UTSA The University of Texas at San Antonio™

College of Business

Working Paper SERIES

Date: October 2, 2019 **WP#** 0230MSS

BAYESIAN ANALYSIS OF A MULTIVARIATE DENSITY RATIO MODEL

Victor De Oliveira
The University of Texas at San Antonio
Victor.deoliveira@utsa.edu

BAYESIAN ANALYSIS OF A MULTIVARIATE DENSITY RATIO MODEL

Victor De Oliveira[1]
Department of Management Science and Statistics
The University of Texas at San Antonio
San Antonio, TX 78249, USA
victor.deoliveira@utsa.edu

September 20, 2019

Abstract

Nowadays data are abundant, often being multivariate and coming from different sources. The goal in many of these situations is the efficient combination or fusion of the information from the different sources to answer the question(s) of interest, which results in more efficient and reliable inferences than using a single source. A useful approach to achieve this goal is the use of the so–called density ratio model, which makes minimal assumptions about the several multivariate distributions involved. In this project we propose to investigate methods to perform Bayesian inference from several related multivariate data sources based on a multivariate density ratio model. To test the practical applicability of the proposed methodology, we plan to use a previously analyzed dataset to quantify the effect of height and age on weight of germ cell testicular cancer patients.

Key words: Data fusion; Weighted systems of distributions; Empirical likelihood; Logistic–normal distribution; semiparametric model.

JEL Classifications: C16, C21

[1]This project was partly funded by the University of Texas at San Antonio, Office of the Vice President for Research. I thank Benjamin Kedem for providing the testicular germ cell tumor data, reading a preliminary version of this manuscript and providing fruitful feedback.

1 Introduction

Nowadays data are abundant, often being multivariate and coming from different sources. Examples include case–control data, weather measurements of different quantities from different instruments, and several multivariate time series. The goal in many of these situations is the efficient combination or fusion of the information from the different sources to answer the question(s) of interest. Methods to combine data from different sources are of paramount importance, since inferences from multiple data sources are more efficient and reliable than inferences from a single source. A useful approach to achieve this goal is the use of the so–called density ratio model. This consists of a system of distributions that represents multiple data sources, in which one distribution serves as the *reference* distribution and the rest are *distortions* or deviations from the reference distribution. The proposed model includes both finite– and infinite–dimensional parameters, so it is by construction semiparameteric. Under this approach different inferential goals about these distributions, being univariate or multivariate, can be accomplished with minimal distributional assumptions, specifically without assuming linearity or normality. Examples of the use of *univariate* density ratio models for different inferential tasks include Anderson (1979), Qin and Zhang (1997), Fokianos, Kedem, Qin and Short (2001) and Kedem, Lu, Wei and Williams (2008). A recent overview of these works in given in Kedem, De Oliveira and Sverchkov (2017).

A frequentist approach for the analysis of the *bivariate* density ratio model was developed in Kedem, Kim, Voulgaraki and Graubard (2009), and extended to the general *multivariate* density ratio model in Voulgaraki, Kedem and Graubard (2012). The multivariate extension of the density ratio model is useful for several reasons. First, it provides a way to determine and quantify the difference between two or more multivariate distributions making minimal assumptions about the data generating mechanisms, unlike traditional methods that rely on linearity and normality (e.g., MANOVA). Second, it provides estimates of the distributions that generate the different data sources using the combined data, rather than only one data source. This results in more efficient and reliable estimates. Third, it provides a way to perform semiparametric regression in a way similar to the Nadaraya–Watson kernel regression method. For the latter, once the distribution of one of the data sources has been estimated, this can be used to estimate the required conditional expectation (regression function) by treating one of the variables as the response and the others as the covariates; see Voulgaraki et al. (2012). The analysis of the density ratio model has mainly used the frequentist approach that relies on the notion of empirical likelihood (Owen, 2001) and the methodology developed in Vardi (1982, 1985) for non–parametric inference in biased sampling models.

In spite of its versatility, the Bayesian approach was not explored for the analysis of density ratio models until the recent article by De Oliveira and Kedem (2017) that studies the case of multiple univariate samples. As multiple multivariate datasets have become the norm in

recent years, the extension of the aforementioned Bayesian methodology to the multivariate case is important. Hence this work aims at developing such extension along the lines of the univariate model, but also providing inferences that are particular to the multivariate case. Specifically, the goals of this article are threefold. First, we propose a model that relies on a non–parametric likelihood, similar to that in Kitamura (2007), and a flexible prior for the model components. This prior is tailored to reflect a desirable smoothness property among the elements of the non–parametric component. Second, we develop a Markov chain Monte Carlo algorithm to sample from the posterior distribution of all the model unknowns. Third, we describe methods to perform Bayesian inference for the tasks of estimation of all the multivariate distributions of the multivariate samples, and testing the equidistribution hypothesis (that the distributions of the multiple samples are all equal). In addition, it is shown that, under the density ratio model and a condition about the data, inferences about marginal and conditional distributions as well as the distributions of derived variables are readily available from the estimated multivariate distributions. Finally, we briefly discuss the problem of estimating, for a given multivariate distribution, the conditional expectation of one of the variables given the rest (regression function) in a way that avoids density estimation, and instead used directly the estimated cdfs.

The proposed methodology is illustrated by a case–control study that involves the trivariate data set of testicular germ cell tumor data analyzed by Voulgaraki et al. (2012).

2 A Weighted System of Multivariate Distributions

Below we formulate the extension to the multivariate case of the sampling mechanism described in De Oliveira and Kedem (2017) for the univariate case. We assume we have at our disposal $q+1$ multivariate independent random samples in $\mathbb{R}^p$. The density ratio model asserts that the distribution of one of the samples, considered as the reference, is (essentially) completely unspecified, while the distributions of the other samples are distortions of the reference distribution. Specifically, these random samples follow the sampling structure

$$\begin{aligned}
\mathbf{X}_{01}, \mathbf{X}_{02}, \ldots, \mathbf{X}_{0n_m} &\overset{\text{iid}}{\sim} G(\boldsymbol{x}), \\
\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1} &\overset{\text{iid}}{\sim} G_1(\boldsymbol{x}) \\
\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2} &\overset{\text{iid}}{\sim} G_2(\boldsymbol{x}) \\
&\vdots \\
\mathbf{X}_{q1}, \mathbf{X}_{q2}, \ldots, \mathbf{X}_{qn_q} &\overset{\text{iid}}{\sim} G_q(\boldsymbol{x})
\end{aligned}$$

where for $i = 0, 1, \ldots, q$ and $j = 1, 2 \ldots, n_i$ we have that $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})' \in \mathbb{R}^p$ is the j^{th} datum from the i^{th} sample, and $\boldsymbol{x} = (x_1, \ldots, x_p)'$. The p–dimensional cdf $G(\boldsymbol{x})$ plays the role of the reference distribution, for which we make minimal assumptions (see below), while

the cdfs $G_1(\boldsymbol{x}), \ldots, G_q(\boldsymbol{x})$ are exponential distortions of $G(\boldsymbol{x})$ given by

$$dG_i(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}_i' \boldsymbol{h}(\boldsymbol{x})) dG(\boldsymbol{x})}{\int_{\mathbb{R}^p} \exp(\boldsymbol{\beta}_i' \boldsymbol{h}(\boldsymbol{y})) dG(\boldsymbol{y})}, \qquad i = 1, \ldots, q, \tag{1}$$

where $\boldsymbol{\beta}_i := (\beta_{i1}, \ldots, \beta_{id})' \in \mathbb{R}^d$ are unknown parameters, $d \geq p$, and $\boldsymbol{h}(\boldsymbol{x}) := (h_1(\boldsymbol{x}), \ldots, h_d(\boldsymbol{x}))'$ is a known multivariate vector–valued function. The distributions of all the samples depend on the 'nonparametric' component of the model, $G(\boldsymbol{x})$, while the distribution of the i^{th} distorted sample also depends on the 'parametric' component, $\boldsymbol{\beta}_i$ (as well as the known function $\boldsymbol{h}(\boldsymbol{x})$). It was remarked in De Oliveira and Kedem (2017) that the tilt function $\boldsymbol{h}(\boldsymbol{x})$ is the least interpretable component of the density ratio model. Many choices are possible, but here we follow Kedem et al. (2009) and Voulgaraki et al. (2012) in using $d = p$ and $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}$ as the default. Finally, to simplify the notation that follows we collect all the data from the $q+1$ multivariate independent samples into the single vector

$$\begin{aligned} \boldsymbol{t} &= (\boldsymbol{t}_1', \boldsymbol{t}_2', \ldots, \boldsymbol{t}_n')' \\ &= (\boldsymbol{x}_{01}', \ldots, \boldsymbol{x}_{0n_0}', \boldsymbol{x}_{11}', \ldots, \boldsymbol{x}_{1n_1}', \ldots, \boldsymbol{x}_{q1}', \ldots, \boldsymbol{x}_{qn_q}')', \end{aligned}$$

where $n := n_0 + n_1 + \ldots + n_q$ is the total size of the combined samples.

The modeling of the reference cdf $G(\boldsymbol{x})$ is aimed at making minimal assumptions. Similarly as in Kitamuta (2007) and De Oliveira and Kedem (2017), we assume this distribution belongs to the family of all discrete distributions with support at the observed data vectors, i.e., the 'nonparametric' family of distributions

$$\mathcal{G} = \Big\{ G(\boldsymbol{x}) = \sum_{k=1}^{n} p_k \mathbf{1}\{\boldsymbol{t}_k \leq \boldsymbol{x}\} \ : \ p_k > 0 \text{ for all } k, \text{ and } \sum_{k=1}^{n} p_k = 1 \Big\},$$

where $\mathbf{1}\{E\}$ denotes the indicator function of event E, and $\mathbf{1}\{\boldsymbol{t}_k \leq \boldsymbol{x}\} := \prod_{r=1}^{p} \mathbf{1}\{t_{kr} \leq x_r\}$ for $\boldsymbol{t}_k = (t_{k1}, \ldots, t_{kp})'$. It is clear that most distributions on $\mathbb{R}^p$ with support contained in the convex hull of the observed data can be well approximated by a member of $\mathcal{G}$, provided n is large. So this family is quite flexible and makes few assumptions.

We assume that $\boldsymbol{t}_k \neq \boldsymbol{t}_{k'}$ for any $k \neq k'$ (no two data points are equal), which is the usual case with continuous data. So for any $i = 0, \ldots, q$ and $j = 1, \ldots, n_i$ it holds that there is a unique index $k(ij) \in \{1, \ldots, n\}$ for which $\boldsymbol{t}_{k(ij)} = \boldsymbol{x}_{ij}$; this provides the map between the $\boldsymbol{x}$ and $\boldsymbol{t}$ data labels.

2.1 Likelihood

When $G(\boldsymbol{x})$ belongs to $\mathcal{G}$ it holds that $p_k = dG(\boldsymbol{t}_k)$, the probability of $\boldsymbol{t}_k$ under G. Also, from (1) follows that

$$G_i(\boldsymbol{x}) = \sum_{k=1}^{n} \left(\frac{p_k \exp(\boldsymbol{\beta}_i' \boldsymbol{h}(\boldsymbol{t}_k))}{\sum_{l=1}^{n} p_l \exp(\boldsymbol{\beta}_i' \boldsymbol{h}(\boldsymbol{t}_l))} \right) \mathbf{1}\{\boldsymbol{t}_k \leq \boldsymbol{x}\}, \qquad i = 1, \ldots, q, \tag{2}$$

so for any $i = 1, \ldots, q$ and $j = 1, \ldots, n_i$ we have that the probability of $\boldsymbol{x}_{ij}$ under G_i is

$$dG_i(\boldsymbol{x}_{ij}) = \frac{p_{k(ij)} \exp(\boldsymbol{\beta}_i' \boldsymbol{h}(\boldsymbol{t}_{k(ij)}))}{\sum_{l=1}^{n} p_l \exp(\boldsymbol{\beta}_i' \boldsymbol{h}(\boldsymbol{t}_l))}.$$

The semiparametric multivariate density ratio model defined above is parametrized by $(\boldsymbol{\beta}', \boldsymbol{p}_-')' \in \mathbb{R}^{qd} \times \mathbb{S}^{n-1}$, where $\boldsymbol{\beta} := (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_q')'$, $\boldsymbol{p} = (p_1, \ldots, p_n)' = (\boldsymbol{p}_-', p_n)$, with $\boldsymbol{p}_- = (p_1, \ldots, p_{n-1})'$, $p_n = 1 - \sum_{k=1}^{n-1} p_k$ and

$$\mathbb{S}^{n-1} = \{\boldsymbol{p}_- \in \mathbb{R}^{n-1} \ : \ p_k > 0 \text{ for all } k, \text{ and } \sum_{k=1}^{n-1} p_k < 1\},$$

is the unit simplex in $\mathbb{R}^{n-1}$. Then the likelihood function of $(\boldsymbol{\beta}', \boldsymbol{p}_-')'$ based on the (combined) $q+1$ samples $\boldsymbol{t}$ is give by

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{p}_-; \boldsymbol{t}) &= \prod_{j=1}^{n_0} dG(\boldsymbol{x}_{0j}) \cdot \prod_{j=1}^{n_1} dG_1(\boldsymbol{x}_{1j}) \cdots \prod_{j=1}^{n_q} dG_q(\boldsymbol{x}_{qj}) \\ &= \prod_{k=1}^{n_0} p_{k(0j)} \cdot \prod_{j=1}^{n_1} \frac{p_{k(1j)} \exp(\boldsymbol{\beta}_1' \boldsymbol{h}(\boldsymbol{t}_{k(1j)}))}{\sum_{l=1}^{n} p_l \exp(\boldsymbol{\beta}_1' \boldsymbol{h}(\boldsymbol{t}_l))} \cdots \prod_{j=1}^{n_q} \frac{p_{k(qj)} \exp(\boldsymbol{\beta}_q' \boldsymbol{h}(\boldsymbol{t}_{k(qj)}))}{\sum_{l=1}^{n} p_l \exp(\boldsymbol{\beta}_q' \boldsymbol{h}(\boldsymbol{t}_l))} \\ &= \frac{\prod_{k=1}^{n} p_k \cdot \exp\left(\boldsymbol{\beta}_1' \sum_{j=1}^{n_1} \boldsymbol{h}(\boldsymbol{t}_{k(1j)}) + \cdots + \boldsymbol{\beta}_q' \sum_{j=1}^{n_q} \boldsymbol{h}(\boldsymbol{t}_{k(qj)})\right)}{\left(\sum_{l=1}^{n} p_l \exp(\boldsymbol{\beta}_1' \boldsymbol{h}(\boldsymbol{t}_l))\right)^{n_1} \cdots \left(\sum_{l=1}^{n} p_l \exp(\boldsymbol{\beta}_q' \boldsymbol{h}(\boldsymbol{t}_l))\right)^{n_q}}, \end{aligned} \tag{3}$$

for $\boldsymbol{\beta} \in \mathbb{R}^{qd}$ and $\boldsymbol{p}_- \in \mathbb{S}^{n-1}$, and zero otherwise.

2.2 Prior

The Bayesian approach has the ability to incorporate into the analysis contextual or subjective prior information. We assume that $\boldsymbol{\beta} := (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_q')'$ and $\boldsymbol{p}_-$ are independent a priori. In addition, we assume that $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q$ are independent a priori and for $i = 1, \ldots, q$, $\boldsymbol{\beta}_i \sim \mathrm{N}_d(\boldsymbol{b}_{0i}, B_{0i})$ with $\boldsymbol{b}_{0i} \in \mathbb{R}^d$ and B_{0i} a positive matrix, so

$$\pi(\boldsymbol{\beta}) \propto \Big(\prod_{i=1}^{q} |B_{0i}|\Big)^{-1/2} \exp\Big(-\frac{1}{2}\sum_{i=1}^{q} (\boldsymbol{\beta}_i - \boldsymbol{b}_{0i})' B_{0i}^{-1} (\boldsymbol{\beta}_i - \boldsymbol{b}_{0i})\Big). \tag{4}$$

A possible default (neutral) prior is obtained by setting $\boldsymbol{b}_{0i} = \boldsymbol{0}_d$ and $B_{0i} = v_{0i} I_d$ for all i, with $v_{0i} > 0$ and I_d the $d \times d$ identity matrix. This represents the prior belief that all $q+1$ samples are equally distributed, with the v_{0i} moderating the strength of this belief. In this case we have

$$\pi(\boldsymbol{\beta}) \propto \Big(\prod_{i=1}^{q} v_{0i}\Big)^{-d/2} \exp\Big(-\frac{1}{2}\sum_{i=1}^{q} \frac{\boldsymbol{\beta}_i' \boldsymbol{\beta}_i}{v_{0i}}\Big). \tag{5}$$

Recall that for any $k = 1, \ldots, n$, p_k is the probability of $\boldsymbol{t}_k$ under the reference distribution G. If $\boldsymbol{t}_k$ and $\boldsymbol{t}_{k'}$ are two support points that are close to each other, we expect p_k and $p_{k'}$ to also be close to each other. This is a key smoothness property that a sensible prior for $\boldsymbol{p}_-$

should satisfy. In addition, it may be desirable to use a prior that is 'non–informative' in the the sense that allows the support points to be equally likely on average, or approximately so. With these desiderata in mind, we aim to construct a prior for $\boldsymbol{p}_-$ that allows for the following properties to hold for any $k, k' = 1, \ldots, n-1$:

(a) Neutrality: It holds exactly or approximately that $E(p_k) = 1/n$;

(b) Smoothness: It holds that $\text{corr}(p_k, p_{k'}) \to 1$ as $||\boldsymbol{t}_k - \boldsymbol{t}_{k'}|| \to 0$.

In order to respect the restriction that $\boldsymbol{p}_-$ must belong to the simplex $\mathbb{S}^{n-1}$, a prior is placed on a suitable transformation of $\boldsymbol{p}_-$, which in turn induces a prior for $\boldsymbol{p}_-$. We use the one–to–one transformation $H : \mathbb{S}^{n-1} \to \mathbb{R}^{n-1}$ given by

$$H(\boldsymbol{p}_-) = \Big(\log\Big(\frac{p_1}{p_n}\Big), \ldots, \log\Big(\frac{p_{n-1}}{p_n}\Big)\Big)',$$

which was proposed and studied by Aitchison and Shen (1980) to analyze compositional data. We would use the normal prior

$$H(\boldsymbol{p}_-) \sim \text{N}_{n-1}\big(\boldsymbol{m}_0, V_0\big),$$

where $\boldsymbol{m}_0$ and V_0 are hyperparameters, so the induced prior for $\boldsymbol{p}_-$ is the so–called logistic–normal distribution with pdf

$$\pi(\boldsymbol{p}_-) \propto \Big(\prod_{k=1}^{n} p_k\Big)^{-1} \exp\Big\{ -\frac{1}{2}\big(H(\boldsymbol{p}_-) - \boldsymbol{m}_0\big)' V_0^{-1} \big(H(\boldsymbol{p}_-) - \boldsymbol{m}_0\big)\Big\}\, \mathbf{1}\{\boldsymbol{p}_- \in \mathbb{S}^{n-1}\}. \quad (6)$$

As default choices we propose using $\boldsymbol{m}_0 = -\frac{m_0}{2}\mathbf{1}_{n-1}$ and $(V_0)_{kk'} = m_0 K_0(||\boldsymbol{t}_k - \boldsymbol{t}_{k'}||)$, with $m_0 > 0$ a hyperparameter, $\mathbf{1}_{n-1}$ a vector of ones, and $K_0(u)$ a known continuous and isotropic correlation function in $\mathbb{R}^p$ with the property $\lim_{u\to\infty} K_0(u) = 0$. One possible choice for $K_0(u)$ to use a correlation function with compact support (one that vanishes beyond a certain threshold distance) which would speed up sampling from the posterior distribution of $(\boldsymbol{\beta}', \boldsymbol{p}_-')'$, to be discussed in the next section. With these hyperparameter choices and from the formulas of the moments of the logistic–normal distribution (Aitchison and Shen, 1980; De Oliveira and Kedem, 2017) we have that for any $k, k' = 1, \ldots, n-1$

$$E\Big(\frac{p_k}{p_n}\Big) = 1 \quad \text{and} \quad \text{corr}\Big(\frac{p_k}{p_n}, \frac{p_{k'}}{p_n}\Big) = \frac{\exp\big(m_0 K(||\boldsymbol{t}_k - \boldsymbol{t}_{k'}||)\big) - 1}{\exp(m_0) - 1},$$

so this prior brings about the desired smoothness property for the ratios p_k/p_n. Since these ratios all have a common denominator and $\sum_{k=1}^{n} p_k = 1$, desiderata (a) and (b) hold (approximately) for the p_k.

Remark 1

One alternative approach to construct a prior for $\boldsymbol{p}_-$ could be to use a Markov random field with a neighborhood structure depending on distance. Specifically, by considering $\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n\}$

as a set of 'sites' in $\mathbb{R}^p$, a neighborhood system $\{N_1, \ldots, N_n\}$ could be defined where N_k, the collection of sites that are neighbors of $\boldsymbol{t}_k$, is

$$N_k = \{\boldsymbol{t}_{k'} : 0 < ||\boldsymbol{t}_k - \boldsymbol{t}_{k'}|| < c\}, \qquad \text{for some } c > 0.$$

Granville (1996) used such approach for multivariate density estimation. As before, a Gaussian Markov random field model could be placed on $H(\boldsymbol{p}_-)$, which induces a Markov random field for $\boldsymbol{p}_-$. But achieving the aforementioned desiderata (a)–(b) with a Markov random field prior seems to be quite difficult, since there is no explicit expression for the mean and covariance structure of $\boldsymbol{p}_-$ under such prior, and the implicit correlations may have unintuitive and/or undesirable behaviors (Wall, 2004). Another alternative approach, proposed for the analyses of some one–sample nonparametric Bayesian models in Rubin (1981) and Chamberlain and Imbens (2003), is to use a Dirichlet prior for $\boldsymbol{p}_-$. But this prior only represents negative dependence among the p_k, so it is inadequate for the present model.

3 Posterior Simulation

Based on the likelihood (3) and the choice of prior for $(\boldsymbol{\beta}', \boldsymbol{p}_-')'$ given in (4) and (6), the posterior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{p}_- \mid \boldsymbol{t})$ is proportional to

$$\frac{\exp\left\{\sum_{i=1}^{q}\left[\boldsymbol{\beta}_i'\sum_{j=1}^{n_i}\boldsymbol{h}(\boldsymbol{t}_{k(ij)}) - \frac{1}{2}(\boldsymbol{\beta}_i - \boldsymbol{b}_{0i})'B_{0i}^{-1}(\boldsymbol{\beta}_i - \boldsymbol{b}_{0i})\right] - \frac{1}{2}(H(\boldsymbol{p}_-) - \boldsymbol{m}_0)'V_0^{-1}(H(\boldsymbol{p}_-) - \boldsymbol{m}_0)\right\}}{\left(\sum_{l=1}^{n} p_l \exp(\boldsymbol{\beta}_1'\boldsymbol{h}(\boldsymbol{t}_l))\right)^{n_1} \cdots \left(\sum_{l=1}^{n} p_l \exp(\boldsymbol{\beta}_q'\boldsymbol{h}(\boldsymbol{t}_l))\right)^{n_q}}, \tag{7}$$

for $\boldsymbol{\beta} \in \mathbb{R}^{qd}$ and $\boldsymbol{p}_- \in \mathbb{S}^{n-1}$, and zero otherwise. This posterior distribution is highly non–standard and complex, so Bayesian inference about $(\boldsymbol{\beta}', \boldsymbol{p}_-')'$ requires the use of Markov chain Monte Carlo methods (Gamerman and Lopes, 2006). To sample from this posterior distribution we extend to the multivariate case the Metropolis–Hastings algorithm used in De Oliveira and Kedem (2017), in which the parameters are updated separately for the two blocks $\boldsymbol{\beta}$ and $\boldsymbol{p}_-$.

By inspection of (7) we have that the full posterior distribution of $\boldsymbol{\beta}$ is given by

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{p}_-, \boldsymbol{t}) \propto \frac{\exp\left\{\sum_{i=1}^{q}\left[\boldsymbol{\beta}_i'\sum_{j=1}^{n_i}\boldsymbol{h}(\boldsymbol{t}_{k(ij)}) - \frac{1}{2}(\boldsymbol{\beta}_i - \boldsymbol{b}_{0i})'B_{0i}^{-1}(\boldsymbol{\beta}_i - \boldsymbol{b}_{0i})\right]\right\}}{\left(\sum_{l=1}^{n} p_l \exp(\boldsymbol{\beta}_1'\boldsymbol{h}(\boldsymbol{t}_l))\right)^{n_1} \cdots \left(\sum_{l=1}^{n} p_l \exp(\boldsymbol{\beta}_q'\boldsymbol{h}(\boldsymbol{t}_l))\right)^{n_q}}.$$

Let $(\boldsymbol{\beta}', \boldsymbol{p}_-')'$ denote the current state of the chain, and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*\prime}, \ldots, \boldsymbol{\beta}_q^{*\prime})' \in \mathbb{R}^{qd}$ be the candidate for the first block of the state in the next iteration. The Metropolis–Hastings update of the first block would be done by first simulating a candidate $\boldsymbol{\beta}^*$ using a random–walk with proposal $q_1(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ equal to the $\mathrm{N}_{qd}(\boldsymbol{\beta}, c_1 I_{qd})$ distribution, where $c_1 > 0$ is a tuning constant. After the candidate $\boldsymbol{\beta}^*$ is simulated, this is accepted with probability

$$\alpha_1(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\beta}^* \mid \boldsymbol{p}_-, \boldsymbol{t}) q_1(\boldsymbol{\beta}^*, \boldsymbol{\beta})}{\pi(\boldsymbol{\beta} \mid \boldsymbol{p}_-, \boldsymbol{t}) q_1(\boldsymbol{\beta}, \boldsymbol{\beta}^*)}\right\} = \min\{1, \xi_1\}, \tag{8}$$

where

$$\xi_1 = \left\{\frac{\sum_{l=1}^n p_l \exp(\boldsymbol{\beta}_1'\boldsymbol{h}(\boldsymbol{t}_l))}{\sum_{l=1}^n p_l \exp(\boldsymbol{\beta}_1^{*\prime}\boldsymbol{h}(\boldsymbol{t}_l))}\right\}^{n_1} \cdots \left\{\frac{\sum_{l=1}^n p_l \exp(\boldsymbol{\beta}_q'\boldsymbol{h}(\boldsymbol{t}_l))}{\sum_{l=1}^n p_l \exp(\boldsymbol{\beta}_q^{*\prime}\boldsymbol{h}(\boldsymbol{t}_l))}\right\}^{n_q}$$
$$\times \exp\Big\{\sum_{i=1}^q \Big[(\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_i)' \sum_{j=1}^{n_i} \boldsymbol{h}(\boldsymbol{t}_{k(ij)}) - \frac{1}{2}\Big((\boldsymbol{\beta}_i^* - \boldsymbol{b}_{0i})' B_{0i}^{-1}(\boldsymbol{\beta}_i^* - \boldsymbol{b}_{0i}) - (\boldsymbol{\beta}_i - \boldsymbol{b}_{0i})' B_{0i}^{-1}(\boldsymbol{\beta}_i - \boldsymbol{b}_{0i})\Big)\Big]\Big\},$$

since $q_1(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = q_1(\boldsymbol{\beta}^*, \boldsymbol{\beta})$. If the candidate is not accepted, the next state is set equal to the current state. When the default prior (5) is used the argument in the exponential function above simplifies to

$$\sum_{i=1}^q \Big[(\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_i)' \sum_{j=1}^{n_i} \boldsymbol{h}(\boldsymbol{t}_{k(ij)}) - \frac{1}{2v_{0i}}(\boldsymbol{\beta}_i^{*\prime}\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_i'\boldsymbol{\beta}_i)\Big].$$

The guideline for choosing the tuning constant c_1 is to set it at a value, chosen by trial and error, that results in an empirical acceptance probability (8) in the range 0.3–0.5. This strategy often leads to efficient algorithms when, as in this case, a random–walk Metropolis–Hastings update with a normal proposal is used (Gamerman and Lopes, 2006).

Likewise, it follows from (7) that the full posterior distributions of $\boldsymbol{p}_-$ is

$$\pi(\boldsymbol{p}_- \mid \boldsymbol{\beta}, \boldsymbol{t}) \propto \frac{\exp\big\{-\frac{1}{2}(H(\boldsymbol{p}_-) - \boldsymbol{m}_0)' V_0^{-1}(H(\boldsymbol{p}_-) - \boldsymbol{m}_0)\big\}}{\big(\sum_{l=1}^n p_l \exp(\boldsymbol{\beta}_1'\boldsymbol{h}(\boldsymbol{t}_l))\big)^{n_1} \cdots \big(\sum_{l=1}^n p_l \exp(\boldsymbol{\beta}_q'\boldsymbol{h}(\boldsymbol{t}_l))\big)^{n_q}} \, \mathbf{1}\{\boldsymbol{p}_- \in \mathbb{S}^{n-1}\}.$$

Let $\boldsymbol{p}_-^* = (p_1^*, \ldots, p_{n-1}^*)'$ denote the candidate for the second block of the state in the next iteration. Following De Oliveira and Kedem (2017), we use a Metropolis–Hastings update of $\boldsymbol{p}_-$ by simulating a candidate $\boldsymbol{p}_-^*$ using an independence proposal distribution $q_2(\boldsymbol{p}_-, \boldsymbol{p}_-^*)$ equal to a scaled version of the prior distribution (6), meaning that V_0 is replaced with $c_2 V_0$, with $c_2 > 0$ a tuning constant. After the candidate $\boldsymbol{p}_-^*$ is simulated, this is accepted with probability

$$\alpha_2(\boldsymbol{p}_-, \boldsymbol{p}_-^*) = \min\Big\{1, \frac{\pi(\boldsymbol{p}_-^* \mid \boldsymbol{\beta}, \boldsymbol{t}) q_2(\boldsymbol{p}_-^*, \boldsymbol{p}_-)}{\pi(\boldsymbol{p}_- \mid \boldsymbol{\beta}, \boldsymbol{t}) q_2(\boldsymbol{p}_-, \boldsymbol{p}_-^*)}\Big\} = \min\{1, \xi_2\}, \tag{9}$$

where

$$\xi_2 = \Big(\prod_{i=1}^n \frac{p_i^*}{p_i}\Big)\left\{\frac{\sum_{l=1}^n p_l \exp(\boldsymbol{\beta}_1'\boldsymbol{h}(\boldsymbol{t}_l))}{\sum_{l=1}^n p_l^* \exp(\boldsymbol{\beta}_1'\boldsymbol{h}(\boldsymbol{t}_l))}\right\}^{n_1} \cdots \left\{\frac{\sum_{l=1}^n p_l \exp(\boldsymbol{\beta}_q'\boldsymbol{h}(\boldsymbol{t}_l))}{\sum_{l=1}^n p_l^* \exp(\boldsymbol{\beta}_q'\boldsymbol{h}(\boldsymbol{t}_l))}\right\}^{n_q}$$
$$\times \exp\Big\{\Big(\frac{c_2 - 1}{2c_2}\Big)\Big[\big(H(\boldsymbol{p}_-) - \boldsymbol{m}_0\big)' V_0^{-1}\big(H(\boldsymbol{p}_-) - \boldsymbol{m}_0\big) - \big(H(\boldsymbol{p}_-^*) - \boldsymbol{m}_0\big)' V_0^{-1}\big(H(\boldsymbol{p}_-^*) - \boldsymbol{m}_0\big)\Big]\Big\}.$$

If the candidate is not accepted, the next state is set equal to the current state. The tuning constant $c - 2$ is also chosen by trial and error so that empirical acceptance probability (9) in the range 0.3–0.5. Note that ξ_2 simplifies substantially when $c_2 = 1$ (i.e., when the proposal distribution equals the prior distribution of $\boldsymbol{p}_-$).

We summarize below the MCMC algorithm to simulate a Markov chain $\{(\boldsymbol{\beta}^{(m)}, \boldsymbol{p}_-^{(m)}) : m = 1, \ldots, M\}$ whose equilibrium distribution is $\pi(\boldsymbol{\beta}, \boldsymbol{p}_- \mid \boldsymbol{t})$, where $\boldsymbol{\beta}^{(m)} := (\boldsymbol{\beta}_1^{(m)\prime}, \ldots, \boldsymbol{\beta}_q^{(m)\prime})'$ and $\boldsymbol{p}_-^{(m)} := (p_1^{(m)}, \ldots, p_{n-1}^{(m)})'$:

Algorithm

Step 1. Choose the hyperparameters $\boldsymbol{b}_{0i}$, B_{0i} (for $i = 1, \ldots, q$), $\boldsymbol{m}_0$, V_0, the tuning constants c_1, c_2, and the initial state $(\boldsymbol{\beta}^{(0)\prime}, \boldsymbol{p}_-^{(0)\prime})'$.

For $m = 1, \ldots, M$ do the following:

Step 2. Simulate independently $\boldsymbol{\beta}^* \sim \mathrm{N}_{qd}\big(\boldsymbol{\beta}^{(m-1)}, c_1 I_{qd}\big)$ and $U_1 \sim \mathrm{unif}(0,1)$, and set

$$\boldsymbol{\beta}^{(m)} = \begin{cases} \boldsymbol{\beta}^* & \text{if } \ U_1 < \alpha_1(\boldsymbol{\beta}^{(m-1)}, \boldsymbol{\beta}^*) \\ \boldsymbol{\beta}^{(m-1)} & \text{otherwise} \end{cases},$$

where $\alpha_1(\cdot,\cdot)$ is given by (8).

Step 3. Simulate independently $\mathbf{W}^* = (W_1^*, \ldots, W_{n-1}^*)' \sim \mathrm{N}_{n-1}\big(\boldsymbol{m}_0, c_2 V_0\big)$ and $U_2 \sim \mathrm{unif}(0,1)$, and compute

$$\boldsymbol{p}_-^* = \big(1 + \sum_{i=1}^{n-1} e^{W_i^*}\big)^{-1}\big(e^{W_1^*}, \ldots, e^{W_{n-1}^*}\big)'.$$

Step 4. Set

$$\boldsymbol{p}_-^{(m)} = \begin{cases} \boldsymbol{p}_-^* & \text{if } \ U_2 < \alpha_2(\boldsymbol{p}_-^{(m-1)}, \boldsymbol{p}_-^*) \\ \boldsymbol{p}_-^{(m-1)} & \text{otherwise} \end{cases},$$

where $\alpha_2(\cdot,\cdot)$ is given by (9), and $p_n^{(m)} = 1 - \mathbf{1}'\boldsymbol{p}_-^{(m)}$.

4 Bayesian Inference

4.1 Estimation

Once a large sample $\{(\boldsymbol{\beta}^{(m)\prime}, \boldsymbol{p}_-^{(m)\prime})' : m = 1, \ldots, M\}$ from the posterior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{p}_- \mid \boldsymbol{t})$ is available, Bayesian estimates of of the quantities of interest follow in the usual way. Point estimates and confidence regions of the parameters in $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q$ can be computed in the usual ways using sample averages and quantiles from the corresponding simulated chains. Likewise, for any $\boldsymbol{x} \in \mathbb{R}^p$ a Bayesian estimate of the reference cdf $G(\boldsymbol{x})$ is given by its posterior expectation

$$\hat{G}^B(\boldsymbol{x}) = E\big(G(\boldsymbol{x}) \mid \boldsymbol{t}\big) = \sum_{k=1}^{n} E(p_k \mid \boldsymbol{t}) \mathbf{1}\{\boldsymbol{t}_k \le \boldsymbol{x}\},$$

and Bayesian estimates of the distorted cdfs $G_1(\boldsymbol{x}), \ldots, G_q(\boldsymbol{x})$ are given by

$$\hat{G}_i^B(\boldsymbol{x}) = E\big(G_i(\boldsymbol{x}) \mid \boldsymbol{t}\big) = \sum_{k=1}^{n} E\Big\{ \frac{p_k \exp(\boldsymbol{\beta}_i' \boldsymbol{h}(\boldsymbol{t}_k))}{\sum_{l=1}^{n} p_l \exp(\boldsymbol{\beta}_i' \boldsymbol{h}(\boldsymbol{t}_l))} \mid \boldsymbol{t} \Big\} \mathbf{1}\{\boldsymbol{t}_k \le \boldsymbol{x}\} \quad \text{for } i = 1, \ldots, q,$$

where for each $k = 1, \ldots, n$, $E(p_k \mid \boldsymbol{t})$ and $E\Big\{ \frac{p_k \exp(\boldsymbol{\beta}_i' \boldsymbol{h}(\boldsymbol{t}_k))}{\sum_{l=1}^{n} p_l \exp(\boldsymbol{\beta}_i' \boldsymbol{h}(\boldsymbol{t}_l))} \mid \boldsymbol{t} \Big\}$ are approximated by sample averages computed from the simulated chain

$$E(p_k \mid \boldsymbol{t}) \approx \frac{1}{M} \sum_{m=1}^{M} p_k^{(m)};$$

$$E\Big\{\frac{p_k\exp(\boldsymbol{\beta}_i'\boldsymbol{h}(\boldsymbol{t}_k))}{\sum_{l=1}^n p_l\exp(\boldsymbol{\beta}_i'\boldsymbol{h}(\boldsymbol{t}_l))}\mid \boldsymbol{t}\Big\}\approx\frac{1}{M}\sum_{m=1}^M\frac{p_k^{(m)}\exp(\boldsymbol{\beta}_i^{(m)\prime}\boldsymbol{h}(\boldsymbol{t}_k))}{\sum_{l=1}^n p_l^{(m)}\exp(\boldsymbol{\beta}_i^{(m)\prime}\boldsymbol{h}(\boldsymbol{t}_l))}.$$

Once estimates for the joint cdfs are in place, univariate and multivariate marginal cdfs for any of the variables under study and any of the $q+1$ distributions can also be obtained. For instance, if $\boldsymbol{x}=(x_1,\ldots,x_p)'$ and $\boldsymbol{t}_k=(t_{k1},\ldots,t_{kp})$, then an estimate of the univariate cdf of the first variable under the ith distribution and an estimate of the bivariate cdf of the first two variables under the reference distribution are given, respectively (with a slight abuse of notation), by

$$\begin{aligned}\hat{G}_i^B(x_1) &= \hat{G}_i^B(x_1,\infty,\ldots,\infty)\\ &= \sum_{k=1}^n E\Big\{\frac{p_k\exp(\boldsymbol{\beta}_i'\boldsymbol{h}(\boldsymbol{t}_k))}{\sum_{l=1}^n p_l\exp(\boldsymbol{\beta}_i'\boldsymbol{h}(\boldsymbol{t}_l))}\mid \boldsymbol{t}\Big\}\mathbf{1}\{t_{k1}\le x_1\},\quad \text{for } i=1,\ldots,q\\ \hat{G}^B(x_1,x_2) &= \hat{G}^B(x_1,x_2,\infty,\ldots,\infty) \;=\; \sum_{k=1}^n E(p_k\mid\boldsymbol{t})\mathbf{1}\{t_{k1}\le x_1\}\mathbf{1}\{t_{k2}\le x_2\}.\end{aligned}$$

4.2 Test of Hypotheses

Under the density ratio model, testing the equidistribution hypothesis, namely, that all cdfs $G(\boldsymbol{x}), G_1(\boldsymbol{x}),\ldots,G_q(\boldsymbol{x})$ are equal, amounts to test the null hypothesis

$$\mathrm{H}_0:\boldsymbol{\beta}_1=\cdots=\boldsymbol{\beta}_q=\mathbf{0}_d,$$

against the alternative $\mathrm{H}_1:\boldsymbol{\beta}_i\neq\mathbf{0}_d$ for at least one $i=1,\ldots,q$. Let M_1 be the Bayesian model specified by the likelihood $L(\boldsymbol{\beta},\boldsymbol{p}_-;\boldsymbol{t})$ in (3) and prior $\pi(\boldsymbol{\beta},\boldsymbol{p}_-)$ obtained by the product of (4) and (6), with $(\boldsymbol{\beta}',\boldsymbol{p}_-')'\in\mathbb{R}^{qd}\times\mathbb{S}^{n-1}$, and let M_0 be the Bayesian model specified by the likelihood $L_0(\boldsymbol{p}_-;\boldsymbol{t})=\prod_{k=1}^n p_k\cdot\mathbf{1}\{\boldsymbol{p}_-\in\mathbb{S}^{n-1}\}$ and prior $\pi_0(\boldsymbol{p}_-)$ in (6), with $\boldsymbol{p}_-\in\mathbb{S}^{n-1}$. Testing H_0 versus H_1 is then equivalent to choosing between M_0 and M_1, which would be based on the *Bayes factor* in favor of M_0. If w_0 denotes the prior probability of M_0, then the Bayes factor in favor of M_0 is given by

$$\begin{aligned}\mathrm{BF}_{01}(\boldsymbol{t}) &= \frac{P(\mathrm{M}_0\mid\boldsymbol{t})/(1-P(\mathrm{M}_0\mid\boldsymbol{t}))}{\pi_0/(1-\pi_0)}\\ &= \frac{\pi_1(\mathbf{0}_{qd}\mid\boldsymbol{t})}{\pi_1(\mathbf{0}_{qd})},\end{aligned}\tag{10}$$

where $P(\mathrm{M}_0\mid\boldsymbol{t})$ is the posterior probability of M_0, $\pi_1(\cdot)$ and $\pi_1(\cdot\mid\boldsymbol{t})$ are, respectively, the marginal prior and posterior densities of $\boldsymbol{\beta}$ under M_1. The last identify follows from the prior independence of $\boldsymbol{\beta}$ and $\boldsymbol{p}_-$ and the fact that model M_0 is nested within model M_1; see De Oliveira and Kedem (2017) for details.

As explained in De Oliveira and Kedem (2017), direct computation $\mathrm{BF}_{01}(\boldsymbol{t})$ using (10) is unstable, so we carry out the model selection using the Bayesian Information Criterion

(BIC). Under some regularity conditions and for large samples, the BIC provides a reliable approximation to the logarithm of the Bayes factor that does not depend on the priors for the model parameters (Kass and Raftery, 1995). For the problem of selecting between M_0 and M_1 this becomes

$$\begin{aligned}
\log\big(B_{01}(\boldsymbol{t})\big) &\approx S_{01}(\boldsymbol{t}) \\
&= \log\big(L_0(\hat{\boldsymbol{p}}_-^{(0)};\boldsymbol{t})\big) - \log\big(L_1(\hat{\boldsymbol{\beta}}^{(1)},\hat{\boldsymbol{p}}_-^{(1)};\boldsymbol{t})\big) - \frac{1}{2}(d_0-d_1)\log(n) \\
&= \Big(\frac{qd}{2}-n\Big)\log(n) - \sum_{k=1}^{n}\log(\hat{p}_k^{(1)}) - \sum_{i=1}^{q}\hat{\boldsymbol{\beta}}_i^{(1)\prime}\sum_{j=1}^{n_i}\boldsymbol{h}(\boldsymbol{t}_{k_{ij}}) \\
&\qquad + \sum_{i=1}^{q} n_i \log\Big(\sum_{l=1}^{n}\hat{p}_l^{(1)}\exp(\hat{\boldsymbol{\beta}}_i^{(1)\prime}\boldsymbol{h}(\boldsymbol{t}_l))\Big), \qquad (11)
\end{aligned}$$

where $\hat{\boldsymbol{p}}_-^{(0)}$ and $(\hat{\boldsymbol{\beta}}^{(1)\prime},\hat{\boldsymbol{p}}_-^{(1)\prime})$ are, respectively, the maximum likelihood estimates of $\boldsymbol{p}_-$ and $(\boldsymbol{\beta}',\boldsymbol{p}_-')'$ under M_0 and M_1, and d_i is the number of parameters in M_i for $i=0,1$. It holds that $\hat{\boldsymbol{p}}_-^{(0)} = \frac{1}{n}\mathbf{1}_{n-1}$ (Owen, 2001), while $(\hat{\boldsymbol{\beta}}^{(1)\prime},\hat{\boldsymbol{p}}_-^{(1)\prime})'$ is computed numerically by following a profiling procedure and the method of Lagrange multipliers; see Fokianos et al. (2001) and Voulgaraki et al. (2012) for details.

5 Example

In this section we explore possible connections between the variables age, height and weight and the incidence of testicular germ cell tumor (TGCT), a common cancer among young adult males, using a data set previously analyzed in Voulgaraki et al. (2012). Previous analyses have suggested that increased height may be a risk factor for testicular cancer, while the body mass index[2] did not appear to be associated with testicular cancer. On the other hand, the analysis in Kedem et al. (2009) suggested that when height and weight are considered jointly, they are a significant risk factor. Voulgaraki et al. (2012) included age in their analysis as it may be a potential confounder, given that it is correlated with height and weight and testicular cancer incidence varies by age. See McGlynn and Cook (2010) for further details about the etiology of this disease. The data set consists of age (years), height (centimeters) and weight (kilograms) measurements from $n = 1691$ U.S. males (mostly young). Of those subjects, $n_1 = 763$ were *cases* (had TGCT) and $n_0 = 928$ were *controls* (did not have TGCT).

For this data set $p = 3$, $q = 1$ and the control group is considered the reference. The variables are labeled as (X_1, X_2, X_2) = (age, height, weight). We use the neutral prior distribution described in Section 2.2, where $\boldsymbol{\beta}_1 = (\beta_1,\beta_2,\beta_3)$ and $\boldsymbol{p}_-$ are assumed independent a priori, with $\boldsymbol{\beta}_1 \sim \mathrm{N}_3(\mathbf{0}_3, 50I_3)$ (so $\boldsymbol{b}_{0i} = \mathbf{0}_3$ and $v_{01} = 50$), and $\boldsymbol{p}_- \sim \mathrm{N}_{1690}(-0.005\mathbf{1}_{1690}, 0.01K_0)$

[2] body mass index = weight (kilograms)/height^2 (meters^2).

(so $m_0 = 0.01$). The entries of K_0 are given by the so–called Wendland–Gneiting correlation function

$$K_0(u_{kk'}) = \Big(1 + a\frac{u_{kk'}}{r_0}\Big)\Big(1 - \frac{u_{kk'}}{r_0}\Big)^a \mathbf{1}_{[0,r_0]}(u_{kk'}),$$

where $u_{kk'} = ||\boldsymbol{t}_k - \boldsymbol{t}_{k'}||$, $a = \lceil (p+5)/2 \rceil$, $\lceil \cdot \rceil$ is the ceiling function, and $r_0 > 0$ is a hyperparameter. This choice of a guarantees that $K_0(\cdot)$ is a correlation function in $\mathbb{R}^p$ that is twice differentiable at the origin (Gneiting, 2002). The hyperparameter r_0 can be subjectively chosen based on the density of the observed points $\{\boldsymbol{t}_k\}$ so the matrix K_0 becomes relatively sparse. For the TGCT data we choose $r_0 = 18$, which is close to the average Euclidean distance between pairs of points in $\boldsymbol{t}$.

We ran the MCMC algorithm with the tuning constants $c_1 = 0.00002$ and $c_2 = 0.4$, for $M = 5500$ iterations and a burn–in period of 500. The Metropolis–Hastings updates for $\boldsymbol{\beta}_1$ and $\boldsymbol{p}_-$ had empirical acceptance rates 0.45 and 0.44, respectively. Figure 1 displays a summary of the MCMC output. The first column displays the trace plots of β_1, β_2 and β_3 as well as two components of $\boldsymbol{p}_-$ chosen at random, p_{673} and p_{1055}, and the second column displays the estimated autocorrelation functions of these traces. These show that the algorithm is efficient, as the chain mixes well and has relatively low autocorrelations. The third column displays the estimated marginal posteriors. The posterior means of β_1, β_2 and β_3 are, respectively, -0.0023, 0.0142 and 0.0015, and their equal–tail 95% credible intervals are $(-0.0151, 0.0106)$, $(0.0016, 0.0259)$ and $(-0.0056, 0.0087)$.

Figure 2 displays the estimated marginal cdfs of the variables height (left) and weight (right) for the control group (top) using the density ratio model (DRM), as well as the marginal cdfs of the same variables for the case group (bottom). These were computed by maginalizing the estimated joint cdfs as indicated at the end of Section 4.1. For comparison the respective marginal empirical cdfs of these variables under both control and case groups are also overlaid (blue lines). The empirical estimates for the cdf of a variable under a group use only the data for that variable in that group, while the DRM estimates use the data from all variables in both groups. Except for the top–left plot, the respective DRM and empirical estimates are close.

Figure 3 displays the estimated cdfs of the variables height (left) and weight (right) for the case and control groups. Some differences are apparent, but it is unclear whether these are statistically significant. To determine the latter we test the hypothesis of equidistribution of the multivariate distributions of (X_1, X_2, X_3) for the case and control groups, $\mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$. From the Bayesian Information Criterion in (11) we have that $S_{01}(\boldsymbol{t}) = 6.425$, so

$$B_{01}(\boldsymbol{t}) \approx e^{6.425} = 617.1 \quad \text{and} \quad P(\mathrm{M}_0 \mid \boldsymbol{t}) \approx 0.998,$$

when $\omega_0 = 1/2$ is assumed. Hence, the data strongly support the hypothesis that the joint distribution of (X_1, X_2, X_3) is the same for the case and control groups.

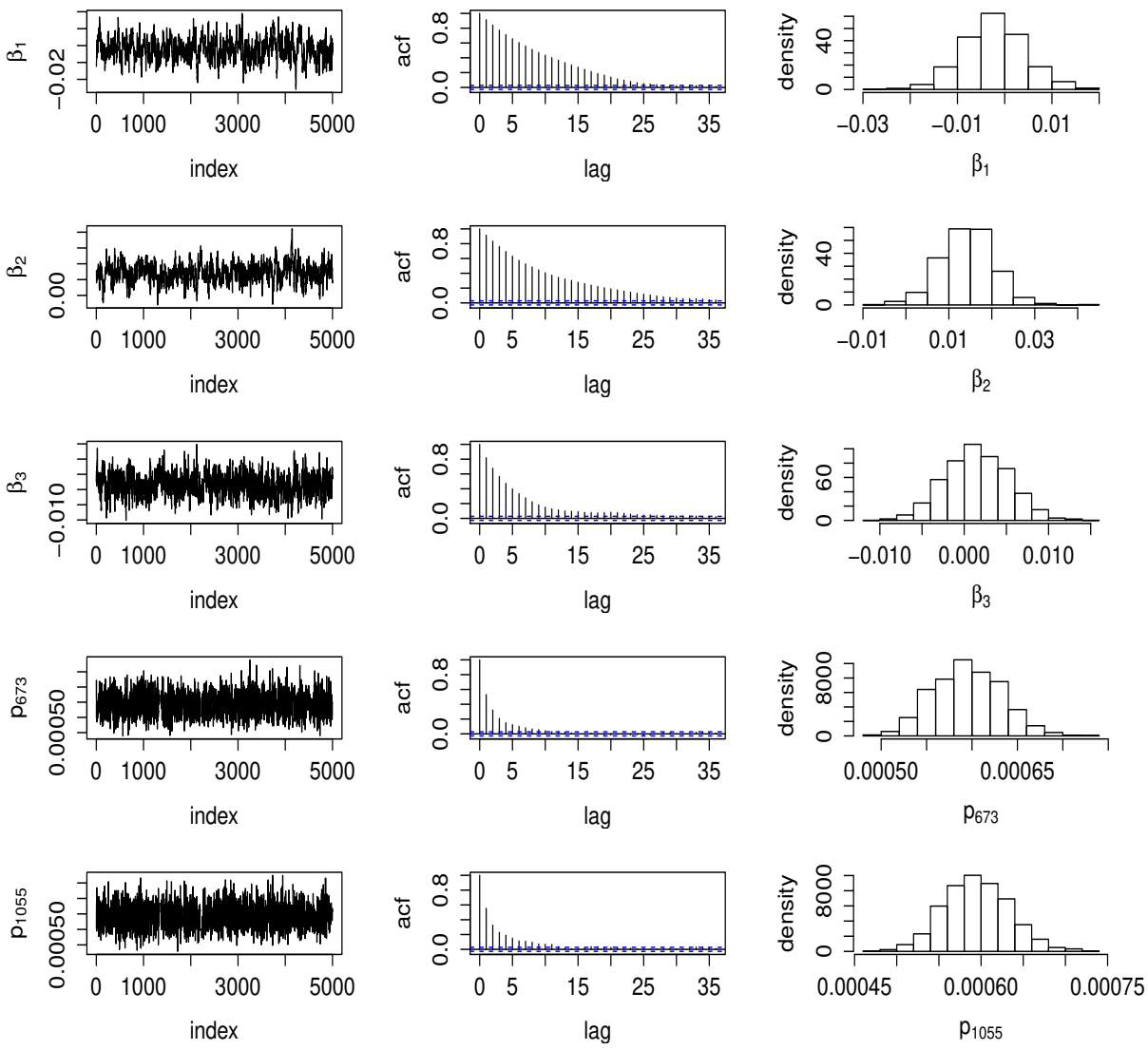

Figure 1: Summary of MCMC output of $\beta_1, \beta_2, \beta_3, p_{673}$ and p_{1055} obtained from the semiparametric density ratio model fitted to the TGCT cancer data.

Remark 2

Some of the conclusions above are at odds with those reached by Kedem et al. (2009) based on a frequentist analysis, using a bivariate DRM and the subset of the same data set that excluded the variable age. This work reported a very small p–value as the result of testing $\bar{\mathrm{H}}_0 : \beta_2 = \beta_3 = 0$ using a likelihood ratio test, which strongly suggests the rejection of $\bar{\mathrm{H}}_0$. One possible explanation is that the claimed χ^2 limiting the distribution of the likelihood ratio statistic does not hold in the multivariate setting.

The present Bayesian analysis might also portray a somewhat mixed message. On the one hand, the cdf estimates in Figure 3 suggest that the marginal distributions of the variable height for the case and control groups may be different; this might also be the case for the

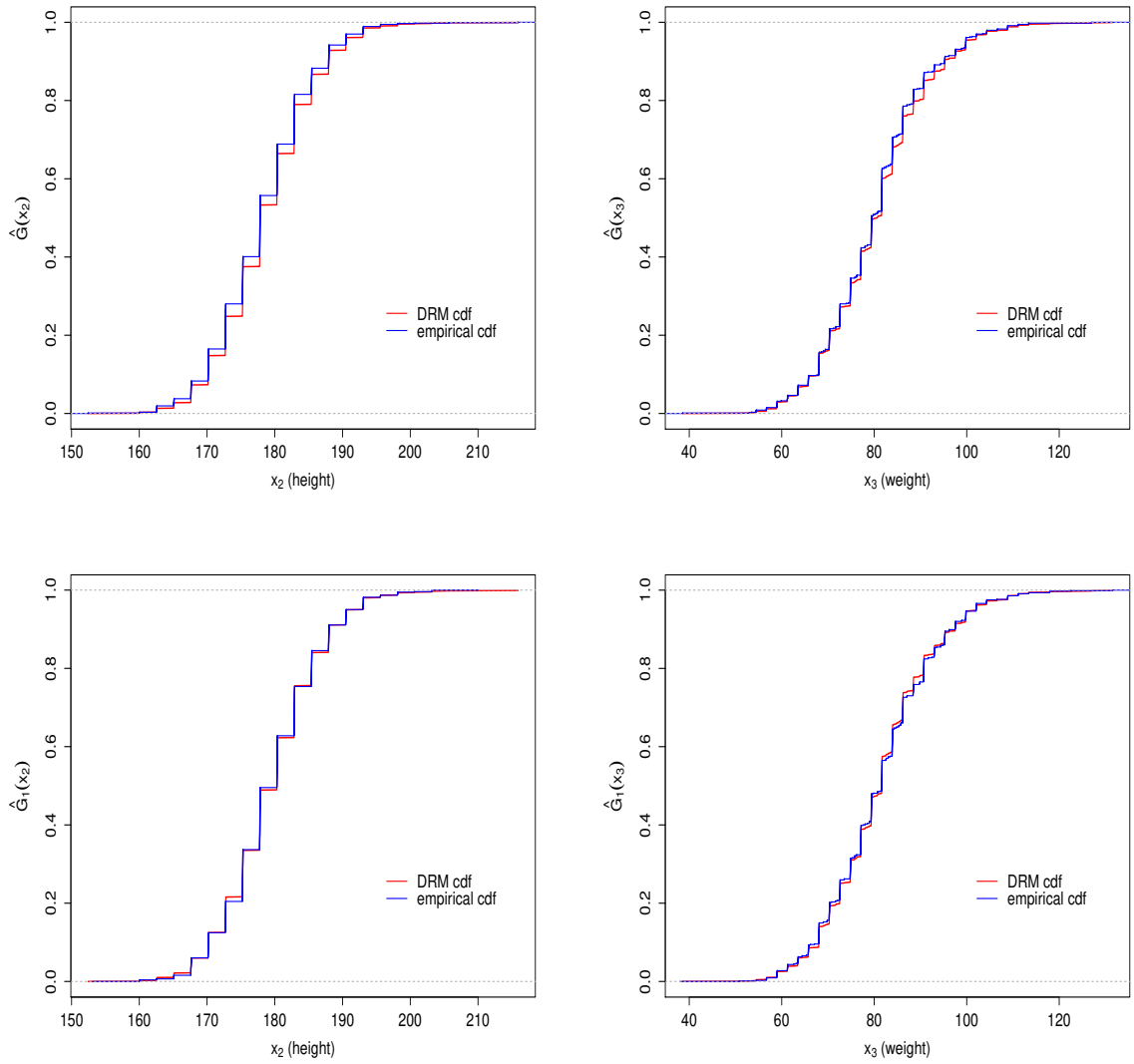

Figure 2: Top: DRM and empirical estimates of the marginal cdfs of the variables height (left) and weight (right) for the control group. Bottom: Same as the top, but for the case group.

variable weight, but to a lesser extent. On the other hand, the test above strongly supports the hypothesis that the joint distribution of the variables age, height and weight are the same for the case and control groups.

Distribution of Derived Variables

Under the DRM the distributions of other variables of interest that are functions of the variables under study can also be estimated. One such variable is the body mass index defined as

$$Z = T(X_1, X_2, X_3) = X_3/(X_2/100)^2.$$

Under the assumption that the joint distribution of (X_1, X_2, X_3) is supported at the finite set $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n\}$, it follows that the support of the distribution of Z is $T(\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n\})$.

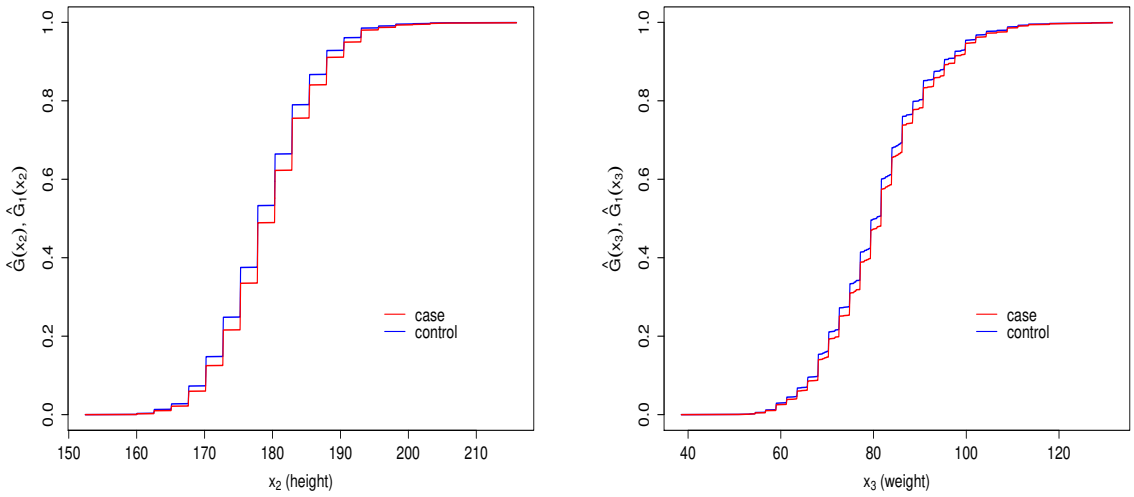

Figure 3: Left: Estimated cdfs of height for the case and control groups. Right: Estimated cdfs of weight for the case and control groups.

Recall that the set $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n\}$ is assumed to contain n distinct elements. For the TGCT data it holds that the set $T(\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n\})$ also contains n distinct elements, so this set is the support of the distribution of Z and the map $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n\} \longrightarrow T(\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n\})$ is one–to–one. Let $T(\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n\}) = \{z_1, \ldots, z_n\}$ and $s(k)$ be the unique index in $\{1, \ldots, n\}$ for which $T(\boldsymbol{t}_{s(k)}) = z_k$, $k = 1, \ldots, n$. Then for the control group we have that $P_G(Z = z_k) = P_G(\mathbf{X} = T^{-1}(\{z_k\})) = p_{s(k)}$, and hence under the DRM the cdf of the body mass index for the control group is given by

$$B(z) = P_G(Z \leq z) = \sum_{k=1}^{n} p_{s(k)} \mathbf{1}\{z_k \leq z\}.$$

By the same token, the cdf of the body mass index for the case group is

$$B_1(z) = \sum_{k=1}^{n} \left(\frac{p_{s(k)} \exp(\boldsymbol{\beta}_1' \boldsymbol{h}(\boldsymbol{t}_{s(k)}))}{\sum_{l=1}^{n} p_l \exp(\boldsymbol{\beta}_1' \boldsymbol{h}(\boldsymbol{t}_l))} \right) \mathbf{1}\{z_k \leq z\}, \qquad z \in \mathbb{R}.$$

Bayesian estimates for these cdfs are readily available from the MCMC sample in the same way of those for $G(\boldsymbol{x})$ and $G_1(\boldsymbol{x})$ once $\{s(1), \ldots, s(n)\}$, a permutation of $\{1, \ldots, n\}$, is obtained; see Section 4.1. Figure 4 displays the estimates of $B(z)$ and $B_1(z)$. These show that the two cdfs are indistinguishable for practical purposes, supporting the conclusion from previous analyses that the distributions of body mass index in the case and control groups are the same; the latter also follows from the acceptance of the equidistribution hypothesis H_0.

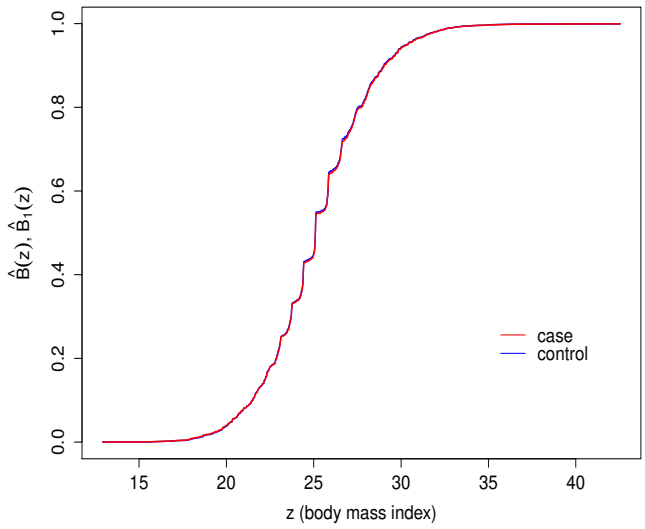

Figure 4: Estimated cdfs of body mass index for the case and control groups.

6 Conclusions and Discussion

In this work we have extended the Bayesian analysis of the univariate density ratio model developed in De Oliveira and Kedem (2017) to the multivariate model studied in Voulgaraki et al. (2012). The extension follows in the footsteps of the univariate model, but some worthwhile additions relevant to the multivariate case were described. First, once the different multivariate cdfs have been estimated, any marginal cdf (being univariate or multivariate) is readily estimated under the density ratio model. Second, the estimation of the cdf of a new variable that is a function of the variables under study is also readily available, as long as the transformation maps the support of the multivariate distribution to the support of the new variable in a one–to–one way. The methodology was illustrate with the re–analysis of the testicular germ cell tumor data analyzed in Voulgaraki et al. (2012).

A problem considered in Voulgaraki et al. (2012) that was not considered here is the estimation of regression functions. By considering one of the variables as the response, say X_p, and the rest of the variables $X_1, \ldots, X_{p-1}$ as covariates, $E_i(X_p \mid X_1, \ldots, X_{p-1})$ (the conditional expectation under the ith distribution) can be estimated from an estimate of the conditional distribution of X_p given $X_1, \ldots, X_{p-1}$. Voulgaraki et al. (2012) chose to estimate the latter using kernel density estimation, but this has several drawbacks, not the least being the fact that multivariate density estimation is a challenging problem, specially in large dimensions. A possible alternative is to express the regression functions in terms of conditional cdfs rather than conditional pdfs. We conjecture that the former should be expressible in terms of the joint cdfs that are already in place, although the procedure might become cumbersome for dimensions larger than three. We plan to investigate this conjecture.

References

Aitchison, J. and Shen, S.M. (1980), Logistic–normal distributions: Some properties and uses. *Biometrika*, 67, 261-272.

Anderson, J.A. (1979), Multivariate logistic compounds. *Biometrika*, 66, 17-26.

Chamberlain, G. and Imbens, G.W. (2003), Nonparametric applications of Bayesian inference. *Journal of Business and Economic Statistics*, 21, 12-18.

De Oliveira, V. and Kedem, B. (2017), Bayesian analysis of a density ratio model. *The Canadian Journal of Statistics*, 45, 274-289.

Fokianos, K., Kedem, B., Qin, J. and Short, D.A. (2001), A semiparametric approach to the one–way layout. *Technometrics*, 43, 56-65.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd. ed. Chapman & Hall/CRC Press, Boca Raton.

Gneiting, T. (2002), Compactly supported correlation functions. *Journal of Multivariate Analysis* 83, 493-508.

Granville, V. (1996), Discriminant analysis and density estimation on the finite d–dimensional grid. *Computational Statistics & Data Analysis*, 22, 27-51.

Kass, R.E. and Raftery, A.E. (1995), Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Kedem, B., De Oliveira, V. and Sverchkov, M. (2017), *Statistical Data Fusion.* World Scientific.

Kedem, B., Kim, E-Y, Voulgaraki, A. and Graubard, B. (2009), Two–dimensional semiparametric density ratio modeling of testicular germ cell data. *Statistics in Medicine*, 28, 2147-2159.

Kedem, B., Lu, G., Wei, R. and Williams, D. (2008). Forecasting mortality rates via density ratio modeling. *Canadian Journal of Statistics*, 36, 193-206.

Kitamura, Y. (2007), Nonparametric likelihood: Efficiency and robustness. *The Japanese Economic Review*, 58, 26-46.

McGlynn, K.A. and Cook, M.B. (2010), The epidemiology of testicular cancer. In: *Male Reproductive Cancers: Epidemiology, Pathology and Genetics*, W.D. Foulkes and K.A. Cooney (eds.), 51-83. Springer.

Owen, A.B. (2001), *Empirical Likelihood.* Chapman & Hall/CRC Press, Boca Raton.

Qin, J. and Zhang, B. (1997), A goodness of fit test for logistic regression models based on case–control data. *Biometrika*, 84, 609-618.

Rubin, D.B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9, 130-134.

Vardi, Y. (1982), Nonparametric estimation in the presence of length bias. *Annals of Statistics*, 10, 616-20.

Vardi, Y. (1985), Empirical distribution in selection bias models. *Annals of Statistics*, 13, 178-203.

Voulgaraki, A., Kedem, B. and Graubard, B.I. (2012), Semiparametric regression in testicular germ cell data. *The Annals of Applied Statistics*, 6, 1185-1208.

Wall, M. (2004), A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, 121, 311-324.