Mathematical Statistics
Stockholm University

Statistical Species Identification

Måns Karlsson
Ola Hössjer

Research Report 2019:7

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se

Mathematical Statistics
Stockholm University
Research Report **2019:7**,
http://www.math.su.se

Statistical Species Identification

Måns Karlsson and Ola Hössjer

Juni 2019

Abstract

Identification of taxa can be significantly assisted by statistical classification in two major ways. First, one may use a statistical model to determine taxon of subjects based on various characteristics or traits. Secondly, when faced with a collection of subjects with common traits measured, one may determine combinations of traits that signify each taxon in question. To this end, we present a general Bayesian approach to classification of observations based on traits, whose measurements follow some (latent) multivariate Gaussian distribution, but might be truncated or even missing and allow for the traits to depend on covariates. It is inspired by liability threshold modelling and Bayesian Quadratic Discriminant Analysis. The approach is paired with two decision rules: one for which classification is forced, and one that allows for uncertainty of classification, including all categories whose posterior probability ratio, compared to the most likely taxon, exceeds a given threshold. Both of these decision rules are evaluated using blockwise Gibbs sampling. Then we show how the reward function corresponding to these two rules can be used for model selection in terms of blockwise cross validation. Finally, we examplify our approach on a data set over four morphologically similar *Acrocephalus*-genus warblers.

1 Introduction

Modern approaches to classification, which can be used for prediction, often utilize machine learning methods, such as Neural Networks (Goodfellow et al., 2016), Support Vector Machines (Cristianini et al., 2000) or Gaussian Processes (Rasmussen and Williams, 2006). This is for good reason; they are highly flexible modelling tools geared towards prediction, the main task of machine learning. A few, albeit not omnipresent, drawbacks with the aforementioned methods are the general need for large data sets, access to large computing power and lack of parameter estimate interpretability. Sometimes, prediction is not the sole interest of a researcher or technician, as it is coupled with an interest in learning about the studied subjects. For the latter purpose, interpretability is instrumental.

The main idea of this paper is to provide a general modelling approach that is a classifier while still being informative about the study subjects, under as many types of imperfect observation as possible, and with controllable caution in the decision making. To achive this, we employ the idea of liability threshold modelling (LTM), where categorical responses are viewed as discrete approximations of a latent liability with a (multivariate) Gaussian distribution (Albert and Chib, 1993), and pair it with a rather flexible decision rule, while also accounting for missing data in various ways. Thereby, we fill a gap between the basic classification procedure of LTM and the efficient but black-box-like machine learning methods. In particular, pairing LTM with flexible decision rules enables new types of models to be fitted, which are related to Bayesian Quadratic Discriminant Analysis (Srivastava et al., 2007). Although the Gaussian distribution is not neccessarily latent in our setting, it simplifies computations.

Our focus is on situations where we have access to data where the categories of the observations are known. This is often referred to as supervised learning (Russell and Norvig, 2016). Thus, our modelling approach is especially suitable to develop a fast and cheap kind of classification, which can complement a more expensive, but more accurate method, such as consulting an expert.

In Sections 2 and 3, we will set up our classification model for the case where we observe continuous trait measurement vectors and then extend it to the case of incomplete or truncated data, where our trait measurement vector may contain a mixture of continuous, integer-valued and ordered categorical values. Essentially, this entails going from an observable to a latent Gaussian distribution of the traits. It turns out that while making this transition, we can introduce a unified approach to handling missing values and the various types of trait measurements, using the truncation concept in various ways.

In order to encompass uncertainty of classification we introduce set-valued reward functions, defined for all possible subsets of categories. In Section 4, an intuitive reward function, which leads to a classifier that simply picks the category with the largest posterior probability, is presented at first. Furthermore, we introduce the *indecisive region* Λ, which is the set of trait and covariate values for which our classifier does not single out any particular category. We illustrate this with a second reward function with a tuning parameter ρ, that governs the conservativeness of our classifier, and hence the size of the indecisive region. Another tuning parameter τ is introduced, with which we can restrict the region in which we trust our classifier. Then, in Section 5 we demonstrate how the reward functions can be used for model selection in terms of blockwise cross-validation.

In Section 6, we test run our method on simulated data. Then, in Section 7, we examplify usage of our method on a data set over four bird species, with only truncated observations available, and evaluate our decision rule. We also visualize the indecisive region Λ. It should be noted that this method is applicable to practically any classification problem where the trait measurements of the study subjects are ordered in some way. Population ecology, as an example, relies heavily on correct identification of taxa, i.e. correct classification, and if traits are shared between taxa, classification is typically harder. This is one potential area of application of our method. We conclude the paper with a discussion (Section 8) on possible extentions and generalizations of the presented method. Some of the mathematical details are put into the appendices.

2 Model formulation, complete data

Suppose we have N different categories, contained in the set $\mathsf{N} = \{1, \ldots, N\}$, with prior probabilities $\pi = (\pi_1, \ldots, \pi_N)$. With full data we measure q traits and p covariates of each subject. Let Y_{ijk} be the measurement of trait k for subject j in category i, where $1 \leq i \leq N$, $1 \leq j \leq n_i$, $1 \leq k \leq q$ and n_i is the number of subjects in category i. We assume that

$$Y_{ij} = (Y_{ij1}, \ldots, Y_{ijq}) \sim \mathrm{N}\left(m_{ij}, \mathbf{\Sigma}_{ij}\right)$$

are independent random vectors having a multivariate normal distribution, with

$$m_{ij} = (m_{ij1}, \ldots, m_{ijq}) \qquad \text{and} \qquad \mathbf{\Sigma}_{ij} = (\Sigma_{ijkl})_{k,l=1}^{q}$$

being the mean vector and the covariance matrix of subject j of category i. Let also

$$x_{ij} = (1, x_{ij1}, \ldots, x_{ijp}) = (x_{ijm})_{m=0}^{p}$$

be the covariate vector of subject j of category i. Trait vectors and covariate vectors of category i are rows in the matrices $\mathbf{Y}_i = \left(Y_{i1}^\top, \ldots, Y_{in_i}^\top\right)^\top$ and $\mathbf{X}_i = \left(x_{i1}^\top, \ldots, x_{in_i}^\top\right)^\top$ respectively. We now proceed by formulating a multivariate and multiple regression model

$$\mathbf{Y}_i = \mathbf{X}_i\mathbf{B}_i + \mathbf{E}_i \tag{1}$$

for category i, where $\mathbf{B}_i = (B_{imk}; m = 0, \ldots, p; k = 1, \ldots, q)$ is the regression parameter matrix, whose first row consists of intercepts for the q traits, m_{ij} is the j^{th} row of $\mathbf{X}_i\mathbf{B}_i$, and $\mathbf{E}_i = \left(E_{i1}^\top, \ldots, E_{in_i}^\top\right)^\top$ is an error term matrix with rows $E_{ij} \sim \mathrm{N}(0, \mathbf{\Sigma}_{ij})$.

For use in the construction of a joint prior, and later the derivation of the marginal posterior distributions of the parameters, the vectorized form of our regression model is needed. Denote the operation of appending columns of a matrix by $\mathrm{vec}(\cdot)$ (note that we may do the inverse operation $\mathrm{vec}^{-1}(\cdot)$ on column vectors) and rewrite (1) as

$$\mathbf{U}_i = \mathrm{vec}(\mathbf{Y}_i) = \mathbf{Z}_i\beta_i + \mathrm{vec}(\mathbf{E}_i) \tag{2}$$

with $\beta_i = \mathrm{vec}(\mathbf{B}_i)$. Denoting an identity matrix of rank q with $\mathbf{I}_q$ and using the matrix tensor product $\otimes$,

$$\mathbf{Z}_i = \mathbf{I}_q \otimes \mathbf{X}_i = \begin{pmatrix} \mathbf{X}_i & 0 & \cdots & 0 \\ 0 & \mathbf{X}_i & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{X}_i \end{pmatrix} \tag{3}$$

is a block-diagonal matrix with q blocks along the diagonal.

Now suppose we have A covariance classes $\alpha = 1, \ldots, A$ for category i such that

$$\mathbf{\Sigma}_{ij} = \mathbf{\Sigma}_i^\alpha \quad \text{if } x_{ij} \in \mathcal{X}^\alpha, \tag{4}$$

where $\mathcal{X} = \mathcal{X}^1 \cup \ldots \cup \mathcal{X}^\alpha$ is a disjoint decomposition of the predictor space $\mathcal{X}$. Assuming a prior on each of the columns of $\mathbf{B}_i$, and letting it be $\mathrm{N}\left((b_{ik0}, \ldots, b_{ikp})^\top = b_{ik}, \mathbf{\Sigma}_{\mathbf{B}_i}\right)$ for $k = 1, \ldots, q$, implies the prior $\mathrm{N}\left(\left(b_{i1}^\top, \ldots, b_{iq}^\top\right)^\top = \beta_{i0}, \mathbf{I}_q \otimes \mathbf{\Sigma}_{\mathbf{B}_i} = \mathbf{\Sigma}_{\beta_i}\right)$ on β_i. Further, assuming prior independence and imposing an Inverse-Wishart distribution $\mathbf{\Sigma}_i^\alpha \sim IW(\nu_0, \mathbf{V}_0)$ on the covariance matrices in (4) for $\alpha = 1, \ldots, A$, we get the joint prior

$$p(\beta_i, \mathbf{\Sigma}_i^1, \ldots, \mathbf{\Sigma}_i^A) = p(\beta_i) \prod_{\alpha=1}^{A} p(\mathbf{\Sigma}_i^\alpha) \tag{5}$$

for the parameters of category i.

2.1 Estimation

Let $\theta_i = \left(\mathbf{B}_i, \mathbf{\Sigma}_i^1, \ldots, \mathbf{\Sigma}_i^A\right)$ represent all parameters of category i. In the following, we assume that $\theta_1, \ldots, \theta_N$ are independent random vectors with probability densities $p(\theta_1), \ldots, p(\theta_N)$ defined in (5). Introducing dependencies is of course possible, and may be important for specific problems, for which we refer the reader to Section 8. From Bayes' Theorem we get an aposteriori density

$$\begin{aligned} p(\theta_i \mid \mathcal{D}_i) &= p(\theta_i) C(\mathcal{D}_i) \prod_{j=1}^{n_i} f\left(y_{ij}; x_{ij}, \theta_i\right) \\ &= p(\theta_i) C(\mathcal{D}_i) \mathcal{L}(\theta_i; \mathcal{D}_i) \\ &\propto p(\theta_i) \mathcal{L}(\theta_i; \mathcal{D}_i) \end{aligned}$$

of θ_i given the complete training data set $\mathcal{D}_i = \{(x_{ij}, Y_{ij});\, j = 1, \ldots, n_i\} = \{\mathbf{X}_i, \mathbf{Y}_i\}$ for category i. The function $\mathcal{L}\left(\theta_i; \mathcal{D}_i\right) = p\left(\mathbf{Y}_i \mid \mathbf{X}_i, \theta_i\right)$ is the likelihood. In the last step we removed the normalizing factor $C(\mathcal{D}_i) = \left(\int p(\theta_i) \mathcal{L}(\theta_i; \mathcal{D}_i)\, \mathrm{d}\theta_i\right)^{-1}$, since it does not depend on θ_i. The Maximum Aposteriori (MAP)-estimator of θ_i is

$$\begin{aligned} \theta_i^{(\mathrm{MAP})} &= \arg\max_{\theta_i} p(\theta_i \mid \mathcal{D}_i) \\ &= \arg\max_{\theta_i} p(\theta_i) \mathcal{L}(\theta_i; \mathcal{D}_i), \end{aligned}$$

whereas the Bayes' estimator of θ_i is

$$\begin{aligned} \theta_i^{(\mathrm{Bayes})} &= \mathbb{E}\left[\theta_i \mid \mathcal{D}_i\right] \\ &= \int \theta_i p(\theta_i \mid \mathcal{D}_i)\, \mathrm{d}\theta_i \\ &= C(\mathcal{D}_i) \int \theta_i p(\theta_i) \mathcal{L}(\theta_i; \mathcal{D}_i)\, \mathrm{d}\theta_i. \end{aligned}$$

Finally, given a new observation $\mathcal{D}^{\mathrm{new}} = (x, Y)$, define the posterior probability of the new observation belonging to category i as

$$p_i = \mathbb{P}(\mathrm{I} = i \mid \mathcal{D}, \mathcal{D}^{\mathrm{new}}) = \frac{\pi_i \omega_i}{\pi_1 \omega_1 + \ldots + \pi_N \omega_N}, \tag{6}$$

where

$$\begin{aligned} \omega_i &= \int f(Y; x, \theta_i) p(\theta_i \mid \mathcal{D}_i)\, \mathrm{d}\theta_i \\ &= C(\mathcal{D}_i) \int f(Y; x, \theta_i) p(\theta_i) \mathcal{L}(\theta_i; \mathcal{D}_i)\, \mathrm{d}\theta_i \end{aligned}$$

are the posterior category weights given $\mathcal{D}^{\mathrm{new}}$ for all categories, before the prior probabilities π_i have been taken into account.

2.2 Monte Carlo Approximations

It is usually difficult to evaluate the normalizing constants $C(\mathcal{D}_i)$ for high-dimensional data sets, and hence also $\theta_i^{(\text{Bayes})}$ and ω_i. However, it is possible to estimate $\theta_i^{(\text{Bayes})}$ and ω_i by Monte Carlo simulation, with

$$\hat{\theta}_i^{(\text{Bayes})} = \frac{1}{R_i} \sum_{r=1}^{R_i} \theta_{ir} \tag{7}$$

and

$$\hat{\omega}_i = \frac{1}{R_i} \sum_{r=1}^{R_i} f(Y; x, \theta_{ir}) \tag{8}$$

respectively, if $\theta_{i1}, \ldots, \theta_{iR_i}$ are R_i replicates drawn from the posterior distribution $p(\theta_i \mid \mathcal{D}_i)$, with $\theta_{ir} = \left(\beta_{ir}, \mathbf{\Sigma}_{ir}^1, \ldots, \mathbf{\Sigma}_{ir}^A\right)$.

We will generate $\theta_{i1}, \ldots, \theta_{iR_i}$ by blockwise Gibbs sampling, and for this we need the conditional posterior distributions of β_i and $\mathbf{\Sigma}_i^\alpha$ for $\alpha = 1, \ldots, A$. To derive those, we need some additional notation. Let $\mathbf{Z}_i^\alpha$, $\mathbf{X}_i^\alpha$, $\mathbf{Y}_i^\alpha$ and $\mathbf{U}_i^\alpha$ denote the submatrices of $\mathbf{Z}_i$, $\mathbf{X}_i$, $\mathbf{Y}_i$ and $\mathbf{U}_i$ corresponding to covariance class α. Recall also that $\mathbf{B}_i = \text{vec}^{-1}(\beta_i)$, meaning that we know $\mathbf{B}_i$ if we know β_i, and vice versa. Using this notation, we may express the conditional posterior of the regression parameters

$$\beta_i \mid \mathbf{U}_i, \{\mathbf{\Sigma}_i^\alpha\}_{\alpha=1}^A \sim \text{N}(\tilde{\beta}, \tilde{\mathbf{\Sigma}})$$

where

$$\tilde{\mathbf{\Sigma}} = \left[\mathbf{\Sigma}_\beta^{-1} + \sum_{\alpha=1}^{A} (\mathbf{\Sigma}_i^\alpha)^{-1} \otimes \left((\mathbf{X}_i^\alpha)^\top \mathbf{X}_i^\alpha\right)\right]^{-1},$$

and

$$\tilde{\beta} = \tilde{\mathbf{\Sigma}} \times \left[\mathbf{\Sigma}_\beta^{-1}\beta_0 + \sum_{\alpha=1}^{A} \left((\mathbf{\Sigma}_i^\alpha)^{-1} \otimes (\mathbf{X}_i^\alpha)^\top\right) \mathbf{U}_i^\alpha\right].$$

Meanwhile, the conditional posteriors of the covariance matrices are

$$\mathbf{\Sigma}_i^\alpha \mid \mathbf{B}_i, \mathbf{Y}_i^\alpha, \mathbf{X}_i^\alpha \sim IW(\nu_0 + n_i^\alpha, \mathbf{V}_0 + \mathbf{S}_i^\alpha),$$

where n_i^α denotes the number of observations in the covariance class α for category i and

$$\mathbf{S}_i^\alpha = (\mathbf{Y}_i^\alpha - \mathbf{X}_i^\alpha \mathbf{B}_i)^\top (\mathbf{Y}_i^\alpha - \mathbf{X}_i^\alpha \mathbf{B}_i) + (\mathbf{B}_i - \mathbf{B}_0)^\top \mathbf{\Sigma}_\mathbf{B} (\mathbf{B}_i - \mathbf{B}_0).$$

For the derivation of the marginal posterior distributions and a detailed description of the specific Monte Carlo-algorithm, we refer to Appendices A and B respectively.

Having computed $\hat{\omega}_1, \ldots, \hat{\omega}_N$, for $\mathcal{D}^{\text{new}}$, we may compute the Monte Carlo-estimated aposteriori probability of $\mathcal{D}^{\text{new}}$ being in category i as

$$\hat{p}_i = \hat{\mathbb{P}}(I = i \mid \mathcal{D}, \mathcal{D}^{\text{new}}) = \frac{\pi_i \hat{\omega}_i}{\pi_1 \hat{\omega}_1 + \ldots + \pi_N \hat{\omega}_N},$$

where $\mathcal{D} = \mathcal{D}_1 \cup \ldots \cup \mathcal{D}_N$ is the complete training data set. We will return to these aposteriori probabilities in Section 4.

3 Model Formulation, obfuscated data

Overall our setup is the same as in Section 2, but we now suppose we only have partial information about the complete training data set $\mathcal{D}$. Due to some obfuscation, which could be rounding, grouping, categorization or lost measurements of some traits, we only know that

$$Y_{ij} \in \mathsf{S}_{ij} = \mathsf{S}_{ij1} \times \cdots \times \mathsf{S}_{ijq},$$

i.e. the complete trait vector Y_{ij} for subject j of category i is contained in a hyperrectangle S_{ij}, whose sides are given by $\{\mathsf{S}_{ijk}\}_{k=1}^{q}$. The sides are sets, ranging in possible size from singletons to infinite intervals of $\mathbb{R}$, and are given by

$$\mathsf{S}_{ijk} = \begin{cases} Y_{ijk}, & k \notin \mathsf{K}_{ij}, \\ (c_{ijk}, d_{ijk}], & k \in \mathsf{K}_{ij}, \end{cases}$$

where $\mathsf{K}_{ij} = \{k; 1 \leq k \leq q;\, Y_{ijk} \text{ obfuscated}\}$. The obfuscations are of three main types. First, if a trait Y_{ijk} is unobserved, written as $Y_{ijk} = \texttt{NA}$, the k:th side to S_{ij} is of infinite length; e.g. $c_{ijk} = -\infty$, $d_{ijk} = \infty$, and we let the interval be open. That is, the interval S_{ijk} equals $\mathbb{R}$. Secondly, a trait may be obfuscated in such a way that interval limits are observed. Rounding is a typical example of this; consider a measurement of a trait $y_{ijk} \in \mathbb{R}^+$ that has been rounded to $z_{ijk} \in \mathbb{Z}^+$. We put $c_{ijk} = z_{ijk} - \tau$ and $d_{ijk} = z_{ijk} + \tau$, which constitute the limits of the interval. Generally, we assume rounding to the midpoint of an interval. We can always scale so the interval is of unit length, which would be equivalent to $\tau = 1/2$. Lastly we have the case when we observe an ordered categorical random variable Z_{ijk}. We assume there is an underlying normally distributed variable Y_{ijk}, and that each category z_{ijk} corresponds to an interval of possible values of y_{ijk}. Count data can be treated as a case of this type of obfuscation, i.e. a trait may be measured in counting occurences of something, and we can handle that type of data similarily as ordered categorical data.

We can treat all types of obfuscations uniformly in the following way. Suppose trait k of subject j of category i is imperfectly observed, i.e. $k \in \mathsf{K}_{ij}$. Let g_k be the number of categories of this trait, which we number as $0, 1, \ldots, g_k - 1$. The observed category is $z_{ijk} \in \{0, 1, \ldots, g_k - 1\}$, where $g_k = 2$ for binary data and $g_k = \infty$ for count data. The corresponding side of S_{ij} is

$$\mathsf{S}_{ijk} = \begin{cases} \left(-\infty, \frac{1}{2}\right], & \text{if } z_{ijk} = 0, \\ \left(z_{ijk} - \frac{1}{2}, z_{ijk} + \frac{1}{2}\right], & \text{if } 1 \leq z_{ijk} \leq g_k - 2, \\ \left(g_k - \frac{3}{2}, \infty\right), & \text{if } z_{ijk} = g_k - 1. \end{cases}$$

Here, a useful trick would be to add auxiliary categories, that never were observed, to take the place of $z_{ijk} = 0$ and $z_{ijk} = g_k - 1$. That ensures all observed intervals are of unit length, although we may let intervals vary in length if there is reason to construct such a model. We also write

$$Z_{ijk} = z(\mathsf{S}_{ijk}) = \begin{cases} 0, & \text{if } \mathsf{S}_{ijk} = \left(-\infty, \frac{1}{2}\right], \\ \frac{c_{ijk}+d_{ijk}}{2}, & \text{if } \mathsf{S}_{ijk} \text{ is bounded}, \\ g_k - 1, & \text{if } \mathsf{S}_{ijk} = \left(g_k - \frac{3}{2}, \infty\right], \end{cases}$$

for the center point of a finite or half-open, infinite S_{ijk}, whereas $z(\mathsf{S}_{ijk}) = Y_{ijk}$ when $Y_{ijk} = \mathsf{S}_{ijk}$ is perfectly observed. We may write the actually observed training data set as

$$\mathcal{D}^{\text{obs}} = \{(x_{ij}, \mathsf{S}_{ij}) \,;\, i = 1, \ldots, N, j = 1, \ldots, n_i\}.$$

Finally, we remark on the importance (or lack thereof) of taking rounding into account. Consider rounding a Gaussian trait Y_{ijk} to $z_{ijk} \in \mathbb{Z}$ for some k, $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. Suppose we have no covariates and that $Y_{ijk} \sim N(\mu_k, \sigma_k^2)$ for $j = 1, \ldots, n_i$. An unbiased estimator of μ_k is the average $\bar{Y}$, whereas $\bar{Z}$ is a biased estimator of μ_k. One can quantify the size of the bias using σ_k and the width of the rouding interval $\mathsf{w} = \eta\sigma_k$ (Tricker, 1984). In short, the larger w is relative to σ_k, the larger our bias will become, as measured by η. Already when $\sigma = \mathsf{w} = 1$, the bias of the magnitude 10^{-9}, and hence, unless we need an extremely precise mean estimate, the bias is small compared to the estimate uncertainty. Therefore, one might permit oneself to use rounded values as true values, if the bias is small enough.

3.1 Estimation

Using $\mathcal{D}_i^{\text{obs}} = \{(x_{ij}, \mathsf{S}_{ij}); j = 1, \ldots, n_i\}$, the posterior distribution of θ_i becomes

$$\begin{aligned} p(\theta_i \mid \mathcal{D}_i^{\text{obs}}) &= p(\theta_i) C(\mathcal{D}_i^{\text{obs}}) \prod_{j=1}^{n_i} p(\mathsf{S}_{ij}; x_{ij}, \theta_i) \\ &= p(\theta_i) C\left(\mathcal{D}_i^{\text{obs}}\right) \mathcal{L}(\theta_i; \mathcal{D}_i^{\text{obs}}) \end{aligned} \tag{9}$$

where the normalizing factor $C\left(\mathcal{D}_i^{\text{obs}}\right)$ is

$$\left(C\left(\mathcal{D}_i^{\text{obs}}\right)\right)^{-1} = \int p(\theta_i) \mathcal{L}(\mathcal{D}_i^{\text{obs}}; \theta_i) \, \mathrm{d}\theta_i$$

and

$$p(\mathsf{S}_{ij}, x_{ij}; \theta_i) = \begin{cases} f(Y_{ij}; x_{ij}, \theta_i), & \mathsf{K}_{ij} = \emptyset, \\ \int_{\mathsf{S}_{ij}} f(y_{ij}; x_{ij}, \theta_i) \prod_{k \in \mathsf{K}_{ij}} \mathrm{d}y_{ijk}, & \mathsf{K}_{ij} \neq \emptyset. \end{cases} \tag{10}$$

Thus, with perfect observations ($\mathsf{K}_{ij} = \emptyset$), we evaluate the density of the trait vector f at the observed point Y_{ij} and the model is exactly as specified in Section 2. Otherwise, we construct a $|\mathsf{K}_{ij}|$-dimensional integral over f and the contribution to the likelihood is this integral, with the function evaluated exactly at the remaining perfectly observed traits, if any exist. In particular, if all traits are imperfectly observed, the integral is q-dimensional. We may approximate the integral in (10) by

$$|\mathsf{S}_{ij}| \, f\left(z(\mathsf{S}_{ij}), x_{ij}; \theta_i\right) = \prod_{k \in \mathsf{K}_{ij}} |\mathsf{S}_{ijk}| \cdot f\left(z(\mathsf{S}_{ij1}), \ldots, z(\mathsf{S}_{ijq}), x_{ij}; \theta_i\right)$$

whenever all $|\mathsf{S}_{ijk}| < \infty$ for $k \in \mathsf{K}_{ij}$, which is the case when employing the trick with auxiliary categories.

Since $p(\mathsf{S}_{ij}, x_{ij}; \theta_i)$ potentially contains integrals of a multivariate Gaussian density function and there in general is a lack of a CDF on closed form for this distribution, the integrals need to be solved numerically. However, in the case of $|\mathsf{K}_{ij}| = 1$, with $\mathsf{S}_{ijk} = (c_{ijk}, d_{ijk}]$, the integral is univariate and thus[1]

$$\begin{aligned} p(\mathsf{S}_{ij}, x_{ij}; \theta_i) = f(Y_{ij(-k)}, x_{ij}; \theta_i) \Bigg[&\Phi\left(\frac{d_{ijk} - m_{ijk}(y_{ij(-k)})}{\sigma_{ijk}}\right) \\ &- \Phi\left(\frac{c_{ijk} - m_{ijk}(y_{ij(-k)})}{\sigma_{ijk}}\right) \Bigg], \end{aligned} \tag{11}$$

[1]The notation with subscript $(-k)$ means dropping element k from a vector; dropping row k from a matrix when not being the last index of a matrix; and dropping column k when being the last index.

where $m_{ijk}(y_{ij(-k)}) = m_{ijk} + \boldsymbol{\Sigma}_{ijk(-k)}\boldsymbol{\Sigma}^{-1}_{ij(-k)(-k)}(y_{ij(-k)} - m_{ij(-k)})$ is the conditional expectation of Y_{ijk} given that $Y_{ij(-k)} = (Y_{ijk'};\, k' \neq k) = y_{ij(-k)}$, $\sigma_{ijk} = \sqrt{\boldsymbol{\Sigma}_{ijkk} - \boldsymbol{\Sigma}_{ijk(-k)}\boldsymbol{\Sigma}^{-1}_{ij(-k)(-k)}\boldsymbol{\Sigma}_{ij(-k)k}}$ is the conditional standard deviation of Y_{ijk} given any value of $Y_{ij(-k)}$, and Φ is the CDF of the univariate Gaussian distribution with mean 0 and standard deviation 1.

Using $\mathcal{D}_i^{\text{obs}}$, we find from (9) that the estimators $\theta_i^{(\text{MAP})}$ and $\theta_i^{(\text{Bayes})}$ are

$$\begin{aligned}\theta_i^{(\text{MAP})} &= \arg\max_{\theta_i} p(\theta_i \mid \mathcal{D}_i^{\text{obs}}) \\ &= \arg\max_{\theta_i} p(\theta_i)\mathcal{L}(\mathcal{D}_i^{\text{obs}}; \theta_i)\end{aligned}$$

and

$$\begin{aligned}\theta_i^{(\text{Bayes})} &= \mathbb{E}\left[\theta_i \mid \mathcal{D}_i^{\text{obs}}\right] = \int \theta_i p(\theta_i \mid \mathcal{D}_i^{\text{obs}})\,\mathrm{d}\theta_i \\ &= C\left(\mathcal{D}_i^{\text{obs}}\right)\int \theta_i p(\theta_i)\mathcal{L}(\mathcal{D}_i^{\text{obs}}; \theta_i)\,\mathrm{d}\theta_i \end{aligned} \tag{12}$$

respectively. Furthermore, redefining $\mathcal{D}^{\text{new}} := (x, \mathsf{S})$ for a new observation and observing the corresponding set K, leads to the posterior category weights

$$\begin{aligned}\omega_i &= \iint_{\mathsf{S}} f(y; x, \theta_i) \prod_{k\in\mathsf{K}} \mathrm{d}y_k\, p(\theta_i \mid \mathcal{D}_i^{\text{obs}})\,\mathrm{d}\theta_i \\ &= C(\mathcal{D}_i^{\text{obs}}) \iint_{\mathsf{S}} f(y; x, \theta_i) \prod_{k\in\mathsf{K}} \mathrm{d}y_k\, p(\theta_i)\mathcal{L}(\mathcal{D}_i^{\text{obs}}; \theta_i)\,\mathrm{d}\theta_i \end{aligned} \tag{13}$$

of this observation.

3.2 Monte Carlo Approximations

The integral over S in (13) is, as mentioned in conjunction with (11), potentially impossible to compute analytically, but could also be well behaved. We can in theory approximate $\theta_i^{(\text{Bayes})}$ in (12) as in (7) by sampling θ_i from $p(\theta_i \mid \mathcal{D}^{\text{obs}})$ a total of R_i times. However, this entails a large number of numerical evaluation of integrals, see (9)-(10). Similarly, we may estimate ω_i for $1 \leq i \leq N$ in (13) through

$$\hat{\omega}_i = \frac{1}{R_i}\sum_{r=1}^{R_i} \int_{\mathsf{S}} f(y; x, \theta_{ir}) \prod_{k\in\mathsf{K}} \mathrm{d}y_k, \tag{14}$$

which in addition to previously presented integrals, involves computation of an integral over S. As an alternative way of computing (12) and (14), we also present an approach where complete data is sampled, based on the

obfuscated data, as a step in the Monte Carlo algorithm; the parameters are sampled as another step in the same algorithm. This allows us to estimate all $\theta_i^{(\text{Bayes})}$ and ω_i under widespread obfuscation, given that we are able to sample $\mathbf{Y}_i$, $i = 1, \ldots, N$. Overall, we want to generate

$$\{\theta_{ir}, Y_{ijkr},\, 1 \le j \le n_i,\, k \in \mathsf{K}_{ij}; Y_{kr}, k \in \mathsf{K}\}_{r=1}^{R_i} \tag{15}$$

from

$$p(\theta_i|\mathcal{D}_i^{\text{obs}}) \prod_{j=1}^{n_i} f\left(y_{ij\mathsf{K}_{ij}r} \mid x_{ij}, \mathsf{S}_{ij}, \theta_i\right) f\left(y_{\mathsf{K}r}, \mid x, Y_{\mathsf{K}^\complement}; \theta_i\right)$$

where $\theta_{ir} = \left(\beta_{ir}, \mathbf{\Sigma}_{ir}^1, \ldots, \mathbf{\Sigma}_{ir}^A\right)$, $y_{ij\mathsf{K}_{ij}r} = (y_{ijkr}; k \in \mathsf{K}_{ij})$, $y_{\mathsf{K}r} = (y_{kr}; k \in \mathsf{K})$ and $Y_{\mathsf{K}^\complement} = (Y_k; k \notin \mathsf{K})$. Note that we do not condition on S in the conditional density of the unobserved traits, as this would introduce a bias in the Monte Carlo estimate of ω_i below.

The details of the specific Gibbs sampling approach we use are presented in Appendix B. Having generated a sample $\theta_{i1}, \ldots, \theta_{iR_i}$, we may compute the estimated category weights of $\mathcal{D}^{\text{new}}$ as

$$\hat{\omega}_i = \frac{1}{R_i} \sum_{r=1}^{R_i} f\left(Y_{\mathsf{K}^\complement}; x, \theta_{ir}\right) \mathbb{1}_{\left\{Y_{\mathsf{K}r} \in \bigtimes_{k\in\mathsf{K}} \mathsf{S}_k\right\}}, \tag{16}$$

where $Y_{\mathsf{K}^\complement}$ is as above, and $Y_{\mathsf{K}r} = (Y_{kr}; k \in \mathsf{K})$. For every θ_{ir}, one could generate many Y_{K} and replace the indicator with an average of the indicators for each sampled Y_{K}.

A potentially more efficient method would be to define $\{y_t\}_{t=1}^T$, where $y_t = (y_{t1}, \ldots, y_{tq})$ with $y_{t\mathsf{K}^\complement} = Y_{\mathsf{K}^\complement}$ and $y_{t\mathsf{K}} \in \bigtimes_{k\in\mathsf{K}} \mathsf{S}_k$, in such a way that $\{y_{tk}, k \in \mathsf{K}\}$ is a grid approximation of $\bigtimes_{k\in\mathsf{K}} \mathsf{S}_k$. Then we can estimate ω_i through

$$\hat{\omega}_i = \frac{\prod_{k\in\mathsf{K}} |\mathsf{S}_k|}{TR_i} \sum_{r=1}^{R_i} \sum_{t=1}^{T} f\left(y_t; x, \theta_{ir}\right). \tag{17}$$

If we use the trick with auxiliary categories, we can choose y_t uniformly at random on $\bigtimes_{k\in\mathsf{K}} \mathsf{S}_k$, as long as we do not have any missing observations, since those are represented with an infinite interval. Thus, (17) is potentially more effcient than (16), but comes at a cost of generality, since (16) is applicable to any new observation.

Finally, the Monte Carlo-estimated aposteriori probability of $\mathcal{D}^{\text{new}} = (x, \mathsf{S})$ being of category i is

$$\hat{p}_i = \hat{\mathbb{P}}(I = i \mid \mathcal{D}^{\text{new}}, \mathcal{D}^{\text{obs}}) = \frac{\pi_i \hat{\omega}_i}{\pi_1 \hat{\omega}_1 + \ldots + \pi_N \hat{\omega}_N} \tag{18}$$

and we may apply (13) with replacement of ω_i by (16) for prediction. If (17) is inserted into (18) we notice that $\prod_{k\in\mathsf{K}} |\mathsf{S}_k|$ and T cancel out, and in case $R_i = R$ for $i = 1, \ldots, N$, also R cancels out.

4 Classification

4.1 Classification with at least one category as output

Let $\mathcal{N} = \mathcal{P}(\mathsf{N}) \setminus \emptyset$ denote the collection of all non-empty subsets of N. Let $\mathrm{I} \in \mathsf{N}$ be the true but unknown category of a future observation $\mathcal{D}^{\text{new}} = (x, Y)$ and let $\hat{\mathrm{I}} \in \mathcal{N}$ be a classifier with $|\hat{\mathrm{I}}| \geq 1$. In order to define $\hat{\mathrm{I}}$ we introduce a reward function $\mathcal{N} \times \mathsf{N} \ni (\mathcal{I}, \mathrm{I}) \mapsto R(\mathcal{I}, \mathrm{I})$ for all $\mathcal{I} \in \mathcal{N}$ and put

$$\begin{aligned}\hat{\mathrm{I}} &= \arg\max_{\mathcal{I}} \mathbb{E}\left[R(\mathcal{I}, \mathrm{I}) \mid \mathcal{D}^{\text{obs}}, \mathcal{D}^{\text{new}}\right] \\ &= \arg\max_{\mathcal{I}} \sum_{i=1}^{N} R(\mathcal{I}, i) p_i\end{aligned}$$

as the optimal Bayesian classifier, with $\mathcal{D}^{\text{obs}} = (\mathcal{D}_1^{\text{obs}}, \ldots, \mathcal{D}_N^{\text{obs}})$ as the complete training data set and p_i is as in (6). So, $\hat{\mathrm{I}}$ is the set in $\mathcal{N}$ that maximizes the expected posterior reward. Each classifier $\hat{\mathrm{I}} = \hat{\mathrm{I}}(\mathcal{D}^{\text{obs}}, \mathcal{D}^{\text{new}})$, viewed as a function of $\mathcal{D}^{\text{new}}$, partitions the test data space into decision regions

$$\Omega_{\mathcal{I}} = \{(x, Y); \hat{\mathrm{I}} = \mathcal{I}\}$$

for all $\mathcal{I} \in \mathcal{N}$. This gives rise to an *indecisive region*

$$\Lambda = \bigcup_{|\mathcal{I}|>1} \Omega_{\mathcal{I}},$$

where we cannot distinguish one particular category with acceptable confidence, only eliminate some of the categories with low degree of belief.

There is considerable freedom in choosing the reward function R and we will examine two particular choices. Let

$$R(\mathcal{I}, \mathrm{I}) = \begin{cases} 0; & \mathrm{I} \notin \mathcal{I} \\ 1/|\mathcal{I}|; & \mathrm{I} \in \mathcal{I} \end{cases}$$

which has expected posterior reward

$$\mathbb{E}\left[R(\mathcal{I}, \mathrm{I}) \mid \mathcal{D}^{\text{obs}}, \mathcal{D}^{\text{new}}\right] = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} p_i$$

and optimal classifier

$$\hat{\mathrm{I}} = \{(N)\} = \arg\max_{i} \pi_i \omega_i \tag{19}$$

where $p_{(1)} < \ldots < p_{(N)}$ are the ordered posterior category probabilities. Notice that $\Lambda = \emptyset$, i.e. this reward function leads to a consistent, but potentially overzealous, classifier.

Our second reward function

$$R(\mathcal{I}, \mathrm{I}) = \mathbb{1}_{\{\mathrm{I} \in \mathcal{I}\}} - \rho |\{i \in \mathcal{I}; i \neq (N)\}| p_{(N)}$$

has expected posterior reward

$$\mathbb{E}\left[R(\mathcal{I}, \mathrm{I}) \mid \mathcal{D}^{\text{obs}}, \mathcal{D}^{\text{new}}\right] = \sum_{i \in \mathcal{I}} p_i - \rho \left(|\mathcal{I}| - \mathbb{1}_{\{(N) \in \mathcal{I}\}}\right) p_{(N)}$$

which is maximized by

$$\hat{\mathrm{I}} = \left\{i; p_i \geq \rho p_{(N)}\right\} = \left\{i; \pi_i \omega_i \geq \rho \pi_{(N)} \omega_{(N)}\right\}. \tag{20}$$

Thus we can tune the conservativeness by picking ρ adequately, as it specifies an upper bound on the fraction of the largest category probability other category probabilities may attain and still be excluded. If we choose $\rho = 0$, we get the classifier $\hat{\mathrm{I}} = \mathsf{N}$ which means $\mathbb{P}(\mathrm{I} \in \hat{\mathrm{I}}) = 1$ for all new observations, but that prediction method does not provide any information at all. The other extreme, choosing $\rho = 1$, leads to $\hat{\mathrm{I}} = \{(N)\}$, and thus our classifier will be the same as (19). In conclusion, our first classifier is a special case of the second.

Since (19) and (20) are both functions of ω_i, it suffices to estimate ω_i by Monte Carlo according to (8), (16) or (17) in order to get Monte Carlo estimates of $\hat{\mathrm{I}}$ and $\{\Omega_{\mathcal{I}}, \mathcal{I} \in \mathcal{N}\}$.

4.2 Classification allowing for empty outputs

If none of the N categories support test data $\mathcal{D}^{\text{new}}$ we would like to include $\emptyset$ as a possible output of the classifier $\hat{\mathrm{I}}$, so that $\hat{\mathrm{I}} \in \mathcal{P}(\mathsf{N})$. To this end, we denote the posterior weight of (13) as $\omega_i(x, \mathsf{S})$ in order to emphasize the dependence on the test data set $\mathcal{D}^{\text{new}} = (x, \mathsf{S})$. Then let

$$\bar{\omega}_i(x, \mathsf{S}) = \iint p(\theta_i \mid \mathcal{D}_i^{\text{obs}}) p(\mathsf{S}', x; \theta_i) \, \mathrm{d}\theta_i \mathrm{d}\mathsf{S}' \tag{21}$$

where the outer integral is taken over all S' such that $\omega_i(x, \mathsf{S}') \leq \omega_i(x, \mathsf{S})$. We interpret $\bar{\omega}_i(x, \mathsf{S})$ as a p-value of test data (x, S) for category i, i.e. the probability of observing an obfuscated trait vector S' of category i with covariate vector x, whose posterior weight $\omega_i(x, \mathsf{S}')$ is at most as large as that of (x, S). As such, it is a measure of the degree of outlyingness of $\mathcal{D}^{\text{new}}$. Then, for a given value of ρ, we generalize the classifier (20) to

$$\hat{\mathrm{I}} = \left\{i; \pi_i \omega_i \geq \rho \pi_{(N)} \omega_{(N)} \wedge \pi_i \bar{\omega}_i \geq \tau\right\} \tag{22}$$

where $\bar{\omega}_i = \bar{\omega}_i(x, \mathsf{S})$. Notice that (20) is as special case of (22) with $\tau = 0$.

For perfectly observed data $\mathsf{S}' = Y'$ we may approximate (21) by

$$\bar{\omega}_i(x, Y) \approx \int p\left(Y', x; \theta_i^{(\text{Bayes})}\right) \, \mathrm{d}Y' \tag{23}$$

where the integral is taken with respect to all Y' such that $\omega_i(x, Y') \leq \omega_i(x, Y)$. Since

$$Y \mid x, \theta_i \sim N(x\mathbf{B}_i, \mathbf{\Sigma}_i^{\alpha}) \tag{24}$$

and using the same type of approximation as in (23), it follows from (24) that

$$\begin{aligned} \omega_i(x, Y) \approx & \frac{1}{(2\pi)^{q/2} \det\left(\hat{\mathbf{\Sigma}}_i^{\alpha(x)}\right)^{1/2}} \cdot \\ & \cdot \exp\left\{-\frac{1}{2}(Y - x\hat{\mathbf{B}}_i)\left(\hat{\mathbf{\Sigma}}_i^{\alpha(x)}\right)^{-1}(Y - x\hat{\mathbf{B}}_i)^{\top}\right\} \end{aligned} \tag{25}$$

where $\hat{\mathbf{B}}_i$ and $\hat{\mathbf{\Sigma}}_i^{\alpha(x)}$ refer to the Bayes estimators of respective parameter. Now, $(Y - x\hat{\mathbf{B}}_i)\left(\hat{\mathbf{\Sigma}}_i^{\alpha(x)}\right)^{-1}(Y - x\hat{\mathbf{B}}_i)^{\top}$ has a $\chi^2(q)$-distribution with distribution function F, say. Therefore (23)-(25) imply

$$\bar{\omega}(x, Y) \approx 1 - F\left((Y - x\hat{\mathbf{B}}_i)\left(\hat{\mathbf{\Sigma}}_i^{\alpha(x)}\right)^{-1}(Y - x\hat{\mathbf{B}}_i)^{\top}\right)$$

which indeed can be interpreted as a p-value for (x, Y) to be an outlier, with the currently chosen τ.

4.3 Choosing ρ and τ

Choosing ρ is intentionally a subjective matter. In the case of a known cost of misclassifications, and a known cost of having a large indecisive region, one could certainly compute which ρ to use, to have the minimal expected cost. However, many applications are more vague, where the user may only believe that it is worse to misclassify, i.e. predict a category which the observation is not, than to not be precise, i.e. only rule out categories with sufficiently low degrees of belief, without being able to put a value on the cost. Together with the choice of prior category probabilitis $(\pi_1, \ldots, \pi_N)$ a user may completely accomodate the prior beliefs about the classification problem at hand. Indeed, $(\pi_1, \ldots, \pi_N)$ captures the apriori belief about how expected an observation from each category is relative the others and ρ is intended to represent the user's idea of the cost of misclassification, and finally τ represents how much outlyingness we accept without losing trust in out classifier.

A potential risk for misclassification is observing a subject of a category not even considered for classification. To allow for mitigation of this, we introduced τ as a cut-off value for the trait distributions. Indeed, the value of τ determines how large deviations in trait measurements we accept without becoming suspicious that we actually observe a subject from an unconsidered category. Choosing $\tau = 0$ allows us to classify any point in the whole trait space, i.e. we believe the parametrization is perfect and that no unconsidered categories can occur.

Finally, we remark that the parameter ρ could be used as an interesting measure of how managable a classification problem is. For any classification problem, we may compute the highest ρ-value for which $\mathrm{I} \in \hat{\mathrm{I}}$ in $\psi\,\%$ of our test cases. A more managable problem would then be one where the ρ-value is higher. It can be used as a model selection tool too; for any combination of traits and/or covariates, we compute the highest ρ-value for which $\mathrm{I} \in \hat{\mathrm{I}}$ in $\psi\,\%$ of our test cases, and then choose a model where ρ is acceptably high, and the model is as parsimonious as possible.

5 Model selection using cross-validation

We will present an approach to model selection using κ-fold cross-validation. It can be used to select covariates and/or traits to use from a larger set, based on predictive performance and parsimonity of the model.

The idea of κ-fold cross-validation is well established within the scientific community, and used for a very wide range of model families, see e.g. Wood (2017, p. 256). Choosing $\kappa = n_i$ for category i is the basic form of cross-validation, but it is computationally expensive, since one has to fit as many models $\sum_{i=1}^{N} n_i$ as there are observations. Since the method under study is already quite computationally intensive, in particular under widespread obfuscation, large q and large data sets, we recommend using κ-fold cross-validation with κ a bit smaller. An interesting conference proceeding by Kohavi (1995) examines cross-validation in general when choosing between classifiers. The author concludes that $\kappa < n_i$ is generally preferred when picking the best classifier using cross-validation.

To perform κ-fold cross-validation in general for our kind of models, we begin by choosing $\kappa \in \mathbb{Z}_+$ independently of i. Then create fold l for species i by choosing uniformly at random a set $J_{il} \subset \{1, \ldots, n_i\}$ comprising n_i/κ or $n_i/\kappa + 1$ observations of $\mathcal{D}_i$, the training data at hand for category i. Fit the model using training data set $\mathcal{D}_{i(-l)} = \mathcal{D} \setminus \{(x_{ij}, Y_{ij})\,; j \in J_{il}\}$ and predict the category of all observations in J_{il}; denote the classification of individual $j \in J_{il}$ of species i by $\hat{\mathrm{I}}_{(-il)ij}$. This procedure is repeated for $N\kappa$ test data

sets $\mathcal{D}_{(-il)}$, when $i = 1, \ldots, N$ and $l = 1, \ldots, \kappa$. To assess the predictive performance of the M models under consideration, let $w_i > 0$ be weights such that $\sum_{i=1}^{N} w_i = 1$, and compute

$$R_m^{\mathsf{cv}} = \sum_{i=1}^{N} \frac{w_i}{n_i} \sum_{l=1}^{\kappa} \sum_{j \in J_{il}} R(\hat{\mathrm{I}}_{(-il)ij}, i), \tag{26}$$

using the reward function that corresponds to a prespecified value of ρ, for $m = 1, \ldots, M$, and save the values. One could e.g. use the weights $w_i = n_i / \sum_{a=1}^{N} n_a$ or $w_i = 1/N$, depending on whether it is more valuable to be able to predict one category or not. Having computed $\{R_m^{\mathsf{cv}}; m = 1, \ldots, M\}$, the best classifier is

$$m^\star = \arg\max_m \left(R_1^{\mathsf{cv}}, \ldots, R_M^{\mathsf{cv}}\right).$$

When having computed R_m^{cv}, for $m = 1, \ldots, M$, one has the possibility to choose a more parsimonious model that performes almost as well as $m^\star$, under the usual principles of simplicity against predictive performance.

6 Simulation study of trait correlation effects

To orient ourselves regarding the effect of correlated traits, we performed the following study. First, we generated a vector of 100 correlation values, evenly spaced on $(-0.99, 0.99)$. For each correlation value, we generated $K = 10$ parameter sets with $N = 4$, $\sum_{i=1}^{N} n_i = 400$, $p = 1$, $q = 2$ and $A = 2$. The category of each observation was chosen uniformly at random on N. For both the covariance classes, the correlation between Y_{ij1} and Y_{ij2} was set to the current value from the 100 different correlation values, whereas the variance was set to 2 for both traits in the first covariance class and to 1 for both traits in the second covariance class. We imposed these values on the covariance matrices for all N categories. Moreover, the regression parameter matrices $B_i = \mathrm{vec}^{-1}(\beta_i)$ are drawn from the prior distributions

$$\begin{aligned}
\beta_i &\sim N\left(\beta_{0i}, \mathbf{I}_4\right), \\
\beta_{01} &= (10, 1, 10, -2)^\top, \\
\beta_{02} &= (10, 1, 13, -2)^\top, \\
\beta_{03} &= (13, 1, 13, -2)^\top, \\
\beta_{04} &= (13, 1, 10, -2)^\top
\end{aligned}$$

and set to the same value across correlations, for each of the K parameter sets. Then, for each parameter set, we generated a test data set, according to

the specified parameter values. Further, we generated another $L = 10$ training data sets for each parameter set, and fitted models to each of the L data sets. These fits were then evaluated on the test data set (i.e. the category of each observation in the test data set was predicted, given covariates and trait values) for the current estimated parameter values, and the reward was computed according to

$$\bar{R} = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \bar{R}_{kl} \tag{27}$$

for both reward functions. Here,

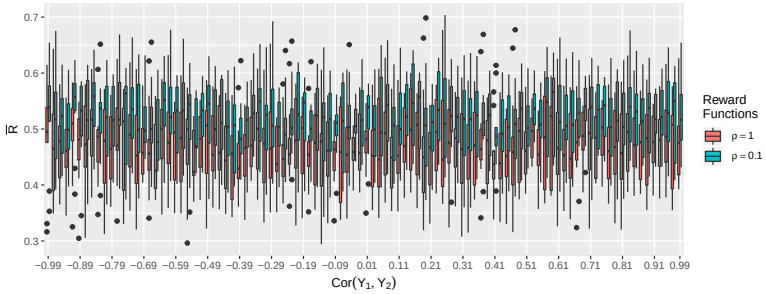
$$\bar{R}_{kl} = \frac{1}{\sum_{i=1}^{N} n_i} \sum_{i=1}^{N} \sum_{j=1}^{n_{ijk}} R(\hat{\mathrm{I}}_{ijkl}, i)$$

where $\hat{\mathrm{I}}_{ijkl}$ is the predicted category of the j:th among all n_{ijk} observations of category i in test data set k, based on estimates from training data set l. The interpretation of $\bar{R}$ is thus, when picking $\rho = 1$ and $\tau = 0$, the probability of classifying correctly, i.e. omitting all categories i from $\hat{\mathrm{I}}$ except the true category I. Based on the simulation study underlying Figure 1 (a), we conclude that correlation between traits has a small effect on the performance of the classifier. When fitting a simple linear regression model with $\bar{R}$ as response and the absolute value of the correlation as explanatory variable, the slope is statistically significant; $\mathbb{E}\left[\bar{R}\right]$ increases with the absolute value of the correlation, with about 0.001 units of $\bar{R}$ per 0.1 correlation units. We conclude that the effect of correlation of traits is very small on the predictive performance.

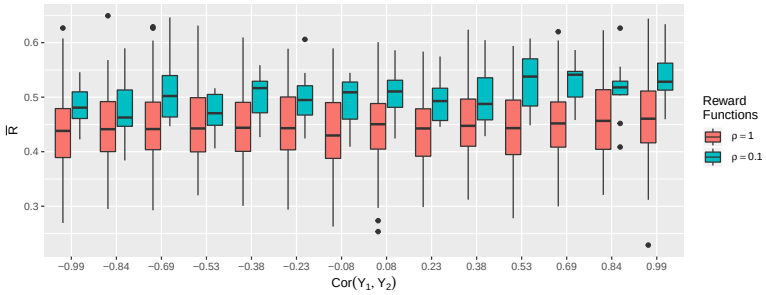
We also did a study with all parameters as before, but for 14 correlation values evenly spaced on $(-0.99, 0.99)$, $K = 100$ and $L = 10$. The corresponding plot is in Figure 1 (b). Most noticable is the lower variance in reward between correlation values, which essentially is due to a smoothing effect by increasing K. Recalling (27) we notice that as $K \to \infty$, $\mathrm{Var}(\bar{R}) \to 0$ by the law of large numbers. No statistically significant effect of $|\mathrm{Cor}(Y_1, Y_2)|$ on $\mathbb{E}\left[\bar{R}\right]$ was distinguishable.

7 A real data example

We examplify usage of the presented model on trait based species identification of four morphologically similar warblers in the *Acrocephalus* genus; Eurasian Reed Warbler (*Acrocephalus scirpaceus*), Marsh Warbler (*Acrocephalus palustris*), Blyth's Reed Warbler (*Acrocephalus dumetorum*) and



(a) Average reward for 100 different correlation values, with $K = 10$, $L = 10$.



(b) Average reward for 14 different correlation values, with $K = 100$, $L = 10$.

Figure 1: Although the plots do not reveal it, we detected a slight positive correlation between $|\text{Cor}(Y_1, Y_2)|$, and $\bar{R}$ for the simulation in subplot (a), meaning that classification is a little easier using correlated traits.

Paddyfield Warbler (*Acrocephalus agricola*). This problem has been approached before by Walinder et al. (1988) and Malmhagen et al. (2013), at least partially using the same data, but in statistically much less rigourous ways.

A row $x_{ij} = (1, x_{ij1})$ of the design matrix $\mathbf{X}_i$ includes as second entry a binary *age* (levels are *juvenile* and *adult*), thus $p = 1$. Traits are *wing length* rounded to millimeters, second primary *notch length* rounded to half millimeters and *notch-wing position*, which is ordered categorical with $g_3 = 18$ levels, making $q = 3$. For those with rusty bird topography, these are all measurements of different parts of the wings. Since all measurements are rounded or ordered categorical, we have no complete observations, only various truncated observations. Enumeration of individuals within species is according to order of appearance.

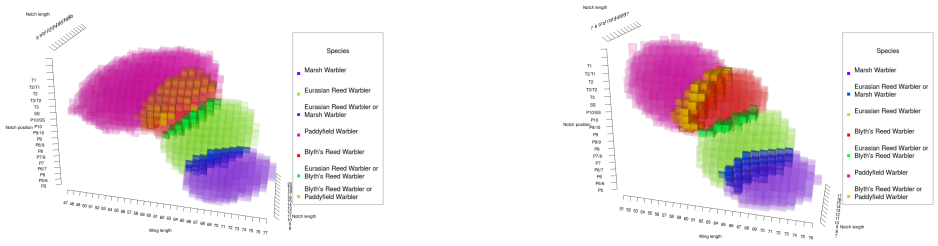
We let the covariance matrices $\mathbf{\Sigma}_{ij}$ depend on j through our only covariate *age*, writing $\mathbf{\Sigma}_i^{\texttt{juv}}$ and $\mathbf{\Sigma}_i^{\texttt{ad}}$ for juveniles and adults respectively. Since *age* is a binary covariate, we have two different covariance matrices for each species. It can be noted that in this specific case, it would be equivalent to model juveniles and adults separately with $N = 8$ and $p = 0$, rather than $N = 4$ and $p = 1$. Our general presentation still allows us to handle this with $N = 4$ categories, as it is a special case of when there are several ($A = 2$) covariance classes with common regression parameters. We impose the prior distributions

$$\begin{aligned}\mathbf{\Sigma}_i^{\texttt{juv}}, \mathbf{\Sigma}_i^{\texttt{ad}} &\overset{\text{i.i.d.}}{\sim} IW(\nu_0, \mathbf{V}_0)\\ \beta_i &\sim N\left(\text{vec}(\mathbf{B}_{i0}), \mathbf{I}_q \otimes \mathbf{\Sigma}_B^{-1}\right)\\ \pi_i &= \frac{1}{4},\ i = 1, \ldots, 4\end{aligned}$$

on β_i, $\mathbf{\Sigma}_i^{\texttt{juv}}$ and $\mathbf{\Sigma}_i^{\texttt{ad}}$, for $i = 1, \ldots, 4$, with the parameter values given in Appendix C. In short, the prior is somewhat informative for categories that have few observations.

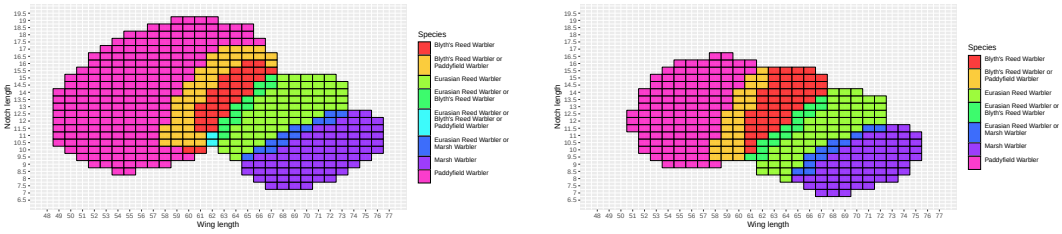
Implementing the model in R (R Core Team, 2018) and generating samples from the posterior with the described blockwise Gibbs sampling algorithm, we focus on presenting the decision regions here, whereas the parameter estimates can be found in Appendix C. Since we have a $q = 3$-dimensional trait space, we can visualize the decision regions given by the two considered decision rules, including the indecisive region Λ. For this application, the trait measurements are always obfuscated in some way, so we discretize the trait space when visualizing the decision regions. We also present the decision regions for the cases when some traits are unobserved, to stress the usability of the method also under incomplete data. One can also straightforwardly

compute decsion regions for when covariates are unobserved, but we omit it here.



(a) Adult birds. (b) Juvenile birds.

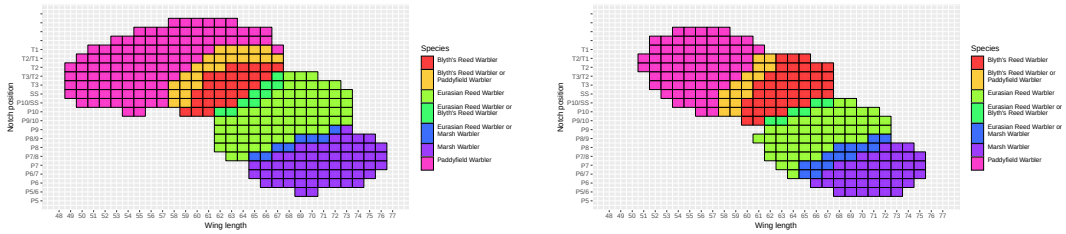
Figure 2: Decision regions when observing all three traits of the Acrocephalus warblers, using $\rho = 0.1$ and $\tau = 0.001$. The indecisive region Λ is less transparent, and colored according to which species one is unsure about. The probability of observing an individual that falls in the indecisive region is 0.0797 for (a) and 0.0795 for (b). The decision region of Paddyfield Warbler partially engulfs Blyth's Reed Warbler for adult birds, reflecting the large uncertainty in the parameter estimates for adult Paddyfield Warblers. Notice also that we introduce unnamed categories for notch position, as the predictive posterior distribution requires this.



(a) Adult birds. (b) Juvenile birds.

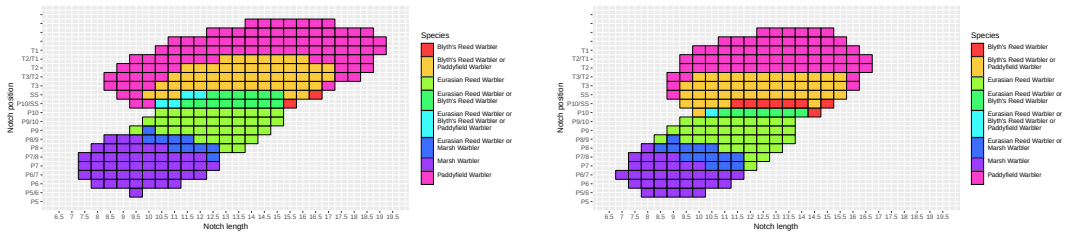
Figure 3: Decision regions when only observing wing and notch length.

Summarizing Figure 2, we have a relatively small indecisive region, even for rather low values of ρ. In particular, the indecisive region is generally located in low-density areas of the trait space, meaning that it is overall unlikely to observe any subject with traits that make a single category distinguishable. Moreover, the indecisive region rule out two categories in every case. Hence, the classifier is quite effective.



(a) Adult birds. (b) Juvenile birds.

Figure 4: Decision regions when only observing wing length and notch position.



(a) Adult birds. (b) Juvenile birds.

Figure 5: Decision regions when only observing notch length and notch position.

Choosing $\rho = 1$ and $\tau = 0$, i.e. doing traditional classification with $|\hat{\mathrm{I}}| = 1$, we estimated the probabilities of classifying wrongly as

$$\hat{\mathbb{P}}\left(\hat{\mathrm{I}} \neq \mathrm{I} \mid x_1 = 0\right) = 0.0251, \quad \hat{\mathbb{P}}\left(\hat{\mathrm{I}} \neq \mathrm{I} \mid x_1 = 1\right) = 0.0262.$$

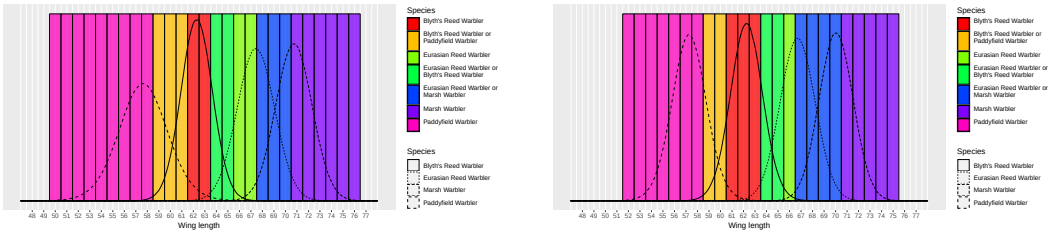
Considering the case of $\rho = 0.1$ and $\tau = 0.001$, we find that

$$\hat{\mathbb{P}}\left(\mathrm{I} \notin \hat{\mathrm{I}} \mid x_1 = 0\right) = 0.0065, \quad \hat{\mathbb{P}}\left(\mathrm{I} \notin \hat{\mathrm{I}} \mid x_1 = 1\right) = 0.0068$$

and that

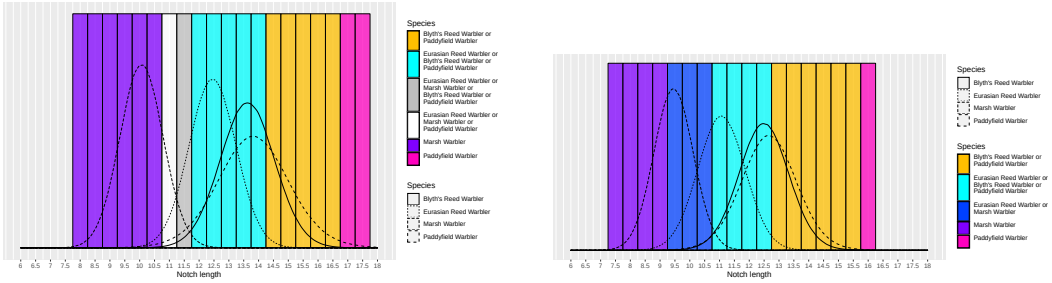
$$\hat{\mathbb{P}}\left(|\hat{\mathrm{I}}| > 1 \mid x_1 = 0\right) = 0.0795, \quad \hat{\mathbb{P}}\left(|\hat{\mathrm{I}}| > 1 \mid x_1 = 1\right) = 0.0797.$$

Overall, this means that by introducing the classification rule using ρ and τ, we have managed a 83.5% decrease in probability of excluding the true species from the set of possible species, at the cost of in 7.91% of the cases not pinpointing one species, for juvenile birds. For adult birds, we manage a 77.7% decrease at the cost of being indecisive in 7.93% of the cases. In all considered cases, we at most choose two species as members of $\hat{\mathrm{I}}$, when



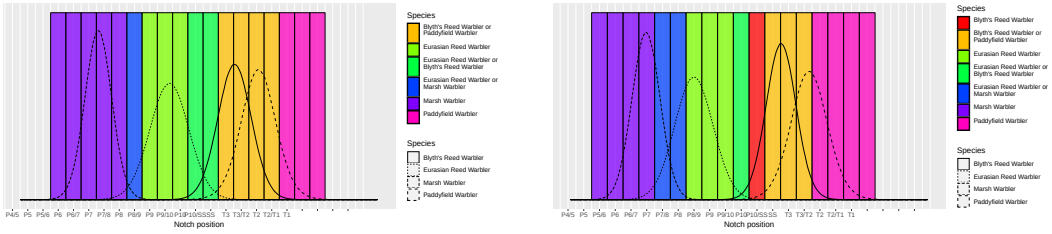
(a) Adult birds.

(b) Juvenile birds.



(c) Adult birds.

(d) Juvenile birds.



(e) Adult birds.

(f) Juvenile birds.

Figure 6: In (a) and (b), decision regions when only wing length is observed are shown; in (c) and (d) decision regions when only notch length is observed are shown; and in (e) and (f) decision regions when only notch position is observed are shown. In all plots, kernel density estimates of each aposteriori trait distribution for each species is shown with black lines of different types; it highlights the larger degree of separation in the traits wing length and notch position.

observing all three traits. Further lowering ρ trades larger indecisive regions for even lower probabilities of excluding the correct species, whereas lowering τ restricts the part of the trait space we feel confident classifying in.

The estimated probabilities presented above hold true when observing

birds of each species with equal probability. Without going into details, we did some quick tests of the performance of the classifier under very skewed prior probability distributions, e.g. $\pi = (1 - \pi_2 - \pi', \pi_2, \pi'/2, \pi'/2)$ with π' small. The decision regions became very similar, indicating that our classifier is very insensitive to prior category probabilities.

Subclassifiers can straightforwardly be derived. If one e.g. observes only birds of species $i = 1, 2$ with equal probability,

$$\begin{aligned}
\hat{\mathbb{P}}\left(|\hat{\mathrm{I}}| > 1 \mid x_1 = 0, \mathrm{I} \in \{1,2\}\right) &= 0.0487 \\
\hat{\mathbb{P}}\left(|\hat{\mathrm{I}}| > 1 \mid x_1 = 1, \mathrm{I} \in \{1,2\}\right) &= 0.0118 \\
\hat{\mathbb{P}}\left(\mathrm{I} \notin \hat{\mathrm{I}} \mid x_1 = 0, \mathrm{I} \in \{1,2\}\right) &= 0.0044 \\
\hat{\mathbb{P}}\left(\mathrm{I} \notin \hat{\mathrm{I}} \mid x_1 = 1, \mathrm{I} \in \{1,2\}\right) &= 0.0018.
\end{aligned}$$

In summary, the probability of being indecisive $\mathbb{P}(Y \in \Lambda)$ and the probability of being wrong $\mathbb{P}(\mathrm{I} \notin \hat{\mathrm{I}})$ is lower than overall for this pair of species, indicating that most of the probability mass of Λ is in other regions than the overlap of these two species' densities. One may compute probabilities for any combination of species, and for any apriori distribution over categories.

One of the species, $i = 4$ (Paddyfield Warbler), has considerably less data available than the other species ($n_4 = 31$, $n_4^{\mathtt{juv}} = 19$ and $n_4^{\mathtt{ad}} = 12$), which is the main reason for this species to seemingly have greater variation in its trait values than the other species, as parameter uncertainty is large. Graphically, one can notice in every Figure concerning adult birds, that the region colored in this species' color is elongated and larger than the other species. It also causes Λ to increase in size, therfore increasing the probability of observing a bird that we cannot determine the species of by these traits. One could mitigate this by choosing less vague priors for the covariance matrices $\mathbf{\Sigma}_4^{\mathtt{juv}}$ and $\mathbf{\Sigma}_4^{\mathtt{ad}}$, if one has apriori knowledge to use in such a construction, or collect more data on this species.

8 Discussion

Throughout this paper, we have set up and analysed a classification problem where classification is aided by two sets of observations for each subject; its trait vector Y and its covariate vector x, where x is informative about the interpretation of Y. Since the trait values often are of various kinds and obfuscated, we set up a uniform framework for these situations using a latent multivariate Gaussian distribution. To formalize the classification, we introduce reward functions and two parameters $\rho \in [0, 1]$ and $\tau \in [0, 1)$. The

choice of ρ affects the size and location of the indecisive region Λ, which is everywhere in observation space where our classifer does not have sufficient information to rule out all but one category, whereas τ puts a limit on how much we allow an observation to deviate and still accept classification. The effect of correlation between traits and the loss of information through typical obfuscations are illustrated through a simulation study and an example of a real world situation where this procedure is applicable.

Overall, there are two main usages of the method presented in this paper. First, one may use a fitted model to classify new observations with statistical rigour. Secondly, one may derive distinguishing characteristics of the categories considered. An example of the usefulness of the second case would be an ornithologist with a set of birds of known taxa, who doesn't know what morphologically separates these taxa. Using this method, she may extract for which trait measurements there is a high probability of certain taxa and thereby create (and write down) an identification key. Further, if there are too many combinations of trait levels to memorize, the model may perform the classification in an automized way. This highlights the mathematical equivalence of the two activities.

An adjustment that is conceptually critical but often negligable numerically, is to correct all latent Gaussian distributions by truncating them to only positive values for some traits. The trait *wing length* in our real world example has to be positive by definition, and hence we should adjust our Gaussian distribution accordingly. However, considering the parameter estimates (see Appendix C), it would make a practically undetectable difference in our example. For other cases, it could indeed be of larger importance.

The reliance on training data with known categories can potentially be relaxed, or at least partially relaxed. To take a step towards unsupervised learning, one would need to add a clustering layer to the model. Potentially, one could specify the number of clusters to identify, specify a range of integers within which the number of clusters should be located or leave it completely open. This would allow the method to be used in situations where it is not known how to classify observations at all, and thus investigate wheter there is support for several categories in a given data set.

Modifying the method slightly in order to handle repeated measurements is a straightforward task. The benefit with repeated measurements of the traits is a better understanding of the magnitude of the measurement error, and greater confidence in the average as being close to the true value. One could then integrate the number of measurements into the classification method, with direct effects on the size of the indecisive region.

As mentioned in Section 2, we assume independence between the regression parameters within different categories. This allows the effect of a co-

variate to vary between the categories, as opposed to forcing the same effect of a covariate across categories. However, in Appendix C, one may find the posterior means of our covariate effects in our real data example of Section 7, and notice that the effect is similar for some traits across categories, and even across traits to some extent. Arguably, this indicates that there is a general effect of our covariate, and therefore we could construct a model that allows for such a general effect, although we did not feel confident in assuming this apriori in our example.

Acknowledgements

The authors are grateful for the data on *Acrocephalus* warblers provided by Falsterbo Bird Observatory, Lennart Karlsson and Björn Malmhagen, and for helpful methodological feedback from Felix Wahl and Mathias Lindholm.

A The posterior distribution

We start by formulating a Lemma which will be useful in the proof of Proposition 1. Denote an identity matrix of rank n with $\mathbf{I}_n$.

Lemma 1. *Let* $\mathbf{Z}$ *be a block-diagonal* $nq \times (p+1)q$ *matrix, where there are* q *blocks* $\mathbf{X}$*, which are* $n \times (p+1)$*-matrices, along the diagonal. Let* $\mathbf{\Sigma}$ *be a symmetric, positive definite* $q \times q$*-matrix. Then it holds that*

$$\mathbf{Z}^\top \left(\mathbf{\Sigma} \otimes \mathbf{I}_n\right)^{-1} \mathbf{Z} = \left(\mathbf{Z} \left(\mathbf{\Sigma} \otimes \mathbf{I}_{p+1}\right)^{-1}\right)^\top \mathbf{Z}.$$

Proof. We prove the lemma by iterated use of the mixed-product property of the tensor product. Since $\mathbf{Z} = (\mathbf{I}_q \otimes \mathbf{X})$, the left hand side becomes

$$\begin{aligned}
\left(\mathbf{I}_q \otimes \mathbf{X}\right)^\top \left(\mathbf{\Sigma} \otimes \mathbf{I}_n\right)^{-1} \left(\mathbf{I}_q \otimes \mathbf{X}\right) &= \left(\mathbf{I}_q \otimes \mathbf{X}^\top\right) \left(\mathbf{\Sigma}^{-1} \otimes \mathbf{I}_n\right) \left(\mathbf{I}_q \otimes \mathbf{X}\right) \\
&= \left(\mathbf{I}_q \mathbf{\Sigma}^{-1} \otimes \mathbf{X}^\top \mathbf{I}_n\right) \left(\mathbf{I}_q \otimes \mathbf{X}\right) \\
&= \left(\mathbf{\Sigma}^{-1} \otimes \mathbf{X}^\top\right) \left(\mathbf{I}_q \otimes \mathbf{X}\right) \\
&= \mathbf{\Sigma}^{-1} \mathbf{I}_q \otimes \mathbf{X}^\top \mathbf{X} \\
&= \mathbf{\Sigma}^{-1} \otimes \mathbf{X}^\top \mathbf{X}
\end{aligned}$$

and the right hand side becomes

$$\begin{aligned}
\left((\mathbf{I}_q \otimes \mathbf{X})\left(\boldsymbol{\Sigma} \otimes \mathbf{I}_{p+1}\right)^{-1}\right)^\top (\mathbf{I}_q \otimes \mathbf{X}) &= \left(\mathbf{I}_q \boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}\mathbf{I}_{p+1}\right)^\top (\mathbf{I}_q \otimes \mathbf{X}) \\
&= \left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}\right)^\top (\mathbf{I}_q \otimes \mathbf{X}) \\
\{\text{by symmetry of } \boldsymbol{\Sigma}\} &= \left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}^\top\right)(\mathbf{I}_q \otimes \mathbf{X}) \\
&= \boldsymbol{\Sigma}^{-1}\mathbf{I}_q \otimes \mathbf{X}^\top\mathbf{X} \\
&= \boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}^\top\mathbf{X}
\end{aligned}$$

which proves the lemma already in the third equalities. □

The simplification of the expressions will be used in the forthcoming Proposition 1, but also noticable is that the RHS of the equality can computationally be considerably faster when implementing this method, than the LHS, which is the canonical parametrization.

In Rossi et al. (2012) we can find the posterior distribution of the regression parameter vector β in a multivariate multiple regression with only one covariance class. Since we have A covariance classes, we generalize this result slightly. All notation is as in the main paper, but we omit the index i since we consider a general case. To breifly recapitulate, we have n observations of q traits explained by p covariates, contained in $\mathbf{Y}$ and $\mathbf{X}$ respectively. We construct $\mathbf{U}$ and $\mathbf{Z}$ as in (2) and (3) respectively. The observations are distributed over A covariance classes and we denote observations belonging to covariance class α by adding a superscript to $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ or $\mathbf{U}$ accordingly. Our regression parameters are included in the $(p+1)\times q$ matrix $\mathbf{B}$, which we write in vector form as $\beta = \text{vec}(\mathbf{B})$. Assuming a prior on each of the columns of $\mathbf{B}$, and letting it be $N\left((\gamma_1, \ldots, \gamma_{p+1})_k^\top = \beta_k, \boldsymbol{\Sigma}_\mathbf{B}\right)$ for $k = 1, \ldots, q$, we obtain a prior $N\left(\left(\beta_1^\top, \ldots, \beta_q^\top\right)^\top = \beta_0, \mathbf{I}_q \otimes \boldsymbol{\Sigma}_\mathbf{B} = \boldsymbol{\Sigma}_\beta\right)$ on β. Each observation of a trait vector y with covariate vector x is assumed to follow the distribution $N(x\mathbf{B}, \boldsymbol{\Sigma}^{\alpha(x)})$, i.e. the covariance class is determined by the covariate.

Proposition 1. *Denote the parameter vector of a Bayesian multivariate multiple regression model with A covariance classes by $\theta = \left(\beta, \boldsymbol{\Sigma}^1, \ldots, \boldsymbol{\Sigma}^A\right)$, where β is the regression parameter vector and $\boldsymbol{\Sigma}^1, \ldots, \boldsymbol{\Sigma}^A$ are the A covariance matrices. Let the prior of θ be $p(\theta) = p(\beta)\prod_{\alpha=1}^A p(\boldsymbol{\Sigma}^\alpha)$, where $\beta \sim N(\beta_0, \boldsymbol{\Sigma}_\beta)$ and $\boldsymbol{\Sigma}^\alpha \sim IW(\nu_0, \mathbf{V}_0)$ for $\alpha = 1, \ldots, A$. Then the posterior*

distribution of $\beta \mid \mathbf{U}, \mathbf{Z}, \boldsymbol{\Sigma}^1, \ldots, \boldsymbol{\Sigma}^\alpha$ *is* $N(\tilde{\beta}, \tilde{\boldsymbol{\Sigma}})$*, where*

$$\tilde{\boldsymbol{\Sigma}} = \left[\boldsymbol{\Sigma}_\beta^{-1} + \sum_{\alpha=1}^{A} (\boldsymbol{\Sigma}^\alpha)^{-1} \otimes (\mathbf{X}^\alpha)^\top \mathbf{X}^\alpha \right]^{-1}$$

$$\tilde{\beta} = \tilde{\boldsymbol{\Sigma}} \times \left[\boldsymbol{\Sigma}_\beta^{-1} \beta_0 + \sum_{\alpha=1}^{A} \left((\boldsymbol{\Sigma}^\alpha)^{-1} \otimes (\mathbf{X}^\alpha)^\top \right) \mathbf{U}^\alpha \right].$$

Proof. By applying Bayes' theorem

$$\begin{aligned} p\left(\beta \mid \mathbf{U}, \mathbf{Z}, \boldsymbol{\Sigma}^1, \ldots \boldsymbol{\Sigma}^A\right) &\propto \exp\left\{ -\frac{1}{2} (\beta - \beta_0)^\top \boldsymbol{\Sigma}_\beta^{-1} (\beta - \beta_0) \right\} \cdot \\ &\quad \cdot \prod_{\alpha=1}^{A} \exp\left\{ -\frac{1}{2} \left(\mathbf{U}^\alpha - \mathbf{Z}^\alpha \beta \right)^\top \left(\boldsymbol{\Sigma}^\alpha \otimes \mathbf{I}_{p+1} \right)^{-1} \left(\mathbf{U}^\alpha - \mathbf{Z}^\alpha \beta \right) \right\} \\ &= \exp\left\{ -\frac{1}{2} \beta \mathbf{C} \beta + \beta \mathbf{D} \right\} \end{aligned}$$

where, as by the proof of Lemma 1,

$$\begin{aligned} \mathbf{C} &= \boldsymbol{\Sigma}_\beta^{-1} + \sum_{\alpha=1}^{A} \left(\mathbf{Z}^\alpha (\boldsymbol{\Sigma}^\alpha \otimes \mathbf{I}_n)^{-1} \right)^\top \mathbf{Z}^\alpha \\ &= \boldsymbol{\Sigma}_\beta^{-1} + \sum_{\alpha=1}^{A} (\boldsymbol{\Sigma}^\alpha)^{-1} \otimes (\mathbf{X}^\alpha)^\top \mathbf{X}^\alpha \end{aligned}$$

and

$$\begin{aligned} \mathbf{D} &= \boldsymbol{\Sigma}_\beta^{-1} \beta_0 + \sum_{\alpha=1}^{A} (\mathbf{Z}^\alpha)^\top (\boldsymbol{\Sigma}^\alpha \otimes \mathbf{I}_n)^{-1} \mathbf{U}^\alpha \\ &= \boldsymbol{\Sigma}_\beta^{-1} \beta_0 + \sum_{\alpha=1}^{A} \left((\boldsymbol{\Sigma}^\alpha)^{-1} \otimes (\mathbf{X}^\alpha)^\top \right) \mathbf{U}^\alpha. \end{aligned}$$

Consequently,

$$\beta \mid \mathbf{U}, \mathbf{Z}, \boldsymbol{\Sigma}^1, \ldots, \boldsymbol{\Sigma}^A \sim N(\tilde{\beta}, \tilde{\boldsymbol{\Sigma}})$$

where

$$\begin{aligned} \tilde{\beta} &= \mathbf{C}^{-1} \mathbf{D} \\ &= \left[\boldsymbol{\Sigma}_\beta^{-1} + \sum_{\alpha=1}^{A} (\boldsymbol{\Sigma}^\alpha)^{-1} \otimes (\mathbf{X}^\alpha)^\top \mathbf{X}^\alpha \right]^{-1} \times \left[\boldsymbol{\Sigma}_\beta^{-1} \beta_0 + \sum_{\alpha=1}^{A} \left((\boldsymbol{\Sigma}^\alpha)^{-1} \otimes (\mathbf{X}^\alpha)^\top \right) \mathbf{U}^\alpha \right] \end{aligned}$$

and

$$\tilde{\boldsymbol{\Sigma}} = \mathbf{C}^{-1} = \left[\boldsymbol{\Sigma}_\beta^{-1} + \sum_{\alpha=1}^{A} (\boldsymbol{\Sigma}^\alpha)^{-1} \otimes (\mathbf{X}^\alpha)^\top \mathbf{X}^\alpha \right]^{-1}.$$

□

Note how the covariance matrix $\tilde{\mathbf{\Sigma}}$ is a multivariate version of a weighted average of the prior covariance matrix $\mathbf{\Sigma}_\beta$ and the other A covariance matrices. A considerable gain in computational speed is achieved by reducing the matrix expressions from their original form to the one presented in Lemma 1.

B Gibbs sampling details

The focus of this section is Procedure 1, in which we describe in detail how to generate a sample of size R_i from the posterior distribution of the parameter vector θ_i, using blockwise Gibbs sampling. It describes the general case, i.e. when we have obfuscated trait measurements in $\mathcal{D}^{\text{obs}}$. For the case with perfectly observed trait measurements, we skip the sampling of $\mathbf{Y}_i$ and use the observed values instead, otherwise the procedure is the same. Applying the procedure to data from each category i will yield all the samples we need from the posterior distribution to perform classification.

In Procedure 1, $TN(\mu, \mathbf{\Sigma}, \mathsf{S})$ refers to the truncated Gaussian distribution, where μ is the mean vector, $\mathbf{\Sigma}$ is the covariance matrix and S describes the truncation limits. Simulating from this distribution can be done exactly using rejection sampling, or approximately using an inner Gibbs algorithm. Depending on the case, either approach can be preferred, as the tradeoff is exact sampling versus efficiency. Also, more advanced algorithms such as Importance Sampling-techniques can be used in this step.

Procedure 1 The Monte Carlo approach to sampling the parameters' posterior distirbution under obfuscation.

Input: $\mathcal{D}^{\text{obs}}$

Output: A sample of size R_i from the posterior distribution of θ_i.

for $\alpha = 1 \to A$ **do**
  draw $\boldsymbol{\Sigma}_{i0}^{\alpha} \sim IW\left(\nu_0, \mathbf{V}_0\right)$
end for
draw $\beta_{i0} \sim N\left(\beta_{i0}, \boldsymbol{\Sigma}_{\beta_i}\right)$
$\theta_{i0} \leftarrow \left(\beta_{i0}, \boldsymbol{\Sigma}_{i0}^{1}, \ldots, \boldsymbol{\Sigma}_{i0}^{A}\right)$
for $r = 1 \to R_i$ **do**
  for $j = 1 \to n_i$ **do**
    draw $\mathbf{Y}_{ij,r-1} \mid x_{ij}, \mathsf{S}_{ij}, \theta_{i,r-1} \sim TN\left(\mathbf{X}_{ij}\mathbf{B}_{i0}, \boldsymbol{\Sigma}_{ij0}, \mathsf{S}_{ij}\right)$
  end for
  $\mathbf{U}_{i,r-1} \leftarrow \text{vec}(\mathbf{Y}_{i,r-1})$
  draw $\beta_{ir} \mid \mathbf{U}_{i(r-1)}, \left\{\boldsymbol{\Sigma}_{i(r-1)}^{\alpha}\right\}_{\alpha=1}^{A} \sim N\left(\tilde{\beta}, \tilde{\boldsymbol{\Sigma}}\right)$
  for $\alpha = 1 \to A$ **do**
    draw $\boldsymbol{\Sigma}_{ir}^{\alpha} \mid \mathbf{U}_{i(r-1)}^{\alpha}, \mathbf{Z}_i^{\alpha}, \beta_{ir} \sim IW(\nu_0 + n_i^{\alpha}, \mathbf{V}_0 + \mathbf{S}_i^{\alpha})$
  end for
  $\theta_{ir} \leftarrow \left(\beta_{ir}, \boldsymbol{\Sigma}_{ir}^{1}, \ldots, \boldsymbol{\Sigma}_{ir}^{A}\right)$
  save θ_{ir}
end for

Table 1: Prior parameter values for the real world data example.

$$\nu_0 = 10 \qquad \mathbf{V}_0 = \begin{pmatrix} 15 & 5 & 5 \\ 5 & 15 & 5 \\ 5 & 5 & 15 \end{pmatrix} \qquad \mathbf{\Sigma}_B = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mathbf{B}_{10} = \begin{pmatrix} 66 & 11 & 108 \\ 1 & 1 & 1 \end{pmatrix} \qquad \mathbf{B}_{20} = \begin{pmatrix} 70 & 9.5 & 105 \\ 1 & 0.5 & 1 \end{pmatrix} \qquad \mathbf{B}_{30} = \begin{pmatrix} 57 & 12.5 & 115 \\ 1 & 1 & 1 \end{pmatrix} \qquad \mathbf{B}_{40} = \begin{pmatrix} 62 & 12.5 & 113 \\ 1 & 1 & 1 \end{pmatrix}.$$

C Results of *Acrocephalus* analysis

Section 7 of the paper contains an example of using our method on a real world data set. The parameters' of the prior distributions used are in Table 1, the mean of the parameters posterior distributions are in Table 2, and some quantiles of the parameters' posterior distributions are in Table 3.

Table 2: Monte Carlo estimated means of parameter posterior distributions.

$$\mathbf{B}_1 = \begin{pmatrix} 66.70 & 11.06 & 107.99 \\ 0.73 & 1.39 & 2.29 \end{pmatrix} \quad \mathbf{B}_2 = \begin{pmatrix} 70.05 & 9.45 & 104.96 \\ 0.69 & 0.61 & 0.67 \end{pmatrix} \quad \mathbf{B}_3 = \begin{pmatrix} 57.27 & 12.66 & 115.31 \\ 0.37 & 1.16 & 0.82 \end{pmatrix} \quad \mathbf{B}_4 = \begin{pmatrix} 62.26 & 12.49 & 113.53 \\ 0.03 & 1.13 & 1.03 \end{pmatrix}$$

$$\boldsymbol{\Sigma}_1^{\mathtt{juv}} = \begin{pmatrix} 2.35 & 0.63 & 0.07 \\ 0.63 & 0.63 & 0.52 \\ 0.07 & 0.52 & 1.48 \end{pmatrix} \quad \boldsymbol{\Sigma}_2^{\mathtt{juv}} = \begin{pmatrix} 2.19 & 0.40 & -0.05 \\ 0.40 & 0.44 & 0.19 \\ -0.05 & 0.19 & 0.80 \end{pmatrix} \quad \boldsymbol{\Sigma}_3^{\mathtt{juv}} = \begin{pmatrix} 1.99 & 0.41 & 0.18 \\ 0.41 & 0.71 & 0.17 \\ 0.18 & 0.17 & 0.92 \end{pmatrix} \quad \boldsymbol{\Sigma}_4^{\mathtt{juv}} = \begin{pmatrix} 2.29 & 0.26 & -0.42 \\ 0.26 & 0.87 & 0.29 \\ -0.42 & 0.29 & 1.38 \end{pmatrix}$$

$$\boldsymbol{\Sigma}_1^{\mathtt{ad}} = \begin{pmatrix} 2.67 & 0.54 & 0.46 \\ 0.54 & 0.63 & 0.49 \\ 0.46 & 0.49 & 1.65 \end{pmatrix} \quad \boldsymbol{\Sigma}_2^{\mathtt{ad}} = \begin{pmatrix} 2.51 & 0.49 & 0.21 \\ 0.49 & 0.53 & 0.16 \\ 0.21 & 0.16 & 0.78 \end{pmatrix} \quad \boldsymbol{\Sigma}_3^{\mathtt{ad}} = \begin{pmatrix} 1.89 & 0.60 & 0.11 \\ 0.60 & 0.85 & 0.32 \\ 0.11 & 0.32 & 1.22 \end{pmatrix} \quad \boldsymbol{\Sigma}_4^{\mathtt{ad}} = \begin{pmatrix} 4.63 & 0.94 & 0.89 \\ 0.94 & 1.45 & 0.58 \\ 0.89 & 0.58 & 1.33 \end{pmatrix}$$

Table 3: Quantiles of the posterior distribution for the parameters of the *Acrocephalus* model. The parameters are indexed as $(\cdot)_{imk}$, where $i = 1, \ldots, N$, $m = 0, \ldots, p$ and $k = 1, \ldots, q$, with $N = 4$, $p = 1$ and $q = 3$. The order of i is *Eurasian reed warbler*, *Marsh warbler*, *Blyth's reed warbler* and *Paddyfield warbler*; the order of m is intercept then *age*; and the order of k is *wing length*, *notch length* then *notch position*.

Quantile	$\mathbf{B}_{101}$	$\mathbf{B}_{111}$	$\mathbf{B}_{102}$	$\mathbf{B}_{112}$	$\mathbf{B}_{103}$	$\mathbf{B}_{113}$	$\mathbf{\Sigma}^{\texttt{juv}}_{111}$	$\mathbf{\Sigma}^{\texttt{juv}}_{112}$	$\mathbf{\Sigma}^{\texttt{juv}}_{113}$	$\mathbf{\Sigma}^{\texttt{juv}}_{122}$	$\mathbf{\Sigma}^{\texttt{juv}}_{123}$	$\mathbf{\Sigma}^{\texttt{juv}}_{133}$	$\mathbf{\Sigma}^{\texttt{ad}}_{111}$	$\mathbf{\Sigma}^{\texttt{ad}}_{112}$	$\mathbf{\Sigma}^{\texttt{ad}}_{113}$	$\mathbf{\Sigma}^{\texttt{ad}}_{122}$	$\mathbf{\Sigma}^{\texttt{ad}}_{123}$	$\mathbf{\Sigma}^{\texttt{ad}}_{133}$
2.5%	66.69	0.70	10.99	1.27	107.87	1.95	2.31	0.55	-0.11	0.57	0.42	1.28	2.61	0.43	-0.03	0.55	0.35	1.30
50%	66.70	0.73	11.07	1.38	108.00	2.29	2.35	0.63	0.07	0.63	0.52	1.48	2.67	0.54	0.47	0.62	0.47	1.59
97.5%	66.72	0.76	11.13	1.52	108.13	2.56	2.38	0.72	0.23	0.71	0.64	1.66	2.74	0.64	0.90	0.72	0.74	2.35

Quantile	$\mathbf{B}_{201}$	$\mathbf{B}_{211}$	$\mathbf{B}_{202}$	$\mathbf{B}_{212}$	$\mathbf{B}_{203}$	$\mathbf{B}_{213}$	$\mathbf{\Sigma}^{\texttt{juv}}_{211}$	$\mathbf{\Sigma}^{\texttt{juv}}_{212}$	$\mathbf{\Sigma}^{\texttt{juv}}_{213}$	$\mathbf{\Sigma}^{\texttt{juv}}_{222}$	$\mathbf{\Sigma}^{\texttt{juv}}_{223}$	$\mathbf{\Sigma}^{\texttt{juv}}_{233}$	$\mathbf{\Sigma}^{\texttt{ad}}_{211}$	$\mathbf{\Sigma}^{\texttt{ad}}_{212}$	$\mathbf{\Sigma}^{\texttt{ad}}_{213}$	$\mathbf{\Sigma}^{\texttt{ad}}_{222}$	$\mathbf{\Sigma}^{\texttt{ad}}_{223}$	$\mathbf{\Sigma}^{\texttt{ad}}_{233}$
2.5%	69.99	0.57	9.42	0.54	104.86	0.39	2.07	0.35	-0.19	0.41	0.13	0.69	2.28	0.40	-0.15	0.47	-0.01	0.50
50%	70.05	0.69	9.45	0.61	104.96	0.66	2.19	0.40	-0.05	0.44	0.19	0.80	2.51	0.49	0.20	0.53	0.16	0.76
97.5%	70.10	0.81	9.48	0.68	105.04	1.00	2.32	0.44	0.08	0.47	0.25	0.92	2.76	0.60	0.62	0.60	0.34	1.16

Quantile	$\mathbf{B}_{301}$	$\mathbf{B}_{311}$	$\mathbf{B}_{302}$	$\mathbf{B}_{312}$	$\mathbf{B}_{303}$	$\mathbf{B}_{313}$	$\mathbf{\Sigma}^{\texttt{juv}}_{311}$	$\mathbf{\Sigma}^{\texttt{juv}}_{312}$	$\mathbf{\Sigma}^{\texttt{juv}}_{313}$	$\mathbf{\Sigma}^{\texttt{juv}}_{322}$	$\mathbf{\Sigma}^{\texttt{juv}}_{323}$	$\mathbf{\Sigma}^{\texttt{juv}}_{333}$	$\mathbf{\Sigma}^{\texttt{ad}}_{311}$	$\mathbf{\Sigma}^{\texttt{ad}}_{312}$	$\mathbf{\Sigma}^{\texttt{ad}}_{313}$	$\mathbf{\Sigma}^{\texttt{ad}}_{322}$	$\mathbf{\Sigma}^{\texttt{ad}}_{323}$	$\mathbf{\Sigma}^{\texttt{ad}}_{333}$
2.5%	61.84	-0.48	12.23	0.80	113.23	0.63	1.31	0.08	-0.22	0.47	-0.06	0.60	1.34	0.32	-0.26	0.61	0.10	0.87
50%	62.26	0.03	12.49	1.13	113.53	1.03	1.94	0.39	0.17	0.69	0.16	0.90	1.85	0.59	0.11	0.83	0.32	1.20
97.5%	62.69	0.56	12.75	1.46	113.85	1.43	3.04	0.83	0.63	1.08	0.44	1.37	2.64	0.99	0.47	1.17	0.60	1.71

Quantile	$\mathbf{B}_{401}$	$\mathbf{B}_{411}$	$\mathbf{B}_{402}$	$\mathbf{B}_{412}$	$\mathbf{B}_{403}$	$\mathbf{B}_{413}$	$\mathbf{\Sigma}^{\texttt{juv}}_{411}$	$\mathbf{\Sigma}^{\texttt{juv}}_{412}$	$\mathbf{\Sigma}^{\texttt{juv}}_{413}$	$\mathbf{\Sigma}^{\texttt{juv}}_{422}$	$\mathbf{\Sigma}^{\texttt{juv}}_{423}$	$\mathbf{\Sigma}^{\texttt{juv}}_{433}$	$\mathbf{\Sigma}^{\texttt{ad}}_{411}$	$\mathbf{\Sigma}^{\texttt{ad}}_{412}$	$\mathbf{\Sigma}^{\texttt{ad}}_{413}$	$\mathbf{\Sigma}^{\texttt{ad}}_{422}$	$\mathbf{\Sigma}^{\texttt{ad}}_{423}$	$\mathbf{\Sigma}^{\texttt{ad}}_{433}$
2.5%	56.64	-0.74	12.24	0.40	114.80	0.05	1.28	-0.29	-1.34	0.49	-0.10	0.77	2.36	-0.17	-0.17	0.76	-0.01	0.68
50%	57.27	0.36	12.66	1.16	115.32	0.81	2.17	0.24	-0.38	0.83	0.26	1.30	4.32	0.84	0.80	1.35	0.52	1.24
97.5%	57.89	1.47	13.10	1.89	115.83	1.61	4.00	0.92	0.26	1.52	0.83	2.46	9.07	2.55	2.53	2.77	1.50	2.60

References

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomos Response Data. *Journal of the American Statistical Association*, 88(422):669–679.

Cristianini, N., Shawe-Taylor, J., and Others (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*, volume 1.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 1137–1145. Montreal, Canada.

Malmhagen, B., Karlsson, M., and Menzie, S. (2013). Using wing morphology to separate four species of Acrocephalus warblers in Scandinavia. *Ringing & Migration*, 28(1):63–68.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for Machine Learning*. The MIT press.

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2012). *Bayesian statistics and marketing*. John Wiley & Sons.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Srivastava, S., Gupta, M. R., and Frigyik, B. A. (2007). Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research*, 8(Jun):1277–1305.

Tricker, A. (1984). Effects of rounding on the moments of a probability distribution. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 33(4):381–390.

Walinder, G., Karlsson, L., and Persson, K. (1988). A new method for separating Marsh Warblers Acrocephalus palustris from Reed Warblers A. scirpaceus. *Ringing & Migration*, 9(1):55–62.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.