^2 cannot be simultaneously included as regressors in the model, since this leads to perfect collinearity. To deal with these issues, Scheffé (1958, 1963) proposed a family of models that are suitable for modeling mixture data. The first-order (linear) and second-order Scheffé models, respectively, for a continuous dependent variable y are defined as

$$y = \sum_{i=1}^{q} \beta_i x_i + \varepsilon \tag{1}$$

and

$$y = \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j}^{q} \sum \beta_{ij} x_i x_j + \varepsilon, \tag{2}$$

where ε indicates the error term.

The models in Equations (1)-(2) can be used if the total amount is fixed or does not affect the response. However, they are not suitable if the amount of a mixture affects the response.

Piepel and Cornell (1985) introduced mixture-amount models to deal with situations in which the response depends on the total amount of a mixture as well as on the ingredient proportions. They recognized the similarity of a mixture-amount experiment to a mixture experiment with one process variable (the amount in this case) and adapted models developed by Scheffé (1963) for mixture experiments with process variables.

Following Piepel and Cornell (1985), assume that we have acquired mixture data at r different values of the amount variable A, denoted by $A_1, A_2, \ldots, A_r$ ($r \geq 2$), and that the relation between the response and the ingredient proportions is modeled by a Scheffé model with p mixture parameters, $\beta_1, \beta_2, \ldots, \beta_p$. If the total amount of the mixture affects the impact of mixture proportions, the parameters corresponding to the mixture ingredients in a model need to vary with A. Thus, each mixture parameter β_m, $m = 1, \ldots, p$, has to depend on the total amount. Using this reasoning, one can create a *mixture-amount model* from the assumed Scheffé model by allowing the mixture parameters β_m to be a function $\beta_m(A)$ of the amount A, for $m = 1, \ldots, p$. One possible parametric model for the dependence of the mixture parameters on the amount is the polynomial function,

$$\beta_m(A) = \beta_m^0 + \sum_{k=1}^{K} \beta_m^k A^k. \tag{3}$$

The parameter β_m^k represents the k^{th} order effect of the amount on β_m.

As an example, we present a model for mixture-amount data for $q = 2$ ingredients based on the second-order Scheffé model given in Equation (2) and using the expression in Equation (3) with $K = 2$ to write the mixture parameters as a function of the amount:

$$\begin{aligned} y &= \beta_1(A)x_1 + \beta_2(A)x_2 + \beta_3(A)x_1x_2 + \varepsilon \\ &= \beta_1^0 x_1 + \beta_2^0 x_2 + \beta_3^0 x_1x_2 + \sum_{k=1}^{2} (\beta_1^k x_1 + \beta_2^k x_2 + \beta_3^k x_1x_2)A^k + \varepsilon. \end{aligned} \tag{4}$$

This model contains first- and second-order effects of the mixture components and linear and quadratic effects of the amount variable. The terms in the mixture-amount model in Equation (4) have the following interpretation:

- if the amount variable A is centered around zero, $\beta_1^0 x_1 + \beta_2^0 x_2 + \beta_{12}^0 x_1x_2$ represents the linear and nonlinear blending properties of the mixture components at the average value of the total amount;
- $(\beta_1^1 x_1 + \beta_2^1 x_2 + \beta_{12}^1 x_1x_2)A$ represents the linear effect of the total amount on the linear and nonlinear blending properties of the mixture components;
- $(\beta_1^2 x_1 + \beta_2^2 x_2 + \beta_{12}^2 x_1x_2)A^2$ represents the quadratic effect of the total amount on the linear

and nonlinear blending properties of the mixture components.

In general, the parameters β_i^k and β_{ij}^k of the terms involving $x_i A^k$ and $x_i x_j A^k$ $(k = 1, 2)$ in Equation (4) are measures of the effect of changing the total amount of the mixture on the linear and nonlinear blending properties of the mixture ingredients. For general q and k, we have

$$y = \sum_{i=1}^{q} \beta_i^0 x_i + \sum\sum_{i<j}^{q} \beta_{ij}^0 x_i x_j + \sum_{k=1}^{K} \left[\sum_{i=1}^{q} \beta_i^k x_i + \sum\sum_{i<j}^{q} \beta_{ij}^k x_i x_j \right] A^k + \varepsilon. \tag{5}$$

To emphasize the fact that the mixture parameters are assumed to be some parametric functions of the amount, we call the models above parametric. When the amount of a mixture does not affect the blending properties of the mixture components but only causes a constant change in the magnitude of the response (that is, all β_i^k are equal and all $\beta_{ij}^k = 0$), Equation (5) reduces to

$$y = \sum_{i=1}^{q} \beta_i^0 x_i + \sum\sum_{i<j}^{q} \beta_{ij}^0 x_i x_j + \sum_{k=1}^{K} \beta_0^k A^k + \varepsilon, \tag{6}$$

where $\beta_0^k = \beta_1^k = \cdots = \beta_q^k$. In this case, the amount does not affect the impact of the proportions on the response, but it has a direct impact itself.

The models specified above are typically used for mixture-amount data. However, there are a number of issues with them. First, the number of parameters in the final model grows rapidly with q and K. Furthermore, K has to be specified a-priori, which is not always easy to do. Third, using a large value for K may yield highly volatile functions $\beta_m(A)$. Finally, in addition to polynomial functions, there is a wide variety of other specifications that one may want to consider. To avoid all these issues, in this paper, we introduce a non-parametric specification for $\beta_m(A)$ based on Gaussian processes. This approach does not require an a-priori selection of the shape of the functions $\beta_m(A)$.

Below, we first discuss Gaussian processes in general. In Section 3, we incorporate the Gaussian process in the mixture-amount model.

2.2 Gaussian processes

A Gaussian process (GP) defines a distribution over functions. Denote such a distribution by $P(f)$ for some function f, $f : \chi \to \mathbb{R}$. Then, $P(f)$ is a Gaussian process if for any finite subset of χ, the marginal distribution over that finite subset has a multivariate Gaussian distribution (Bishop, 2006; Rasmussen and Williams, 2005). We can therefore write $f(\mathbf{x}) \sim \mathcal{N}\big(m(\mathbf{x}), \mathbf{\Omega}(\mathbf{x}, \mathbf{x})\big)$, $\mathbf{x} \subset \chi$, for a mean function $m(\mathbf{x})$ and covariance function $\mathbf{\Omega}(\mathbf{x}, \mathbf{x})$. As a result, a Gaussian process is parameterized by its mean and covariance functions. Note that f can be infinite-dimensional and

therefore Gaussian processes extend multivariate Gaussian distributions to infinite dimensionality.

After some mean is assumed for $f(\mathbf{x})$, the covariance function $\mathbf{\Omega}(\mathbf{x}, \mathbf{x})$ completely defines the behavior of $f(\mathbf{x})$ for different values of $\mathbf{x}$. The function $\mathbf{\Omega}(\mathbf{x}, \mathbf{x})$ parameterizes our beliefs about the smoothness of $f(\mathbf{x})$ with respect to $\mathbf{x}$. Different $\mathbf{\Omega}(\mathbf{x}, \mathbf{x})$ functions could represent many different kinds of nonlinearity and lead to different shapes of $f(\mathbf{x})$ (Rasmussen and Williams (2005), see also Duvenaud et al. (2013); Salimans (2012); Wilson and Adams (2013)). In general, any real-valued function $\mathbf{\Omega}(\mathbf{x}, \mathbf{x})$ is acceptable to describe a covariance function provided the resulting covariance matrix is positive semi-definite.

By estimating the parameters defining the mean and covariance functions of f, we in fact acquire knowledge concerning the distribution of f. Note that, in this process, we do not assume any parametric form for the function f itself. Prior beliefs about the structure of the function f can be incorporated by choosing a particular covariance function. As a result, Gaussian processes are very flexible and can be used to represent many different regression models that would have an infinite number of parameters if formulated in a conventional manner (Neal, 1999).

Prediction for Gaussian processes is easy if the mean and covariance functions are known. Suppose that we already know the function's values at $\mathbf{x}$ and wish to predict the function's value at a new observation $\mathbf{x}^*$, i.e., $f(\mathbf{x}^*)$. Recall that for any function f drawn from a Gaussian process prior with the mean and covariance functions given by $m(\cdot)$ and $\mathbf{\Omega}(\cdot, \cdot)$, respectively, the marginal distribution over any finite subset of χ is multivariate Gaussian. Therefore, the joint distribution of f at the observed data $\mathbf{x}$ and at the new data point $\mathbf{x}^*$ can be written as

$$\begin{pmatrix} f(\mathbf{x}) \\ f(\mathbf{x}^*) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(\mathbf{x}) \\ m(\mathbf{x}^*) \end{pmatrix}, \begin{pmatrix} \mathbf{\Omega}(\mathbf{x}, \mathbf{x}) & \mathbf{\Omega}(\mathbf{x}, \mathbf{x}^*) \\ \mathbf{\Omega}(\mathbf{x}^*, \mathbf{x}) & \mathbf{\Omega}(\mathbf{x}^*, \mathbf{x}^*) \end{pmatrix} \right),$$

where $m(\cdot)$ and $\mathbf{\Omega}(\cdot, \cdot)$ denote the mean and covariance functions evaluated at either the observed data $\mathbf{x}$ or at the new data $\mathbf{x}^*$. Conditioning the joint Gaussian prior distribution on the observations gives

$$\begin{aligned} f(\mathbf{x}^*)|\mathbf{x}, \mathbf{x}^*, f(\mathbf{x}) \sim \mathcal{N}\big(& m(\mathbf{x}^*) + \mathbf{\Omega}(\mathbf{x}^*, \mathbf{x})\mathbf{\Omega}(\mathbf{x}, \mathbf{x})^{-1}(f(\mathbf{x}) - m(\mathbf{x})), \\ & \mathbf{\Omega}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{\Omega}(\mathbf{x}^*, \mathbf{x})\mathbf{\Omega}(\mathbf{x}, \mathbf{x})^{-1}\mathbf{\Omega}(\mathbf{x}, \mathbf{x}^*)\big), \end{aligned} \tag{7}$$

which is the posterior predictive distribution of $f(\mathbf{x}^*)$ for any input $\mathbf{x}^*$. Function values $f(\mathbf{x}^*)$ can be sampled from the joint posterior distribution by evaluating the mean and covariance functions in Equation (7) (Rasmussen and Williams, 2005).

Gaussian processes are conceptually simple and flexible, and they often exhibit a good performance in various applications. Thus, it is not surprising that they are widespread in many

different areas ranging from simple regressions and classifications (Gattiker et al., 2015; Neal, 1997, 1999; Williams, 1999) or multi-task learning (Bonilla et al., 2007; Boyle and Frean, 2005; Melkumyan and Ramos, 2011) to visualisation of high dimensional data (Lawrence, 2004), density estimation (Leonard, 1978; Riihimäki and Vehtari, 2014) or human motion modeling (Wang et al., 2008). However, Gaussian processes have hitherto not been used in the context of mixture-amount models.

3 Model

3.1 Derivation

As discussed above, a straightforward way to model mixture-amount data is to specify the dependence of mixture parameters on the amount explicitly, like in Equation (3). In this section, we present an elegant way to model $\beta_m,\ m = 1, \ldots, p$, as a function of the amount A using Gaussian processes (Rasmussen and Williams, 2005), where we do not explicitly assume any functional form.

We denote the set of observed amount values in the data as $\vec{A} = (A_1, A_2, \ldots, A_r)'$, $r \leq N$, where N is the total number of observations. The latent function linking the mixture parameters β_m to the amount is given by $\beta_m(A)$. Our approach specifies a (prior) distribution directly over these functions, where the correlation structure in $\mathbf{\Omega}(\cdot, \cdot)$ is specified using only one positive parameter (τ). This parameter determines how quickly the mixture parameters vary with respect to the amount.

Formally, we collect the parameters for all observed amount values in the parameter matrix $\mathbf{B}(\vec{A})$, which contains the p ingredient's and their interactions' effects at different amount values in its rows and different ingredient's and their interactions' effects at a given value of the amount in its columns, i.e.,

$$\mathbf{B}(\vec{A}) = \begin{pmatrix} \beta_1(A_1) & \beta_2(A_1) & \ldots & \beta_p(A_1) \\ \beta_1(A_2) & \beta_2(A_2) & \ldots & \beta_p(A_2) \\ \vdots & \vdots & \vdots & \vdots \\ \beta_1(A_r) & \beta_2(A_r) & \ldots & \beta_p(A_r) \end{pmatrix} = \begin{pmatrix} \beta_1(\vec{A})' \\ \beta_2(\vec{A})' \\ \vdots \\ \beta_p(\vec{A})' \end{pmatrix}', \tag{8}$$

with $\beta_m(\vec{A}) = \big(\beta_m(A_1), \beta_m(A_2), \ldots, \beta_m(A_r)\big)'$.

If we assume that the response variable is continuous and consider a linear Scheffé model, then

$p = q$ and the response y_i of an observation i can be modeled as

$$y_i = \big(\mathbf{a}_i \mathbf{B}(\vec{A})\big) \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{qi} \end{pmatrix} + \varepsilon_i, \tag{9}$$

where $\mathbf{a}_i$ is a $1 \times r$ row vector indicating which of the amount values corresponds to observation i. The j^{th} element of $\mathbf{a}_i$ is one if the j^{th} amount is used for observation i and zero otherwise. The row vector $\mathbf{a}_i$ selects the appropriate parameters from $\mathbf{B}(\vec{A})$. Using some linear algebra, we can rewrite the model for y_i as

$$y_i = \big(\mathbf{x}_i' \otimes \mathbf{a}_i\big)\, \text{vec}\big(\mathbf{B}(\vec{A})\big) + \varepsilon_i, \tag{10}$$

where $\otimes$ is the Kronecker product, vec$(\cdot)$ denotes the vectorization operator and $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{qi})'$. Stacking all response values gives

$$\mathbf{y} = \mathbf{X}\text{vec}\big(\mathbf{B}(\vec{A})\big) + \boldsymbol{\varepsilon}, \tag{11}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$, $\mathbf{X} = (X_1', X_2', \ldots, X_N')'$ with $X_i = \mathbf{x}_i' \otimes \mathbf{a}_i$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)'$.

The model in Equation (11) resembles a standard linear regression model. The only difference is that we treat the parameter vector vec$\big(\mathbf{B}(\vec{A})\big)$ as a function of the (observed) amount values. To complete the model, we assume that the prior on the parameters $\beta_m(A)$ is a Gaussian process with mean b_m and covariance function $\boldsymbol{\Omega}(\cdot, \cdot)$. We also allow the Gaussian processes to be correlated across m, that is, we allow for correlation between the different mixture ingredient parameters. At a given amount level, the variance-covariance matrix of the p mixture parameters is given by $\sigma^2 \boldsymbol{\Phi}$.

As a result, the parameter matrix at the observed amounts, $\mathbf{B}(\vec{A})$, follows a matrix-normal distribution, that is, $\mathbf{B}(\vec{A})|\sigma^2 \sim \mathcal{MN}\big(\bar{\mathbf{B}}, \boldsymbol{\Omega}, \sigma^2 \boldsymbol{\Phi}\big)$, where

$$\bar{\mathbf{B}} = (\mathbf{1}_{r \times 1} \otimes \mathbf{b}'), \tag{12}$$

with $\mathbf{1}_{r \times 1}$ being a vector of ones of length r, $\mathbf{b} = (b_1, \ldots, b_p)'$ and $\boldsymbol{\Omega}$ denoting a covariance matrix with elements $\boldsymbol{\Omega}(A', A'')$, $\forall A', A'' \in \vec{A}$.

Summarizing all equations and allowing for d additional covariates in matrix $\mathbf{X}_2$, the final

model becomes

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\text{vec}\big(\mathbf{B}(\vec{A})\big) + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \\ \text{vec}\big(\mathbf{B}(\vec{A})\big)|\tau, \mathbf{b}, \boldsymbol{\Phi}, \sigma^2 &\sim \mathcal{N}\big(\text{vec}(\bar{\mathbf{B}}), \sigma^2\boldsymbol{\Phi} \otimes \boldsymbol{\Omega}\big), \\ \boldsymbol{\varepsilon}|\sigma^2 &\sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}). \end{aligned} \tag{13}$$

To complete the model, we specify the following priors:

$$\begin{aligned} \boldsymbol{\beta}_2|\sigma^2 &\sim \mathcal{N}(\mathbf{0}, u \cdot \sigma^2\mathbf{I}), \\ \mathbf{b}|\sigma^2 &\sim \mathcal{N}(\mathbf{0}, u \cdot \sigma^2\mathbf{I}), \\ \boldsymbol{\Phi} &\sim \mathcal{W}^{-1}(\mathbf{P}, \nu), \end{aligned} \tag{14}$$

where u is a scalar that allows us to set the prior uncertainty on $\boldsymbol{\beta}_2$ and $\mathbf{b}$. Here, $\mathcal{W}^{-1}$ indicates the inverse Wishart distribution. We use a diffuse prior on σ^2, and the settings for the prior on τ will be discussed separately in Section 4.4 and provided for each illustration in the results section later in the paper.

Note that to introduce the model, we considered the linear regression setup in Equation (9). However, models other than the linear Scheffé models and models for dependent variables that are not continuous can be developed in a similar manner. In Section 4.3, we work out details for a model in which the dependent variable is binary.

3.2 Variance-covariance structure of the mixture parameters

In this section, we exploit the structure of the variance-covariance matrix of the Gaussian process $(\sigma^2\boldsymbol{\Phi} \otimes \boldsymbol{\Omega})$ to model the correlation across the mixture parameters.

Consider again the mixture parameters stacked in the matrix $\mathbf{B}(\vec{A})$ as in Equation (8). One parameter $\beta_m(A_i)$ specifies the impact of a particular mixture proportion or a cross-product of proportions on the response at a specific value A_i of the amount variable. These parameters are not independent. First, our model incorporates the idea that for amount values that are close to each other, the model parameters are expected to be rather similar. Intuitively, the value of $\beta_m(A')$ should be similar to that of $\beta_m(A'')$ if $A' \approx A''$. In the model, this is captured by the correlation between the parameters $\beta_m(A')$ and $\beta_m(A'')$. The correlation increases when the amounts A' and A'' are closer together. Second, at a given amount A', the parameters $\beta_1(A'), \beta_2(A'), \ldots, \beta_p(A')$ might also exhibit some correlation. For example, the effects of radio advertising and TV advertising may move up and down together as the advertising intensity changes. We allow for this type of correlation by letting $\beta_1(A'), \beta_2(A'), \ldots, \beta_p(A')$ be correlated as well.

We begin by specifying the correlation structure across different amount values for a given parameter β_m. We specify the elements of $\text{Var}\big(\beta_m(\vec{A})\big) = \mathbf{\Omega}$ as

$$\mathbf{\Omega}(A', A'') = \exp\Big(-\frac{1}{2\tau^2}||A' - A''||^2\Big), \quad \tau > 0, \tag{15}$$

where A' and A'' are two amount values and τ denotes a model parameter. The function in Equation (15) is called the squared exponential (Gaussian) kernel. For any pair of amounts, A' and A'', a Gaussian process with this correlation function implies:

- $\beta_m(A')$ and $\beta_m(A'')$ will tend to have high correlation if A' and A'' are "close" to each another, since $||A' - A''||$ will then approach zero and $\mathbf{\Omega}(A', A'') = \exp\big(-\frac{1}{2\tau^2}||A' - A''||^2\big)$ will tend to one,

- $\beta_m(A')$ and $\beta_m(A'')$ will tend to have low correlation if A' and A'' are "far" apart, since $||A' - A''||$ will then be a large positive value and $\mathbf{\Omega}(A', A'') = \exp\big(-\frac{1}{2\tau^2}||A' - A''||^2\big)$ will tend to zero.

In other words, functions drawn from a Gaussian process with the Gaussian kernel will be locally smooth with high probability. This means that the mixture parameter values for amounts that are similar will also be similar. The similarity between the mixture parameters will decrease with the distance between A' and A''.

The parameter τ in Equation (15) controls the smoothness of the function of $\beta_m(A)$ as it determines how quickly $\beta_m(A)$ varies with A. By varying τ, we can in fact capture many different scenarios. By setting this parameter to zero or to a large positive value, we obtain standard models as special cases. In particular, if we take $\tau \to 0$, we allow separate mixture parameters for each value of the amount, as, if $A' \neq A''$ and τ approaches zero, $\mathbf{\Omega}(A', A'') = \exp(-\frac{1}{2\tau^2}||A' - A''||^2)$ tends to zero as well. On the other hand, if we let $\tau \to \infty$, we allow constant mixture parameters (i.e., independent of A), as, when τ increases, $\mathbf{\Omega}(A', A'') = \exp(-\frac{1}{2\tau^2}||A' - A''||^2)$ tends to one. By taking τ values between 0 and ∞, we can describe settings in between constant and separate mixture parameters, without restricting ourselves to any particular function for $\beta_m(A)$.

Instead of using some parametric function to model the mixture parameters in terms of the total amount variable, we only assume that the mixture parameters vary smoothly with the amount. This smoothness is controlled by a single parameter τ that can even be estimated. Such a specification is flexible enough to represent many different parametric forms that would require a very large number of parameters if formulated in a conventional way.

Whereas the correlation matrix $\mathbf{\Omega}$ captures the correlation structure for a given parameter β_m across different amount values, the covariance across the mixture parameters at a given amount is described by matrix $\mathbf{\Phi}$. At a given amount A', we have $\text{Var}\big(\boldsymbol{\beta}(A')\big) = \mathbf{\Phi}$, where $\boldsymbol{\beta}(A') =$

$\big(\beta_1(A'), \ldots, \beta_p(A')\big)'$ is the vector of the mixture parameters at A'. In the context of mixtures, this covariance is expected to be non-zero as a result of the direct impact of the total amount on the response. As an extreme illustration, consider a case where there are only two mixture ingredients, x_1 and x_2. Assume that their proportions have a constant impact on the response, while the amount A has a direct impact. Our proposed model will capture such a case with $\mathbf{\Phi}$ implying equal variances and a perfect correlation between $\beta_1(A)$ and $\beta_2(A)$. To see this, start with the model

$$y = \beta_1(A)x_1 + \beta_2(A)x_2 + \varepsilon.$$

The perfect correlation in combination with equal variances implies that we can write $\beta_1(A) = b_1 + \alpha(A)$ and $\beta_2(A) = b_2 + \alpha(A)$, where $\alpha(A)$ is a one-dimensional Gaussian process with mean zero and b_1 and b_2 are the means of $\beta_1(A)$ and $\beta_2(A)$, respectively. Using the mixture constraint, we can now rewrite the model as

$$\begin{aligned} y &= b_1x_1 + \alpha(A)x_1 + b_2x_2 + \alpha(A)x_2 + \varepsilon \\ &= \alpha(A) + b_1x_1 + b_2x_2 + \varepsilon. \end{aligned}$$

As in practice we usually expect a direct effect of the amount, $\mathbf{\Phi}$ will usually involve non-zero correlations. Note that, as $\mathbf{\Phi}$ is an unrestricted variance-covariance matrix, we need to restrict $\mathbf{\Omega}$ to be a correlation matrix to ensure identification.

4 Estimation

In this section, we discuss the Bayesian estimation of the model parameters in Equation (13) using Markov Chain Monte Carlo (MCMC) sampling. This approach requires taking draws from the joint posterior distribution of the model parameters (Bishop, 2006; Gelman et al., 2013; Greenberg, 2014; Zellner, 1996). However, sampling from the joint posterior density $p\Big(\tau, \mathbf{b}, \mathbf{B}(\vec{A}), \boldsymbol{\beta}_2, \mathbf{\Phi}, \sigma^2 \Big| \mathbf{y}\Big)$ is not directly feasible. Instead, we employ a Gibbs sampler (Casella and George, 1992) and repeatedly sample from conditional posterior distributions.

4.1 Sampling strategy

A straightforward Gibbs sampler where each parameter is drawn from its full conditional posterior is not efficient. The main reason for this is that we expect $\mathbf{B}(\vec{A})$ and τ to be strongly correlated: when τ is large (small) we expect quite similar (different) parameter values across the amount levels and vice versa. At the same time, $\mathbf{B}(\vec{A})$ and $\mathbf{b}$ are also likely to be strongly correlated. To

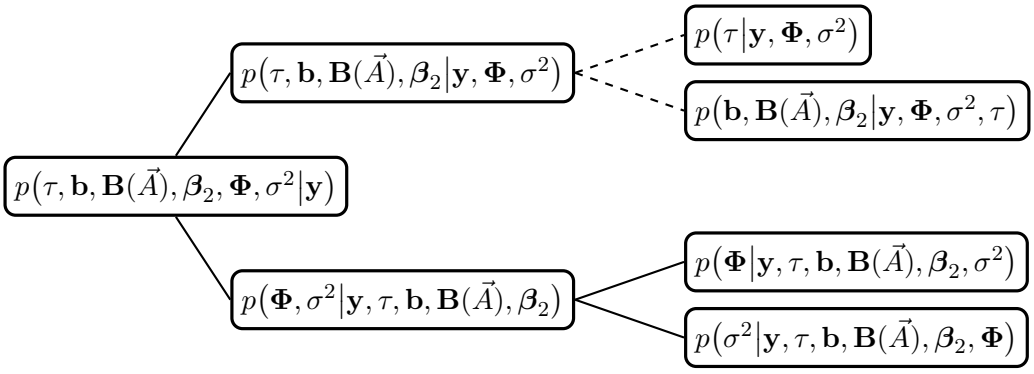


Figure 1: Decomposition of sampling from the joint posterior distribution into sampling from conditional posterior distributions used in the MCMC sampling. Dashed lines symbolize exact decompositions, solid lines symbolize decompositions based on Gibbs sampling

reduce the dependence between the draws of $\mathbf{B}(\vec{A})$ and τ (and $\mathbf{b}$), we use the decomposition

$$p\Big(\tau,\mathbf{b},\mathbf{B}(\vec{A}),\boldsymbol{\beta}_2\Big|\mathbf{y},\mathbf{\Phi},\sigma^2\Big) = p\Big(\mathbf{b},\mathbf{B}(\vec{A}),\boldsymbol{\beta}_2\Big|\mathbf{y},\mathbf{\Phi},\sigma^2,\tau\Big) \times p\Big(\tau\Big|\mathbf{y},\mathbf{\Phi},\sigma^2\Big)$$

and apply Gibbs sampling steps for the latter distributions, where the sampling distribution of τ is not conditional on $\mathbf{B}(\vec{A})$ and $\mathbf{b}$. A Metropolis-Hastings (Chib and Greenberg, 1995) step within a Gibbs sampler is needed to sample τ. As a result, we iteratively sample from the four conditional distributions given at the right-hand side of Figure 1. Figure 1 graphically demonstrates how the sampling from the full posterior distribution is decomposed into iterative sampling from conditional posterior distributions.

In theory, one could treat the Gaussian process as a standard prior on $\text{vec}\big(\mathbf{B}(\vec{A})\big)$. In our case, such an approach is numerically infeasible as some individual parameters may exhibit extreme correlations. This is especially true if some observed amount values are almost the same or if $\tau \to \infty$. The correlation matrix $\mathbf{\Omega}$ then becomes (nearly) singular and the traditional inverse of $\mathbf{\Omega}$ does not exist.

To make the estimation numerically tractable when $\tau \to \infty$ or some observed amount values are nearly identical, we use the singular value decomposition of the correlation matrix $\mathbf{\Omega}$, that is,

$$\mathbf{\Omega} = \mathbf{USV}' = \mathbf{USU}'. \tag{16}$$

Here, $\mathbf{U}$ is a real unitary matrix ($\mathbf{UU}' = \mathbf{I}$ with $\mathbf{I}$ an identity matrix) and $\mathbf{S}$ is a diagonal matrix. If $\mathbf{\Omega}$ is singular, some diagonal elements of $\mathbf{S}$ will be equal to zero. If it is nearly singular, these values will be close to zero. To improve the numerical stability of our estimation procedure, we replace these small diagonal elements by zeros. The threshold that we use to define a non-zero element is 10^{-6}. The corresponding matrix is denoted by $\mathbf{S}^*$. We now have $\mathbf{\Omega} \approx \mathbf{US}^*\mathbf{U}'$. Using

this relation, we define the inverse of $\mathbf{\Omega}$ as

$$\mathbf{\Omega}^{-1} = \mathbf{U}\mathbf{S}^{*-1}\mathbf{U}', \tag{17}$$

where $\mathbf{S}^{*-1}$ is a diagonal matrix containing the reciprocals of all r^* non-zero diagonal entries of $\mathbf{S}^*$. This procedure is known as taking the Moore-Penrose pseudoinverse (Ben-Israel and Greville, 2003).

We next define the Choleski decomposition of $\mathbf{\Omega}$ as

$$\mathbf{\Omega}^{\frac{1}{2}} = \mathbf{U}\mathbf{S}^{*\frac{1}{2}}_{r\times r^*}, \tag{18}$$

where $\mathbf{S}^{*\frac{1}{2}}_{r\times r^*}$ is the $r \times r^*$-dimensional matrix obtained by taking the square root of the non-zero elements of $\mathbf{S}^*$ and dropping the final $(r - r^*)$ zero columns. Note that $\mathbf{\Omega}^{\frac{1}{2}}\mathbf{\Omega}^{\frac{1}{2}\prime}$ indeed approximately equals $\mathbf{\Omega}$. The fact that $\mathbf{\Omega}$ is singular is reflected by the fact that the matrix $\mathbf{\Omega}^{\frac{1}{2}}$ is not square if $r^* \neq r$.

We exploit the singular value decomposition to sample $\mathrm{vec}\big(\mathbf{B}(\vec{A})\big)$ in the appropriate lower dimensional space if $\mathbf{\Omega}$ is (nearly) singular. We define $\mathrm{vec}\big(\mathbf{B}(\vec{A})\big) = \mathrm{vec}\big(\bar{\mathbf{B}}\big) + \big(\mathbf{F} \otimes \mathbf{\Omega}^{\frac{1}{2}}\big)\boldsymbol{\gamma}\big(\vec{A}\big)$ with $\boldsymbol{\gamma}(\vec{A}) \sim \mathcal{N}(\mathbf{0}_{pr^*\times 1}, \sigma^2\mathbf{I}_{pr^*\times pr^*})$, where $\mathbf{F}$ is the lower-triangular matrix resulting from the Choleski decomposition of $\mathbf{\Phi}$, i.e., $\mathbf{\Phi} = \mathbf{F}\mathbf{F}'$. Note that the distribution of $\boldsymbol{\gamma}(\vec{A})$ is a distribution in a lower dimensional space which naturally leads to $\mathrm{vec}\big(\mathbf{B}(\vec{A})\big) \sim \mathcal{N}\big(\mathrm{vec}\big(\bar{\mathbf{B}}\big), \sigma^2\mathbf{\Phi} \otimes \mathbf{\Omega}\big)$ as the implied distribution of $\mathrm{vec}\big(\mathbf{B}(\vec{A})\big)$, just as we defined before. Using this and the definition of $\bar{\mathbf{B}}$ (see Equation (12)) we can write

$$\begin{aligned}
\mathbf{X}\mathrm{vec}\big(\mathbf{B}(\vec{A})\big) &= \mathbf{X}\left(\mathrm{vec}\big(\bar{\mathbf{B}}\big) + \big(\mathbf{F} \otimes \mathbf{\Omega}^{\frac{1}{2}}\big)\boldsymbol{\gamma}\big(\vec{A}\big)\right) \\
&= \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{pmatrix}\mathbf{b} + \mathbf{X}\big(\mathbf{F} \otimes \mathbf{\Omega}^{\frac{1}{2}}\big)\boldsymbol{\gamma}\big(\vec{A}\big) \\
&= \mathbf{Z}\mathbf{b} + \mathbf{X}\big(\mathbf{F} \otimes \mathbf{\Omega}^{\frac{1}{2}}\big)\boldsymbol{\gamma}\big(\vec{A}\big),
\end{aligned}$$

where the matrix $\mathbf{Z} = (\mathbf{x}_1', \ldots, \mathbf{x}_N')'$ contains all explanatory variables in the Scheffé model used, ignoring the fact that different observations correspond to different amounts. Note that $\boldsymbol{\gamma}\big(\vec{A}\big)$ has a lower dimensionality, namely pr^*, than $\mathbf{B}(\vec{A})$, namely pr.

We can now sample $\boldsymbol{\gamma}(\vec{A})$ instead of $\mathrm{vec}\big(\mathbf{B}(\vec{A})\big)$ and avoid numerical issues due to potentially strong correlations. When some of the correlations in $\mathbf{\Omega}$ become too large, the singular value decomposition makes sure that the parameters are sampled in the lower dimensional space.

Applying the above, we rewrite the model in Equation (13) and the priors in Equation (14) as

$$\mathbf{y} = \mathbf{X}\text{vec}\big(\mathbf{B}(\vec{A})\big) + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} = \mathbf{X}\big(\mathbf{F}\otimes\boldsymbol{\Omega}^{\frac{1}{2}}\big)\boldsymbol{\gamma}\big(\vec{A}\big) + \mathbf{Z}\mathbf{b} + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \tag{19}$$

where $\mathbf{X}^* = \Big(\mathbf{X}\big(\mathbf{F}\otimes\boldsymbol{\Omega}^{\frac{1}{2}}\big) \quad \mathbf{Z} \quad \mathbf{X}_2\Big)$ and

$$\boldsymbol{\beta}^* = \begin{pmatrix}\boldsymbol{\gamma}(\vec{A})\\ \mathbf{b}\\ \boldsymbol{\beta}_2\end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}_{p(r^*+1)+d\times 1}, \sigma^2\begin{pmatrix}\mathbf{I}_{pr^*\times pr^*} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & u\mathbf{I}_{p\times p} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & u\mathbf{I}_{d\times d}\end{pmatrix}\right)$$

or, more compactly,

$$\boldsymbol{\beta}^* \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}^*),$$

with $\boldsymbol{\Sigma}^* = \text{diag}(\mathbf{I}_{pr^*\times pr^*}, u\mathbf{I}_{(p+d)\times(p+d)})$. We can next apply standard results to derive all the sampling steps needed for our model.

4.2 Sampling distributions

To obtain the conditional posterior distribution of τ, $p(\tau|\mathbf{y}, \boldsymbol{\Phi}, \sigma^2)$, we need to integrate over the distribution of $\boldsymbol{\gamma}(\vec{A})$, $\mathbf{b}$ and $\boldsymbol{\beta}_2$. The posterior distribution of τ is obtained as $p(\tau|\mathbf{y}, \boldsymbol{\Phi}, \sigma^2) \propto p(\mathbf{y}|\tau, \boldsymbol{\Phi}, \sigma^2)p(\tau)$, where $p(\tau)$ is the prior distribution of τ and

$$\begin{aligned} p\Big(\mathbf{y}\Big|\tau, \boldsymbol{\Phi}, \sigma^2\Big) &= \int_{\boldsymbol{\gamma}(\vec{A}),\mathbf{b},\boldsymbol{\beta}_2} p\Big(\mathbf{y}\Big|\tau, \mathbf{b}, \boldsymbol{\gamma}(\vec{A}), \boldsymbol{\beta}_2, \boldsymbol{\Phi}, \sigma^2\Big)p\Big(\boldsymbol{\gamma}(\vec{A}), \mathbf{b}, \boldsymbol{\beta}_2\Big|\tau, \boldsymbol{\Phi}, \sigma^2\Big)d\boldsymbol{\gamma}(\vec{A})d\mathbf{b}d\boldsymbol{\beta}_2\\ &= \int_{\boldsymbol{\beta}^*} p\Big(\mathbf{y}\Big|\tau, \boldsymbol{\beta}^*, \boldsymbol{\Phi}, \sigma^2\Big)p\Big(\boldsymbol{\beta}^*\Big|\sigma^2\Big)d\boldsymbol{\beta}^*\\ &\propto \exp\Big(-\frac{1}{2\sigma^2}\Big(\mathbf{w}-\mathbf{V}\hat{\beta}\Big)'\Big(\mathbf{w}-\mathbf{V}\hat{\beta}\Big)\Big) \times \Big|\Big(\mathbf{V}'\mathbf{V}\Big)^{-1}\Big|^{\frac{1}{2}}, \end{aligned} \tag{20}$$

where $\mathbf{w} = \big(\mathbf{y}' \quad (\mathbf{0}_{p(r^*+1)+d\times 1})'\big)'$, $\mathbf{V} = \big(\mathbf{X}^{*\prime} \quad (\boldsymbol{\Sigma}^{*-\frac{1}{2}})'\big)'$, $\hat{\beta} = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{w}$ and $\boldsymbol{\Sigma}^{*-1} = \boldsymbol{\Sigma}^{*-\frac{1}{2}\prime}\boldsymbol{\Sigma}^{*-\frac{1}{2}}$. The last step in this derivation follows from standard results for the linear model with a Gaussian prior applied to Equation (19).

Since the resulting posterior for τ, $p(\tau|\mathbf{y}, \boldsymbol{\Phi}, \sigma^2)$, is not of a known type, we apply the random-walk Metropolis-Hastings sampler to sample from it. We set the candidate generating function to be

$$\begin{aligned} \log\big(\tau^{\text{Cand}}\big) &= \log\big(\tau^{\text{Prev}}\big) + \eta,\\ \eta &\sim \mathcal{N}(0, \kappa^2). \end{aligned} \tag{21}$$

The acceptance probability of τ^{Cand} is then calculated as

$$\alpha = \min\left(\frac{p\big(\tau^{\text{Cand}}\big|\mathbf{y},\mathbf{\Phi},\sigma^2\big)g\big(\tau^{\text{Prev}}\big|\tau^{\text{Cand}}\big)}{p\big(\tau^{\text{Prev}}\big|\mathbf{y},\mathbf{\Phi},\sigma^2\big)g\big(\tau^{\text{Cand}}\big|\tau^{\text{Prev}}\big)},1\right), \tag{22}$$

where $g(\cdot)$ is the density of the candidate generating function in Equation (21) (Chib and Greenberg, 1995).

To sample $\boldsymbol{\beta}^*$, we consider the model given in Equation (19). The kernel of the conditional posterior distribution for $\boldsymbol{\beta}^*$ is

$$\boldsymbol{\beta}^*\Big|\mathbf{y},\mathbf{\Phi},\sigma^2,\tau \propto \exp\Big(-\frac{1}{2\sigma^2}\big(\boldsymbol{\beta}^*-\bar{\boldsymbol{\beta}}^*\big)'\big(\mathbf{X}^{*\prime}\mathbf{X}^*+\mathbf{\Sigma}^{*-1}\big)\big(\boldsymbol{\beta}^*-\bar{\boldsymbol{\beta}}^*\big)\Big), \tag{23}$$

where $\bar{\boldsymbol{\beta}}^* = \big(\mathbf{X}^{*\prime}\mathbf{X}^*+\mathbf{\Sigma}^{*-1}\big)^{-1}\mathbf{X}^{*\prime}\mathbf{y}$. This is the kernel of a multivariate normal distribution with mean $\bar{\boldsymbol{\beta}}^*$ and variance-covariance matrix $\sigma^2\big(\mathbf{X}^{*\prime}\mathbf{X}^*+\mathbf{\Sigma}^{*-1}\big)^{-1}$. Since $\bar{\mathbf{B}} = (\mathbf{1}_{r\times 1}\otimes\mathbf{b}')$, $\boldsymbol{\beta}^* = \big(\boldsymbol{\gamma}(\vec{A})'\ \ \mathbf{b}'\ \ \boldsymbol{\beta}_2'\big)'$ and $\text{vec}\big(\mathbf{B}(\vec{A})\big) = \text{vec}\big(\bar{\mathbf{B}}\big) + \big(\mathbf{F}\otimes\mathbf{\Omega}^{\frac{1}{2}}\big)\boldsymbol{\gamma}\big(\vec{A}\big)$, we can obtain draws for $\text{vec}\big(\mathbf{B}(\vec{A})\big)$ from draws for $\mathbf{b}$ and $\boldsymbol{\gamma}(\vec{A})$, see the discussion above.

We sample $\mathbf{\Phi}$ from the inverted Wishart distribution with parameters $\sigma^{-2}\big(\mathbf{B}(\vec{A})-\bar{\mathbf{B}}\big)'\mathbf{\Omega}^{-1}\big(\mathbf{B}(\vec{A})-\bar{\mathbf{B}}\big)+\mathbf{P}$ and $r^*+\nu$, where $\mathbf{P}$ and ν give the prior scale and degrees of freedom, respectively.

Finally, we sample σ^2 from the inverted Gamma-2 distribution with parameter $(\mathbf{y}-\mathbf{X}^*\boldsymbol{\beta}^*)'(\mathbf{y}-\mathbf{X}^*\boldsymbol{\beta}^*)+\sigma^{-2}\boldsymbol{\beta}^{*\prime}\mathbf{\Sigma}^{*-1}\boldsymbol{\beta}^*$ and $N+p(r^*+1)+d$ degrees of freedom.

4.3 Limited dependent variables

The ideas above can be easily generalized to deal with limited dependent variables. When the dependent variable $\mathbf{y}$ is binary, we employ the estimation procedure by Albert and Chib (1993); Allenby and Rossi (1999); McCulloch and Rossi (1994, 2000); Train (2009). Write the model as

$$y_i = \begin{cases} 1 \text{ if} & z_i = X_i\text{vec}\big(\mathbf{B}(\vec{A})\big) + X_{2i}\boldsymbol{\beta}_2 + \varepsilon_i > 0, \\ 0 \text{ if} & z_i = X_i\text{vec}\big(\mathbf{B}(\vec{A})\big) + X_{2i}\boldsymbol{\beta}_2 + \varepsilon_i \leq 0, \end{cases}$$

for $i = 1,\ldots,N$, where $\varepsilon_i \sim \mathcal{N}(0,1)$ and all other details of the model stay the same. The only detail to note is that the variance of ε_i is restricted to one. Therefore, in the notation of the previous sections, we restrict σ^2 to be one. For parameter inference, we sample z_i, $i = 1,\ldots,N$, alongside the other parameters as part of the Gibbs sampler (keeping $\sigma^2 = 1$ fixed). Denote $\mathbf{z} = (z_1,\ldots,z_N)'$. The conditional distribution for the only additional step is given by

$$p\big(\mathbf{z}\big|\mathbf{y},\tau,\mathbf{b},\text{vec}\big(\mathbf{B}(\vec{A})\big),\boldsymbol{\beta}_2,\mathbf{\Phi},\sigma^2=1\big).$$

All elements of $\mathbf{z}$ are independent of each other conditional on the model parameters. Hence, for the i^{th} element z_i, the distribution reduces to

$$z_i \big| y_i, \text{vec}\big(\mathbf{B}(\vec{A})\big), \boldsymbol{\beta}_2 \sim \begin{cases} \mathcal{N}\Big(X_i \text{vec}\big(\mathbf{B}(\vec{A})\big) + X_{2i}\boldsymbol{\beta}_2, 1\Big)\mathcal{I}\big(z_i > 0\big) & \text{if } y_i = 1, \\ \mathcal{N}\Big(X_i \text{vec}\big(\mathbf{B}(\vec{A})\big) + X_{2i}\boldsymbol{\beta}_2, 1\Big)\mathcal{I}\big(z_i \leq 0\big) & \text{if } y_i = 0, \end{cases}$$

with $\mathcal{I}(\cdot)$ denoting the indicator function and $i = 1, \ldots, N$.

4.4 Prior specification for τ

Some care is needed when choosing a prior distribution for τ. In this section, we provide some intuition on how we specify this prior.

First, from Equation (15), the correlation between $\beta_m(A')$ and $\beta_m(A'')$ for $A' \neq A''$ depends only on the difference between A' and A''. Note that for larger amount values, a larger value of τ is required to represent the same level of correlation. The prior required for τ therefore depends on the scale of the amount variable. To deal with this issue, we standardize the total amount variable, by dividing it by its standard deviation. Note that standardization preserves the ranking and the pairwise ratios of the total amount values. Moreover, from the prior distribution for τ when the total amount variable is standardized, we can always derive the corresponding prior distribution for τ for the original amount values. To show this, denote the standard deviation of the amount variable in the data by S. Write then

$$\boldsymbol{\Omega}(A', A'') = \exp\Big(-\frac{1}{2\tau^2}\big|\big|A' - A''\big|\big|^2\Big) = \exp\Big(-\frac{1}{2(\tau/S)^2}\big|\big|A'/S - A''/S\big|\big|^2\Big) = \exp\Big(-\frac{1}{2\tau^{*2}}\big|\big|A^{*'} - A^{*''}\big|\big|^2\Big),$$

where $\tau^* = \tau/S$ and $A^* = A/S$ is the standardized amount value. Now, if we consider the standardized amounts, we begin by specifying a prior distribution for τ^*. Then, to obtain the corresponding distribution for τ for the original amount values, we can use $\tau = \tau^* S$, where the distribution for τ^* is known.

Second, in some cases, we may want to use a prior distribution with a strictly positive domain, that is, a prior that sets zero probability on no correlation. Such a prior is especially useful for data where only one observation per amount value is available. Here, if we allowed the Gaussian process to have zero correlation, we would end up with independent parameters for each individual observation. Naturally, such a specification does not make sense. Even if multiple observations per amount level are available, one may still want to impose such a prior if one expects the amount levels to be somehow related to some unobserved factors in the data generating process.

Finally, when choosing a prior for τ we are in fact specifying a prior on the correlation structure across different amount values. An uninformative prior for τ may sometimes lead to a very

informative specification for the correlations. Therefore, after specifying a prior for τ, it is useful to inspect the implied prior for the correlations (see Gelman (2006); Gilmour and Goos (2009) for a related discussion).

5 Illustrations

In this section, we consider two data sets to illustrate our approach. The first data set describes how mice react to hormone mixtures administered at three different amount levels. The dependent variable here is continuous and describes cornification of the vaginal epithelium. Using this data set, we demonstrate that two common models in the mixture-amount literature are special cases of our model. In these special cases, a certain functional form is assumed for the mixture parameters.

Often, it is not a-priori known how the mixture parameters depend on the total amount meaning that functional form assumptions may not be justified. Therefore, we demonstrate next how, in our approach, we estimate this relationship without making any parametric assumptions. We do so using a realistic data set, which describes to what extent women recognize advertisements run in magazines and/or on television with different intensities as measured by Gross Rating Points (GRPs) (operationalized as the amount variable, with 52 unique values). The dependent variable here is binary and indicates whether an advertising campaign is recognized (1) or not (0).

5.1 Mice experiment

For the first example, we consider data from Claringbold (1955) who presented an experiment involving 10 different mixtures of three distinct hormones administered to 10 groups of 12 mice each. Each hormone mixture was studied at three amount levels, 0.75×10^{-4} μg (A_1), 1.50×10^{-4} μg (A_2) and 3.00×10^{-4} μg (A_3), and so there were 30 experimental runs in total. The response variable of interest is the fraction of mice in each group (out of 12) that responded to each of the 30 mixture-amount combinations. The dependent variable considered is the angular transformation of the fractions, see Claringbold (1955) for details. We replicate the data in Table 1.

We use these data to estimate the parameters of two simple mixture-amount models and to demonstrate that they are special cases of the methodology we introduce in this paper. Consider first a simple linear Scheffé model for the complete dataset, ignoring the amount (see Equation (1)):

$$y_i = \beta_1^0 x_{1i} + \beta_2^0 x_{2i} + \beta_3^0 x_{3i} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_0^2). \tag{24}$$

Next, consider the same linear regression for each observed amount level separately, that is,

$$y_i = \beta_1^1 x_{1i} + \beta_2^1 x_{2i} + \beta_3^1 x_{3i} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_1^2), \qquad \text{if } i \text{ corresponds to } A_1, \tag{25}$$

Hormone proportion			Percent response ($p \times 100$)			Angular response (y)		
x_1	x_2	x_3	A_1	A_2	A_3	A_1	A_2	A_3
1	0	0	17	42	83	24.09	40.20	65.91
0	1	0	58	58	100	49.80	49.80	81.70
0	0	1	25	50	42	30.00	45.00	40.20
$\frac{2}{3}$	$\frac{1}{3}$	0	0	33	75	8.30	35.26	60.00
$\frac{1}{3}$	$\frac{2}{3}$	0	33	33	75	35.26	35.26	60.00
$\frac{2}{3}$	0	$\frac{1}{3}$	0	25	75	8.30	30.00	60.00
$\frac{1}{3}$	0	$\frac{2}{3}$	25	42	42	30.00	40.20	40.20
0	$\frac{2}{3}$	$\frac{1}{3}$	17	33	67	24.09	35.26	54.74
0	$\frac{1}{3}$	$\frac{2}{3}$	33	33	58	35.26	35.26	49.80
$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	17	25	58	24.09	30.00	49.80

Table 1: Data for the mice experiment

$$y_i = \beta_1^2 x_{1i} + \beta_2^2 x_{2i} + \beta_3^2 x_{3i} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_2^2), \qquad \text{if } i \text{ corresponds to } A_2, \tag{26}$$

$$y_i = \beta_1^3 x_{1i} + \beta_2^3 x_{2i} + \beta_3^3 x_{3i} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_3^2), \qquad \text{if } i \text{ corresponds to } A_3. \tag{27}$$

As priors, we take $\boldsymbol{\beta}^j = (\beta_1^j, \beta_2^j, \beta_3^j)' | \sigma_j^2 \sim \mathcal{N}(\mathbf{0}_{3\times 1}, 10^3 \times \sigma_j^2 \mathbf{I}_{3\times 3})$ ($j = 0, 1, 2, 3$ corresponding to the models in Equations (24)-(27), respectively) and the diffuse priors for the variance parameters. We display summary statistics for the posterior sample of $100,000$ draws in Table 2.

	Model			
	Ignoring amount (Eq. (24), $j = 0$)	Considering A_1 only (Eq. (25), $j = 1$)	Considering A_2 only (Eq. (26), $j = 2$)	Considering A_3 only (Eq. (27), $j = 3$)
β_1^j	35.66	12.03	33.68	61.31
	(6.35)	(6.75)	(4.43)	(4.85)
β_2^j	50.49	40.24	40.58	70.59
	(6.34)	(6.75)	(4.44)	(4.87)
β_3^j	34.60	28.47	38.59	36.78
	(6.35)	(6.75)	(4.47)	(4.84)
σ_j^2	241.33	91.54	39.61	47.29
	(67.07)	(53.25)	(22.77)	(27.36)

Table 2: Posterior means and standard deviations (in parentheses) for the parameters of the models in Equations (24)-(27)

The model in Equation (24) assumes the same parameters irrespective of the amount value.

The models in Equations (25)-(27) assume different, and unrelated, parameters for each amount value. We argue that these two cases are nested within the model which we introduce in this paper. As a result, the model based on the Gaussian process prior should be able to replicate the results obtained above. Prior to analysis, to make τ less dependent on the scale of the amount variable (see Section 4.4), we standardize the amount. In order to obtain a model where the mixture parameters are independent of the amount, as in Equation (24), we set a large value for τ ($\tau = 10,000$) in Equation (15). Furthermore, we fix $\mathbf{b} = \mathbf{0}_{pr\times 1}$ and $\mathbf{\Phi} = 10^3 \times \mathbf{I}_{p\times p}$, to match the typical uninformative regression setting. In fact, then the posterior means of $\beta_1^j, \beta_2^j, \beta_3^j, \sigma_j^2$, $j = 0, 1, 2, 3$, will be plain OLS estimates. The second column in Table 3 gives the summary statistics for the posterior sample of $100,000$ draws. We can clearly see that the estimates of the mixture parameters are indeed constant with respect to the amount variable. Furthermore, they are not very different from the parameter estimates obtained for the model in Equation (24), shown in the second column of Table 2.

	Model		
	constant parameters ($\tau = 10,000$)	different parameters per amount ($\tau = 0$)	benchmark models (Eq. (25) - (27))
$\beta_1(A_1)$	35.67	12.02	12.03
	(6.34)	(5.03)	(6.75)
$\beta_1(A_2)$	35.67	33.66	33.68
	(6.34)	(5.05)	(4.43)
$\beta_1(A_3)$	35.67	61.31	61.31
	(6.34)	(5.05)	(4.85)
$\beta_2(A_1)$	50.48	40.23	40.24
	(6.36)	(5.05)	(6.75)
$\beta_2(A_2)$	50.48	40.57	40.58
	(6.36)	(5.04)	(4.44)
$\beta_2(A_3)$	50.48	70.56	70.59
	(6.36)	(5.03)	(4.87)
$\beta_3(A_1)$	34.61	28.48	28.47
	(6.33)	(5.05)	(6.75)
$\beta_3(A_2)$	34.61	38.59	38.59
	(6.33)	(5.03)	(4.47)
$\beta_3(A_3)$	34.61	36.77	36.78
	(6.33)	(5.04)	(4.84)
σ^2	241.60	50.92	-
	(66.80)	(14.00)	-

Table 3: Posterior means and standard deviations (in parentheses) for parameters obtained from the Gaussian process prior model (columns 2-3) and from the benchmark models (column 4)

To obtain a model with separate independent parameters for each observed amount level, as in Equations (25)-(27), we set $\tau = 0$ in Equation (15) and fix $\mathbf{b} = \mathbf{0}_{pr\times 1}$ and $\mathbf{\Phi} = 10^3 \times \mathbf{I}_{p\times p}$.

The posterior means and standard deviations (in parentheses) for the parameters are given in the third column of Table 3. For ease of comparison, we repeat the estimates from Table 2 (columns 3-5) in the last column of Table 3. The corresponding parameter estimates are again very similar. As a result, the mixture-amount model based on the Gaussian process prior indeed covers the two extreme scenarios described in Equations (24)-(27).

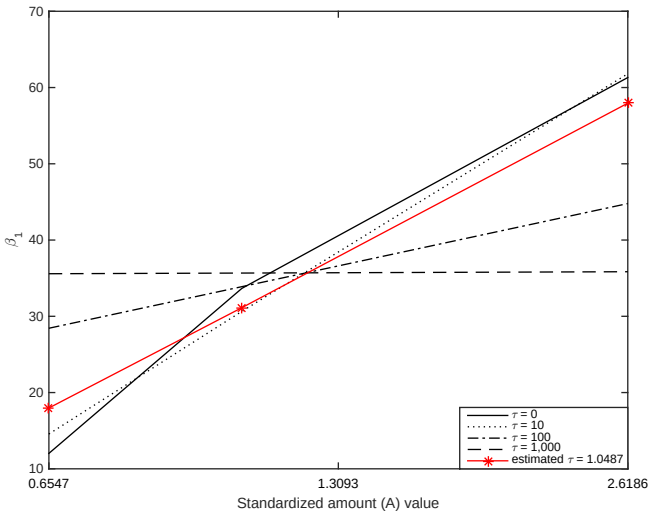
Using $\tau = 0$ and $\tau = 10,000$ led to either independent or constant parameters across amount values. By choosing $0 < \tau < \infty$, we can describe many different intermediate settings without explicitly assuming a particular parametric form for each $\beta_m(A)$. We illustrate this in Figure 2 where we use τ values of $0; 10; 100$ and $1,000$ and plot the corresponding posterior means of the mixture parameters with respect to the standardized amount variable. In the same figure, we plot the posterior means of the mixture parameters for an estimated $\tau = 1.0487$ (in red), obtained using the prior $p(\tau) \sim \ln\mathcal{N}(0.1, 0.4^2)$ and uninformative priors for the other parameters. It is clear that, by changing τ, we can describe many different scenarios and, when $\tau \to \infty$, the mixture parameters no longer change with the amount.

In this section, we demonstrated that our model based on the Gaussian process prior, if τ is chosen accordingly, can in fact replicate two simple models for mixture-amount data. We also showed the mixture parameters for an estimated τ value without getting into detail of how we do this. In the next section, we consider the second data set and demonstrate the estimation of τ together with β_m, $\mathbf{b}$, $\mathbf{\Phi}$ and σ^2.

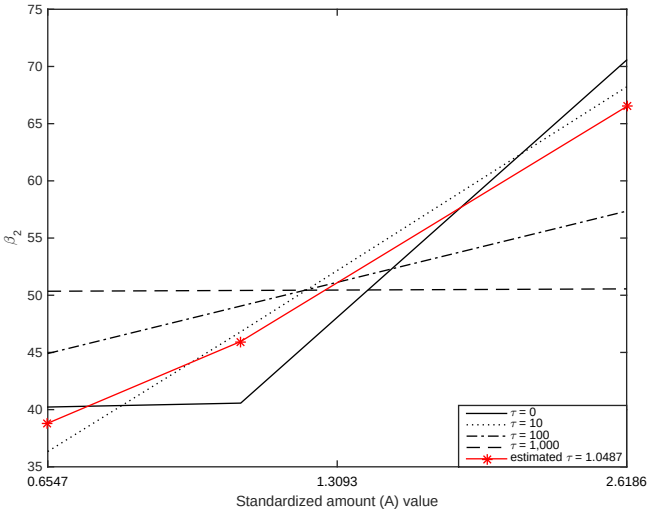
5.2 Advertising campaign recognition

In this section, we consider an application concerning advertising campaign recognition. In a questionnaire, individual female respondents indicated whether they recognized various skin and hair care advertising campaigns. The dependent variable takes the value 1 if a campaign is recognized and 0 otherwise. The mixture variables describing each campaign are proportions of the total advertising exposure (A) in magazine (x_1) and on television (TV) (x_2), which make up 100% for every campaign. There are differences in the advertising campaigns across regions, where these differences are in the total exposure to advertising as well as in the proportions across TV and magazines. For each respondent, we know in which of the regions she lives. As a measure of the advertising campaign exposure, we use Gross Rating Points (GRPs), where a GRP is defined as a percentage of the target audience reached by a campaign (De Pelsmacker et al., 2010).

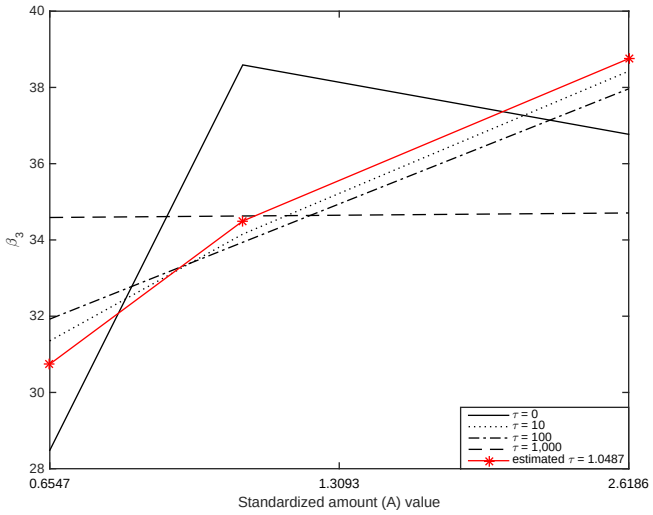
There are 52 advertising campaigns in the data set in total corresponding to 52 unique total amount values. We provide the histogram of the total amount values in our data set in Figure 3, where, for ease of comparison, we give both original and standardized amount values on the lower and upper axes, respectively. As we can see, the share of the ads with less than 300 GRPs is the



(a) $\beta_1(\vec{A})$



(b) $\beta_2(\vec{A})$



(c) $\beta_3(\vec{A})$

Figure 2: Posterior means for the mixture parameters as a function of the standardized amount for different values of τ

largest in our sample, whereas we have much fewer ads with more than 500 GRPs. From Figure 4, we can see the proportions of magazine advertising versus the total amount values (original amount values on the bottom axis and standardized amount values on the top axis). Notice that, in our data set, for large(r) total amount values, the magazine advertising tends to be low(er), and more of the total advertising exposure tends to be invested in TV advertising. Furthermore, there are no observations with a proportion of magazine advertising between 40% and 99%.

The advertising campaigns ran in magazines and/or on TV in the period of June-December, 2011, in the Netherlands and two regions of Belgium, Flanders and Wallonia. For selected campaigns, consumer responses were recorded by means of an online survey at 5 different points in time (the so-called waves). There are approximately 500 respondents per wave and per region. In Flanders, campaigns comprise 4 brands, and there is a total of $9,490$ individual observations. There are 4 brands in Wallonia ($7,786$ individual responses) and 6 brands in the Netherlands ($9,509$ individual responses). Note that, for some brands, there were multiple campaigns. In total, there are $26,785$ responses from $6,679$ respondents in our data set. We provide a subset of the data set in Table 4. Note that our data are somewhat restrictive as each campaign ran in a specific region, for one brand, at one time point only. Within every campaign, there is no variation in the media mix. More information about the data can be found in Aleksandrovs et al. (2015), who use the ads run in Belgium to introduce mixture-amount modeling in the advertising literature.

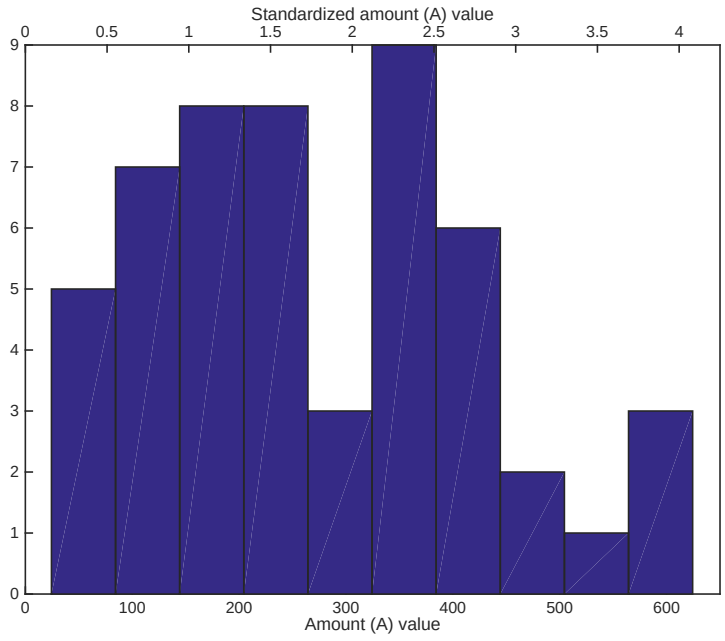


Figure 3: Histogram of the amount values in the advertising campaign data set

Parameter estimation

In this section, we first consider the second-order Scheffé model for the mixture variables (see Equation (2)). As additional control variables, we include dummy variables for the region, brand and wave. The model for the latent variables driving the campaign recognition of the respondents

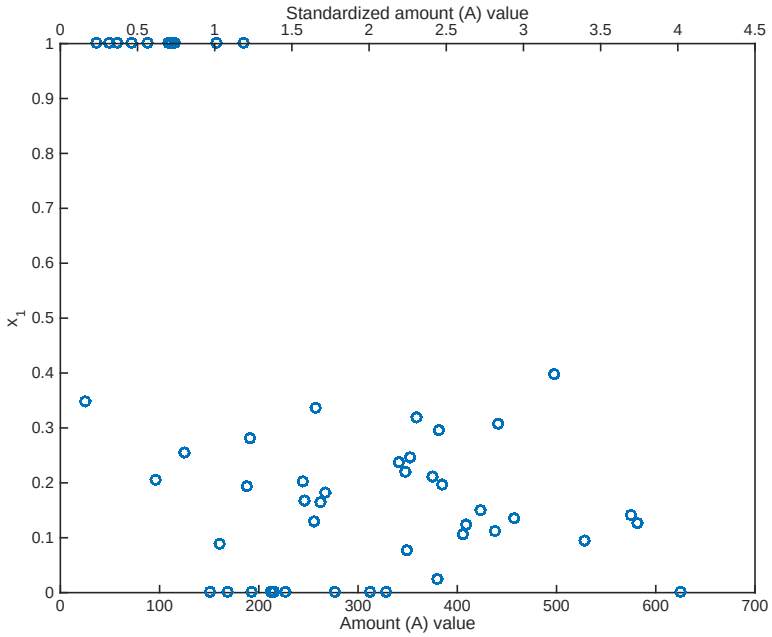


Figure 4: Scatter plot of the proportions of magazine advertising (x_1) and the total amount (A) values

can then be written as

$$\mathbf{z} = \mathbf{X}\text{vec}\big(\mathbf{B}(\vec{A})\big) + \mathbf{D}_{\text{Region}}\boldsymbol{\beta}_{\text{Region}} + \mathbf{D}_{\text{Brand}}\boldsymbol{\beta}_{\text{Brand}} + \mathbf{D}_{\text{Wave}}\boldsymbol{\beta}_{\text{Wave}} + \boldsymbol{\varepsilon}, \tag{28}$$

where we define $\mathbf{X}$ and $\text{vec}\big(\mathbf{B}(\vec{A})\big)$ as in Equation (11) with $\mathbf{x}_i = (x_{1i}, x_{2i}, x_{1i}x_{2i})'$ and $\mathbf{D}_{\text{Region}}$, $\mathbf{D}_{\text{Brand}}$, $\mathbf{D}_{\text{Wave}}$ are matrices with dummy coded columns which correspond to the observations' region, brand and wave, respectively, and $\boldsymbol{\beta}_{\text{Region}}$, $\boldsymbol{\beta}_{\text{Brand}}$ and $\boldsymbol{\beta}_{\text{Wave}}$ are the corresponding vectors of, what we call, non-mixture parameters. The reference region in the model is Wallonia. The reference brand is brand 6 and the reference wave is wave 5.

Following the discussion in Section 4.4, we standardize the amount variable and use $\mathcal{U}(0.75, 2)$ as a prior distribution for τ (note the lower bound that we impose on τ). We use $\mathcal{W}^{-1}(3\times\mathbf{I}_{3\times3}, 7)$ as a prior distribution for $\boldsymbol{\Phi}$, $\mathcal{N}(\mathbf{0}_{11\times1}, 10\times\mathbf{I}_{11\times11})$ as a prior distribution for $(\boldsymbol{\beta}'_{\text{Region}}\ \boldsymbol{\beta}'_{\text{Brand}}\ \boldsymbol{\beta}'_{\text{Wave}})'$ (that is, we use $u = 10$, see Section 4) and a similar distribution for $\mathbf{b}$. Since our dependent variable is a 0/1 variable, we formulate the problem as a choice model and therefore set $\sigma^2 = 1$ (see Section 4.3). Note also that 7 degrees of freedom of the prior distribution of $\boldsymbol{\Phi}$ is the minimum required for the expected value and variance of $\boldsymbol{\Phi}$ to exist, and the prior of $\boldsymbol{\Phi}$ implies $\text{E}(\boldsymbol{\Phi}) = \mathbf{I}_{3\times3}$, where $\mathbf{I}_{3\times3}$ denotes an identity matrix.

As initial values, we use $\tau_{\text{init}} = 1$, $\boldsymbol{\Phi}_{\text{init}} = \mathbf{I}_{3\times3}$, $(\boldsymbol{\beta}'_{\text{Region}}\ \boldsymbol{\beta}'_{\text{Brand}}\ \boldsymbol{\beta}'_{\text{Wave}})' = \mathbf{0}_{11\times1}$ and $\mathbf{b} = \mathbf{0}_{3\times1}$. We set κ in Equation (21) to 0.2, resulting in an acceptance rate of 42.09% in Equation (22),

Campaign	Wave	Region	Brand	ID	GRP_{MAG} (x_1)	GRP_{TV} (x_2)	GRP	Recognition
1	1	Netherlands	2	4791	31.70 (0.25)	92.80 (0.75)	124.50	0
1	1	Netherlands	2	4796	31.70 (0.25)	92.80 (0.75)	124.50	0
1	1	Netherlands	2	4787	31.70 (0.25)	92.80 (0.75)	124.50	1
1	1	Netherlands	2	4810	31.70 (0.25)	92.80 (0.75)	124.50	1
2	1	Netherlands	2	4810	48.70 (0.18)	218.60 (0.82)	267.30	1
...	...	...	...	...	...	...		
12	2	Wallonia	2	2160	86.60 (0.34)	170.00 (0.66)	256.60	0
13	2	Wallonia	5	2160	135.00 (0.31)	305.80 (0.69)	440.80	1
...	...	...	...	...	...	...		
31	3	Netherlands	6	6530	108.10 (1.00)	0.00 (0.00)	108.10	0
...	...	...	...	...	...	...		
51	1	Flanders	4	35613	19.60 (0.21)	76.00 (0.79)	95.60	1
52	5	Flanders	4	6261	0.00 (0.00)	276.50 (1.00)	276.50	0

Table 4: Data for advertising campaign recognition

which is close to the suggested target in Robert and Casella (2010). We use $20,000$ iterations in the estimation and subsequently disregard $10,000$ samples as a burn-in. To show convergence, we plot the Markov chain for τ in Figure 5. The posterior mean, standard deviation and 95% HPD interval for τ are 0.87, 0.10 and [0.75 1.07], respectively. To demonstrate the implied correlation structure for the mixture parameters across different amounts at the posterior mean of τ, we plot the correlation versus the differences in the standardized amount values in Figure 6. The circles denote the implied correlation values at the smallest observed distance (0.0026) and largest observed distance (3.9105) in our data set. In Table 5, we give the posterior estimates of $\boldsymbol{\beta}_{\text{Region}}$, $\boldsymbol{\beta}_{\text{Brand}}$ and $\boldsymbol{\beta}_{\text{Wave}}$. From Table 5 we see that, when controlling for an ad configuration, brand

and wave, on average, women from Wallonia recognize a larger number of campaigns than their Dutch or Flemish counterparts. Further, a larger number of ads is recognized in wave 5 than in other waves, all other covariates in Equation (28) being equal. Finally, the posterior mean of $\mathbf{\Phi}$ is

$$\begin{pmatrix} 2.83 & -0.07 & -2.24 \\ -0.07 & 0.60 & 0.26 \\ -2.24 & 0.26 & 3.73 \end{pmatrix}.$$

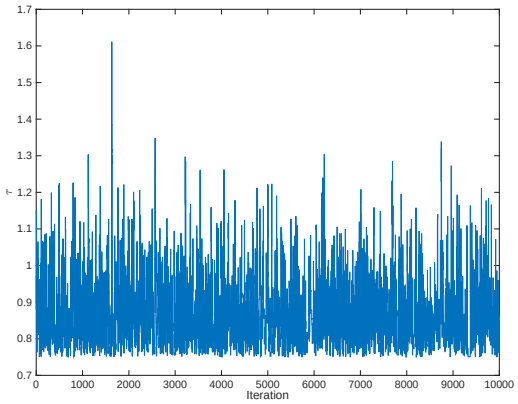


Figure 5: Posterior τ draws after a burn in of $10,000$ observations for the advertising campaign data set

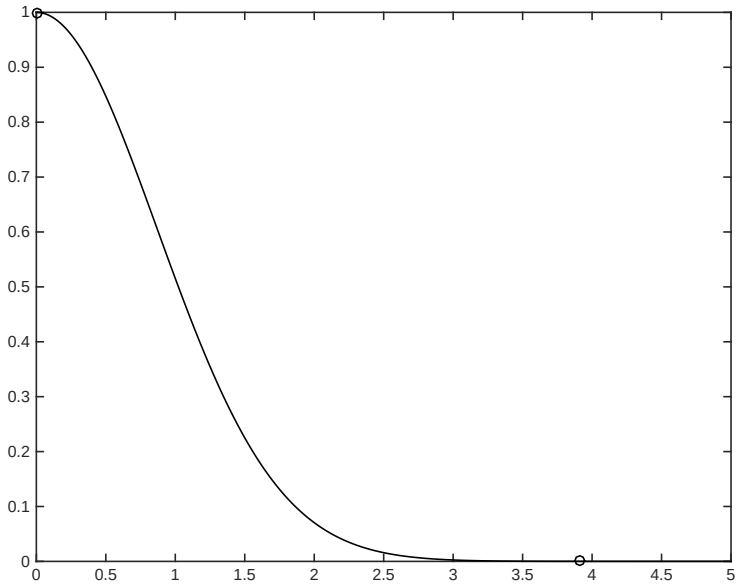


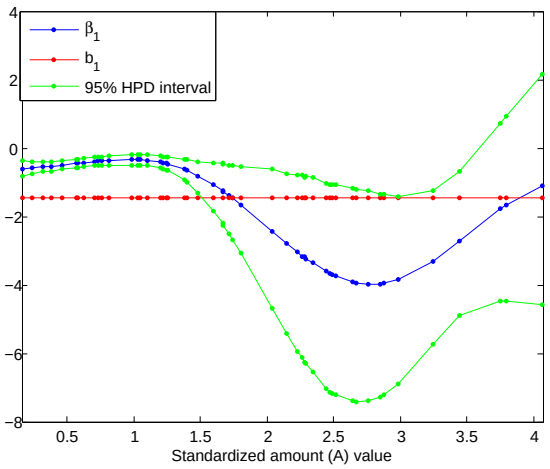
Figure 6: Implied correlation at the posterior mean of τ versus the difference in the standardized amount. The circles denote implied correlation at the smallest and largest observed differences in the advertising campaign data set

$\boldsymbol{\beta_D}$	j	$\mathbb{E}(\boldsymbol{\beta_D}\|\mathbf{y})$	StDev$(\boldsymbol{\beta_D}\|\mathbf{y})$	95% HPD interval	
$\boldsymbol{\beta}_{\text{Region}_j}$	Flanders	−0.17	0.02	−0.22	−0.13
	Netherlands	−0.21	0.03	−0.27	−0.15
$\boldsymbol{\beta}_{\text{Brand}_j}$	1	0.04	0.04	−0.04	0.11
	2	−0.58	0.06	−0.69	−0.46
	3	−0.08	0.03	−0.15	−0.02
	4	0.29	0.03	0.23	0.35
	5	−0.07	0.08	−0.21	0.08
$\boldsymbol{\beta}_{\text{Wave}_j}$	1	−0.18	0.04	−0.25	−0.12
	2	−0.28	0.03	−0.34	−0.21
	3	−0.14	0.03	−0.21	−0.08
	4	−0.10	0.04	−0.18	−0.03

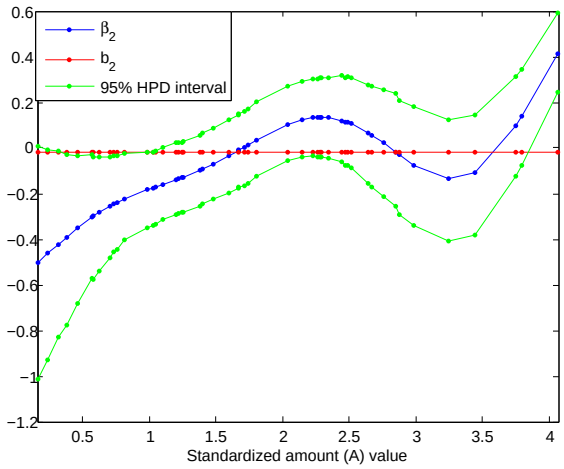
Table 5: Posterior means, standard deviations and 95% HPD intervals for the non-mixture parameters for the advertising campaign data set

In Figure 7, we plot the posterior means of $\beta_1(\vec{A})$, $\beta_2(\vec{A})$ and $\beta_3(\vec{A})$ (blue inner curves) together with 95% HPD intervals (green curves above and below) versus the standardized amount. The horizontal red lines correspond to the posterior means of b_1, b_2 and b_3. The dots on the curves denote the observed amount values in our data set. We see that the effects of both magazine and TV advertising and also the effect of the interaction of the two advertising media vary smoothly with respect to the total advertising exposure. Note that none of the shapes among $\beta_1(\vec{A})$, $\beta_2(\vec{A})$ and $\beta_3(\vec{A})$ is linear or quadratic, which are the functions commonly assumed in standard mixture-amount models in the literature. Furthermore, they are different for different mixture variables. Importantly, the values of $\beta_1(\vec{A})$, $\beta_2(\vec{A})$ and $\beta_3(\vec{A})$ are not constant and differ substantially from the means $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{b}_3$, which demonstrates that the effect of mixture proportions is not constant with respect to the amount. As the data contain a larger number of campaigns with small GRPs (see Figure 3), the uncertainty depicted by the green curves corresponding to the 95% HPD intervals is the smallest for observations at lower GRPs. The lack of both campaigns with larger GRPs and explanatory variables which vary over the campaigns in the data set lead to somewhat wider HPD intervals, especially for the campaigns with larger GRPs.

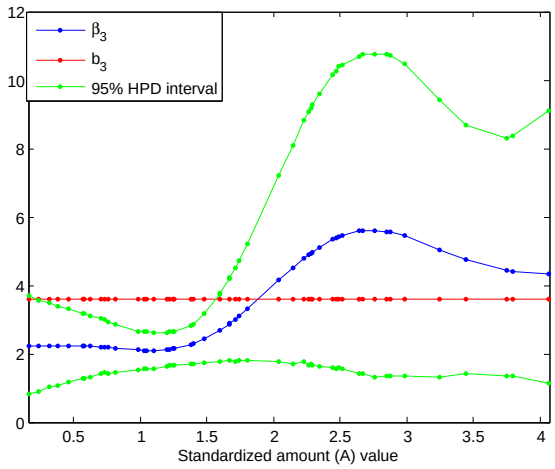
In Figure 8, using the estimated model in Equation (28), we demonstrate how the probability of the campaign recognition changes for different values of the total advertising exposure and the advertising media mix, at the average values for the region, brand and wave. Using the posterior distribution of the parameters, we calculate the probability of recognizing a campaign for values of the standardized amount ranging from 0.15 to 4.1 and the proportion of magazine advertising ranging from 0.1 to 1. The ranges of the amount and magazine proportion match the range in the observed data. To obtain the posterior distribution for the mixture parameters at amount values that are not used in the data, we use Equation (7). As expected, we see that the largest



(a) $\beta_1(\vec{A})$ (Magazines)



(b) $\beta_2(\vec{A})$ (TV)



(c) $\beta_3(\vec{A})$ (Magazines×TV)

Figure 7: Posterior means for the mixture parameters versus the standardized amount

recognition probability is achieved for the largest total advertising exposure values. The large dip in the probability of recognition, at high proportions of magazine advertising and values of the standardized amount of about 3, can be explained by the following. First, as it can be seen from Figure 3, our data set does not contain many campaigns where the total advertising exposure is around 480 GRPs (standardized amount around 3). Furthermore, for such large(r) advertising exposure values, the observed proportion of magazine advertising is always low, see Figure 4. The result of this is that the estimation uncertainty around the recognition probabilities is quite large in this range. To avoid clutter, we do not show this estimation uncertainty in Figure 8. The dip itself is explained by the fact that, in the data set, the campaign recognition was very low at relatively similar observations. An interesting observation is that, when the total advertising exposure is low(er), to maximize the recognition, it seems to be wiser to invest more in magazine advertising. On the other hand, when the total advertising exposure is large, to maximize the

campaign recognition, it is wiser to invest more in TV advertising. This information is important when choosing between advertising media for a given advertising exposure. Then, this type of graph lends itself to evaluating tradeoffs when deciding how much of an advertising budget to allocate in total and per medium.

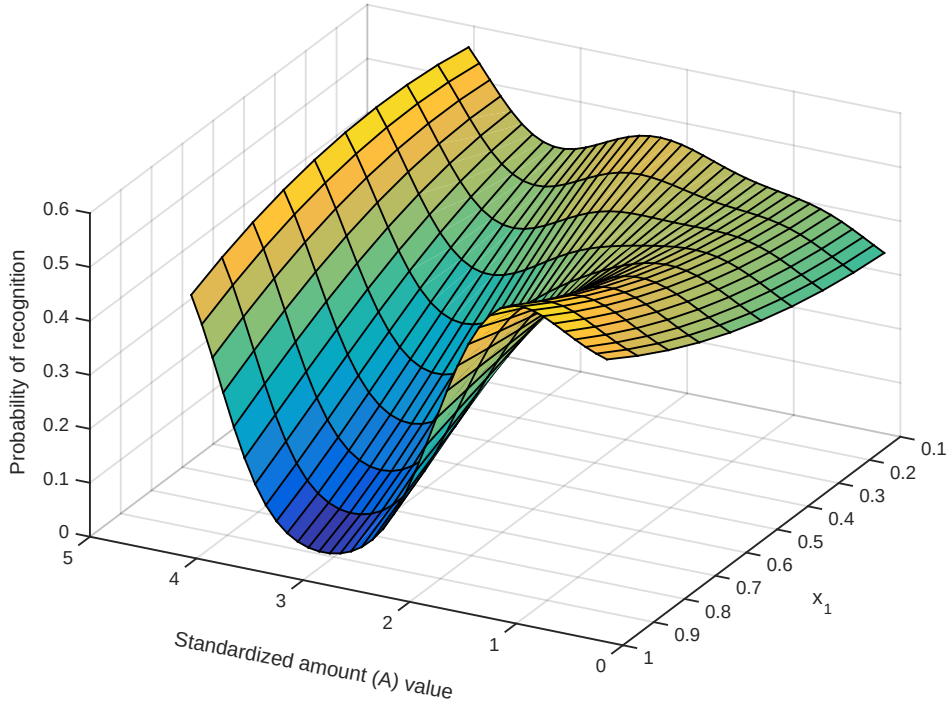


Figure 8: Campaign recognition probabilities for different values of the standardized amount and the proportion invested in magazine advertising (x_1) for the model in Equation (28)

Forecasting performance

In this section, we demonstrate the performance of our model in terms of forecasting recognition for amount values that are not observed. We show that the forecasting performance of our model is superior to that of commonly used mixture-amount models, which all assume that the mixture parameters are parametric functions of the amount. The relative performance of benchmark models deteriorates substantially when more amounts are omitted from the data, but this is much less so for our model. Our model performs best even when half of the observed amount values are omitted, which proves its attractiveness not only for learning the dependence of the mixture parameters on the amount but also for forecasting their values at new amounts.

For this comparison, we consider five parametric models. All these models assume the second-

order Scheffé model for the mixture ingredients, which was introduced in Equation (2), that is,

$$z_i = \beta_1(A)x_{1i} + \beta_2(A)x_{2i} + \beta_{12}(A)x_{1i}x_{2i} + \sum_{j=1}^{2} \beta_{\text{Region}_j} d_{\text{Region}_j,i} + \sum_{j=1}^{5} \beta_{\text{Brand}_j} d_{\text{Brand}_j,i} + \sum_{j=1}^{4} \beta_{\text{Wave}_j} d_{\text{Wave}_j,i} + \varepsilon_i, \quad (29)$$

where $d_{\text{Region}_j,i}$, $d_{\text{Brand}_j,i}$ and $d_{\text{Wave}_j,i}$ are dummy variables which equal one if observation i comes from, respectively, region, brand and wave j, and β_{Region_j}, β_{Brand_j} and β_{Wave_j} are the corresponding parameters. The benchmark models differ in the specification of the dependence of the mixture parameters on A. We consider *linear*, *quadratic* and *cubic* functions. The three specifications are

$$\beta_m(A) = \beta_m^0 + \beta_m^1 A, \quad m \in \{1, 2, 12\}, \quad (30)$$

$$\beta_m(A) = \beta_m^0 + \beta_m^1 A + \beta_m^2 A^2, \quad m \in \{1, 2, 12\}, \quad (31)$$

and

$$\beta_m(A) = \beta_m^0 + \beta_m^1 A + \beta_m^2 A^2 + \beta_m^3 A^3, \quad m \in \{1, 2, 12\}. \quad (32)$$

The two final models we include in our comparison are

$$z_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i}x_{2i} + \alpha_1 A + \sum_{j=1}^{2} \beta_{\text{Region}_j} d_{\text{Region}_j,i} + \sum_{j=1}^{5} \beta_{\text{Brand}_j} d_{\text{Brand}_j,i} + \sum_{j=1}^{4} \beta_{\text{Wave}_j} d_{\text{Wave}_j,i} + \varepsilon_i \quad (33)$$

and

$$z_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i}x_{2i} + \alpha_1 A + \alpha_2 A^2 + \sum_{j=1}^{2} \beta_{\text{Region}_j} d_{\text{Region}_j,i} + \sum_{j=1}^{5} \beta_{\text{Brand}_j} d_{\text{Brand}_j,i} + \sum_{j=1}^{4} \beta_{\text{Wave}_j} d_{\text{Wave}_j,i} + \varepsilon_i, \quad (34)$$

where we assume that the amount does not affect the mixture parameters but only causes a constant change in the response (see Equation (6)). Note that the models in Equations (30)-(34) have in total 17, 20, 23, 15 and 16 parameters, respectively, that need to be estimated, while our model involves estimating $\mathbf{b}$, τ and $\mathbf{\Phi}$, hence, 10 parameters, to estimate the distribution of the mixture parameters and 11 non-mixture parameters. Therefore, in our model, the total number of parameters to be estimated depends only on the number of mixture ingredients and does not vary with the smoothness of the mixture parameters with respect to the amount.

Our aim is to compare the predictive performance of our model to that of the benchmark

models given in Equations (30)-(34), for campaign recognition at amount values that are omitted during the estimation. To this end, we split the sample into an estimation and a test sample. We compute forecasts for our model by first calculating the posterior predictive distribution, as in Equation (7), for the mixture parameters at amount values in the test sample and then combining with the posterior distribution for the non-mixture parameters (see the model in Equation (28)). We also estimate the models in Equations (30)-(34) using Bayesian methodology and use the posterior distributions (10,000 draws) of their parameters for forecasting at amount values in the test sample. As a forecasting performance measure, we use the aggregate mean squared error (MSE), which is defined as

$$\text{MSE} = \frac{1}{n}\sum_{j=1}^{n}(\hat{p}_j - p_j)^2,$$

where $\hat{p}_j - p_j$ is the difference between the predicted proportion of recognition for campaign j ($\hat{p}_j$) and the actual proportion (p_j), averaged across all n campaigns in the test sample.

We investigate the forecasting performance using k-fold cross validation with $k = 52/\{4, 13, 26\} = \{13, 4, 2\}$ over the amount values, where we take 4, 13 or 26 *consecutive* amount values for each test sample. We do so in order to withhold parts of the amounts from the estimation, which makes it more difficult for the models to estimate the pattern of the mixture parameters with respect to the amount. In each round, one of the k subsamples is retained as the test sample and the remaining $k-1$ subsamples are used as the training sample. We calculate the MSE values for all test samples and average them across the k repetitions. We provide the resulting MSE values for our model (GP) and the benchmark models in Table 6.

From Table 6, we can see that our model performs best for all cases in the k-fold cross validation exercise. For each test set size, it leads to smaller MSEs than the benchmark models. When four amount values are omitted, the MSE value of our model is roughly four times smaller (better) than those of the benchmark models. When we omit 26 amount values (which corresponds to half of the amount values observed in the sample), only one of the benchmark models (model in Equation (33)) performs similarly to our model. Note that omitting half of the observed amount values is extreme and is considered here to only evaluate how the models' performances deteriorate if more amount values are held out from the estimation sample. In this case, the forecasting performance of our model stays best. Of the benchmark models, the models with fewer parameters tend to perform better (but not better than our model).

In practice, one could use the Gaussian process prior model to estimate the mixture parameters with respect to the amount variable to obtain some intuition about the possible parametric form. If the mixture parameters resemble a known function in the amount, one may impose it and estimate a standard model like we demonstrated above. This will most likely lead to an improved

fit. Then, there is no need for guessing parametric forms for the mixture parameters with respect to the amount and/or formally testing which model fits best.

test set size	GP	linear (Equation (30))	quadratic (Equation (31))	cubic (Equation (32))	A only (Equation (33))	A and A^2 (Equation (34))
4	0.0107	0.0430	0.0419	0.0437	0.0421	0.0421
13	0.0141	0.0480	0.0508	0.0588	0.0298	0.0422
26	0.0281	0.0417	0.1430	0.1451	0.0282	0.0312

Table 6: MSE values for our model based on Gaussian process (GP) and five benchmark models' forecasts in k-fold cross validation when omitting $4, 13$ or 26 consecutive amounts

6 Conclusion

In this paper, we introduced a new flexible but parsimonious model for mixture-amount data. The current approach to model this kind of data involves strong parametric assumptions for the functional form relating the mixture parameters to the total amount variable. Furthermore, when a flexible parameterization is used in the traditional approach, there are many parameters to estimate. The model that we developed does not require any parametric assumptions concerning the relation between the mixture parameters and the amount. Moreover, there is only one parameter that describes how the mixture parameters vary with respect to the total amount.

Our model is based on so-called Gaussian processes and avoids the necessity to a-priori specify the shape of the dependence of the mixture parameters on the amount. The Gaussian process is used as a prior on the amount-specific mixture parameters. This prior specifies the correlation between the mixture parameters at different amount values. The strength of this correlation controls the variation of the mixture parameters across different amounts.

We demonstrate that our model outperforms standard models from the literature. As we argue, a parametric function relating the mixture parameters to the amount variable is never known a-priori. As a result, we can never be certain whether that parametric assumption is correct. Therefore, in the traditional approach, model comparison and testing procedures are required to choose the final model. This is not needed for the model proposed here. Our modeling approach turns out to be useful to obtain insights in the response to mixture ingredients as well as amounts and has a very good predictive performance.

References

J. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

L. Aleksandrovs, P. Goos, N. Dens, and P. De Pelsmacker. Mixed-media modeling may help optimize campaign recognition, brand interest: How to apply the "Mixture-amount modeling" method to cross-platform effectiveness measurement". *Journal of Advertising Research*, 55(4): 443–457, 2015.

G. Allenby and P. Rossi. Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89(1):57–78, 1999.

A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications*. Springer, 2003.

C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

E. Bonilla, K. Ming, A. Chai, and Ch. Williams. Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems 20*, pages 153–160. MIT Press, 2007.

Ph. Boyle and M. Frean. Dependent Gaussian processes. In *Advances in Neural Information Processing Systems 17*, pages 217–224. MIT Press, 2005.

G. Casella and E. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3): 167–174, 1992.

S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

P. Claringbold. Use of the simplex design in the study of the joint action of related hormones. *Biometrics*, 11(2):174–185, 1955.

J.A. Cornell. *Experiments with Mixtures*. John Wiley & Sons, Inc, 2002.

P. Danaher, T. Dagger, and M. Smith. Forecasting television ratings. *International Journal of Forecasting*, 27(4):1215–1240, 2011.

P. De Pelsmacker, M. Geuens, and J. Van Den Bergh. *Marketing Communications: A European Perspective*. Financial Times Management, 2010.

D. Duvenaud, J. Lloyd, R. Grosse, J. Tenenbaum, and Z. Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1166–1174, 2013.

J. Gattiker, M. Hamada, D. Higdon, M. Schonlau, and W. Welch. Using a Gaussian process as a nonparametric regression model. *Quality and Reliability Engineering International*, 2015.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

S. Gilmour and P. Goos. Analysis of data from non-orthogonal multistratum designs in industrial experiments. *Journal of the Royal Statistical Society, Ser. C*, 58(4):467–484, 2009.

E. Greenberg. *Introduction to Bayesian Econometrics*. Cambridge University Press, 2014.

N. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in Neural Information Processing Systems*, 16(3):329–336, 2004.

T. Leonard. Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society, Ser. B*, 40(2):113–146, 1978.

R. McCulloch and P. Rossi. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1-2):207–240, 1994.

R. McCulloch and P. Rossi. Bayesian analysis of the multinomial probit model. In R. Mariano, T. Schuermann, and M. Weeks, editors, *Simulation-Based Inference in Econometrics*. Cambridge University Press, New York, 2000.

A. Melkumyan and F. Ramos. Multi-kernel Gaussian processes. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pages 1408–1413. AAAI Press, 2011.

R. Neal. *Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification*. Technical Report No. 9702, Dept. of Statistics, University of Toronto, 1997.

R. Neal. Regression and classification using Gaussian process priors. In J. Bernardo, J. Berger, A. Dawid, and A. Smith, editors, *Bayesian Statistics 6*, pages 475–501. Oxford University Press, 1999.

G. Piepel and J. Cornell. Models for mixture experiments when the response depends on the total amount. *Technometrics*, 27(3):219–227, 1985.

C. E. Rasmussen and Ch. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

J. Riihimäki and A. Vehtari. Laplace approximation for logistic Gaussian process density estimation and regression. *Bayesian Analysis*, 9(2):425–448, 2014.

Ch. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2010.

H. Sahrmann, G. Piepel, and J. Cornell. In search of the optimum Harvey Wallbanger recipe via mixture experiment techniques. *The American Statistician*, 41(3):190–194, 1987.

T. Salimans. Variable selection and functional form uncertainty in cross-country growth regressions. *Journal of Econometrics*, 171(2):267–280, 2012.

H. Scheffé. Experiments with mixtures. *Journal of the Royal Statistical Society, Ser. B*, (20): 344–360, 1958.

H. Scheffé. The simplex-centroid design for experiments with mixtures. *Journal of the Royal Statistics Society, Ser. B*, 25(2):235–263, 1963.

K. Train. *Discrete Choice Models with Simulation*. Cambridge University Press, 2009.

J. Wang, D. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298, 2008.

C. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. Jordan, editor, *Learning in Graphical Models*, pages 599–621. MIT Press, 1999.

A. G. Wilson and R. P. Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning (ICML)*, 2013. Oral Presentation.

A. Zellner. *An Introduction to Bayesian Inference in Econometrics*. Wiley-Interscience, 1996.