ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)

A minimum relative entropy based correlation model between the response and covariates

Bhaskar Bhattacharya

Southern Illinois University, Carbondale, USA

and Mohammad Al-talib

Yarmouk University, Irbid, Jordan

[Received May 2015. Final revision June 2016]

Summary. A semiparametric model is presented utilizing dependence between a response and several covariates. We show that this model is optimum when the marginal distributions of the response and the covariates are *known*. This model extends the generalized linear model and the proportional likelihood ratio model when the marginal distributions are *unknown*. New interpretations of known models such as the logistic regression model, density ratio model and selection bias model are obtained in terms of dependence between variables. For estimation of parameters, a simple algorithm is presented which is guaranteed to converge. It is also the same regardless of the choice of the distribution for response and covariates; hence, it can fit a very wide variety of useful models. Asymptotic properties of the estimators of model parameters are derived. Real data examples are discussed to illustrate our approach and simulation experiments are performed to compare with existing procedures.

Keywords: Dependence; Information; Logistic regression; Semiparametric generalized linear model

1. Introduction

Luo and Tsai (2012) described neuropsychological scale data where the response variable is the score from the trail making test (part A) measuring 334 patients' processing speed in seconds, and the covariates are years of education, age and diagnosis. Often these types of data have several covariates (Figs 1 and 2 in Section 7 show scatter plots of scores *versus* years of education and age), have unknown statistical distributions and known statistical procedures fail to work properly. Stamey *et al.* (1989) examined the correlations between the level of prostate-specific antigen and several clinical measures in 97 men who were about to receive a radical prostatectomy. The goal (Hastie *et al.*, 2009) is to predict the logarithm of prostate-specific antigen level, lpsa, from a number of measurements including log-cancer-volume, lcavol, log-prostate-weight, lcp, age and logarithm of capsular penetration, lcp (Figs 5–7 in Section 7 show scatter plots of lpsa *versus* lcavol, lweight and lcp respectively). Although there are moderately high correlations between the response and most covariates, a linear regression model does not consider the non-linearity at the edges of the data. In both examples, our goal is to develop a model for the response variable by utilizing its dependence on the covariates while maintaining all the

Address for correspondence: Bhaskar Bhattacharya, Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408, USA.
E-mail: bhaskar@siu.edu

1369–7412/17/79000

marginal information about the response and the covariates that is obtained from the sample in the model.

For a response Y with covariates Z_i, $1 \leqslant i \leqslant d$, the correlation coefficient $\text{corr}(Y, Z_i) = c_i$ measures any linear relationship between Y and Z_i. When $|c_i|$ is much smaller than 1, then the linear relationship between Y and Z_i is very weak. Then we may think of the relationship between Y and Z_is as non-linear in nature. In this sense, c_i not only measures the strength of the linear but also the non-linear relationship between Y and Z_i with higher or lower c_i referring respectively to more linearity or more non-linearity. Thus c_i can be considered as a global measure of dependence between Y and Z_i. For example, if $E(Y|Z_i) = Z_i^2$ with Z_i being symmetrically distributed around 0, then $c_i = 0$, then the dependence of Y on Z_i is *entirely non-linear*. Fig. 8(c) in Section 8 considers a similar situation with $E(Y|Z) = |Z|$.

In this paper, we develop a model for Y based on $\text{corr}(Y, Z_i) = c_i$, $1 \leqslant i \leqslant d$. We shall see that, when the specified c_is are high, the model that is obtained is almost linear (e.g. prostate data), whereas, when the specified c_is are low, the model that is obtained is non-linear (e.g. the trail making data). In any scientific study, the initial choice of covariates from a vast pool may be difficult. Given a set of covariates, transformations of covariates or increasing the number of correlation constraints with original or transformed variables might prove useful for a better fit of the model developed. See Sections 7 and 8 for discussions on transformations of covariates.

When the marginal distributions of Y and the Z_is are known, our procedure has close connections with the *maximum entropy* (ME) principle, which may be stated as follows:

> 'when selecting a model for a given situation it is often appropriate to express the prior information in terms of constraints. However, one must be careful so that no information other than these specified constraints is used in model selection. That is, other than the constraints that we have, the uncertainty associated with the probability distribution to be selected should be kept at its maximum' (Jaynes, 1957).

The ME principle can be generalized to the concepts of Kullback–Leibler (KL) distance and I-projection, as defined below (Csiszár, 1975). For two probability measures Q and P, the *KL distance* (or, *relative entropy*) between Q and P is defined as

$$I(Q|P) = \begin{cases} \displaystyle\int \ln\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right) \mathrm{d}Q, & \text{if } Q \ll P, \\ \infty, & \text{otherwise.} \end{cases} \tag{1.1}$$

($Q \ll P$ means that Q is absolutely continuous with respect to P.) Although $I(Q|P)$ is not a metric, it is always non-negative and equals 0 if and only if $Q = P$. Hence it is often interpreted as a measure of 'divergence' or 'distance' between Q and P. For a given P and a specified set of probability measures $\mathbb{C}$, it is often of interest to find the $Q^* \in \mathbb{C}$ which satisfies

$$I(Q^*|P) = \inf_{Q \in C} I(Q|P) \qquad (<\infty). \tag{1.2}$$

Such a Q^* is called the *I-projection* of P onto $\mathbb{C}$. Csiszár (1975) has shown that Q^* exists uniquely if $\mathbb{C}$ is convex and variation closed and there is a $Q \in \mathbb{C}$ such that $I(Q|P) < \infty$ (when P is uniform, equation (1.2) becomes the ME principle).

Associations between random variables have been of interest to statisticians and probabilists over several recent decades (Agresti, 2013). Recently, Reshef *et al.* (2011) have proposed a measure of dependence for two-variable relationships, known as the maximal information coefficient. Although the maximal information coefficient captures a wide range of associations both functional and not, it is restricted to only two variables at a time. Szekely *et al.* (2007,

2009) have proposed distance correlation measures of dependence between variables. But this paper proposes models utilizing dependence based on correlations between variables and their marginal distributions.

In Section 2 first we show that the model proposed is *optimum* when the marginal distributions of Y and $\mathbf{Z}$ are *known*. Then we propose a model for Y when the marginal distributions of Y and the Z_is are *unknown*, and we discuss its properties. In Section 3, we describe the relationship of the model with other existing models. Also, new interpretations of known models are obtained. In Section 4, we present an algorithm to estimate the parameters of the model proposed. This algorithm is *guaranteed* to converge and is applicable to *any* distributions of Y and Z_is. Consistency and asymptotic normality of the parameters are addressed in Section 5. In Section 6, we present results from two simulation experiments, for discrete and continuous response cases. In Section 7, two real data examples are considered: one from neuropsychological test scores, and the other involving prostate cancer trials. Discussion and final comments are in Section 8. Further details on duality, proofs of theorems and residual plots are in the supplemental file that is available on line.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

```
http://wileyonlinelibrary.com/journal/rss-datasets
```

2. Model proposed and its properties

Let the response variable Y have cumulative distribution function (CDF) $P_1(y)$, and marginal probability density functions (PDFs) $p_1(y)$. Those for a d-dimensional covariate vector $\mathbf{Z} = (Z_1, \ldots, Z_d)$ are $P_2(\mathbf{z})$ and $p_2(\mathbf{z})$ respectively. First we assume that the marginal distributions of Y and $\mathbf{Z}$ are *known*. As we are interested in the dependence between Y and $\mathbf{Z}$, let $(Y, \mathbf{Z})$ have a (unknown) joint CDF $Q(y, \mathbf{z})$. Let the marginal CDFs of Y and $\mathbf{Z}$ that are derived from $Q(y, \mathbf{z})$ be $Q_1(y)$ and $Q_2(\mathbf{z})$ (PDFs $q_1(y)$ and $q_2(\mathbf{z})$) respectively. Now consider a class $\mathbb{C}$ of all joint probability distributions $Q(y, \mathbf{z})$ of $(Y, \mathbf{Z})$ where

$$\mathbb{C} = \{Q : \mathrm{corr}(Y, Z_i) = c_i, 1 \leqslant i \leqslant d, Q_1(y) = P_1(y), \forall\, y \in \mathbb{R}, Q_2(\mathbf{z}) = P_2(\mathbf{z}), \forall\, \mathbf{z} \in \mathbb{R}^d\}, \tag{2.1}$$

for a *given* vector of constants $\mathbf{c} = (c_1, \ldots, c_d)$, $-1 \leqslant c_i \leqslant 1$. Using the marginal distributions of Y and $\mathbf{Z}$, we can equivalently express

$$\mathbb{C} = \{Q : E(Y\mathbf{Z}) = \mathbf{c}', Q_1(y) = P_1(y), \forall\, y \in \mathbb{R}, Q_2(\mathbf{z}) = P_2(\mathbf{z}), \forall\, \mathbf{z} \in \mathbb{R}^d\}, \tag{2.2}$$

where $\mathbf{c}' = (c_1', \ldots, c_d')$, $c_i' = c_i \surd\{\mathrm{var}(Y)\,\mathrm{var}(Z_i)\} + E(Y)\,E(Z_i)$, $\forall\, i$. Without loss of generality, we shall use $c_i = c_i'$.

The independence model $P = P_1 P_2$ ignores any dependence relationship between Y and $\mathbf{Z}$. We propose a model Q^* for $(Y, \mathbf{Z})$ to be that Q in $\mathbb{C}$ that is *closest* to P in the minimum KL divergence (recall the ME principle). Theorem 1 below shows that the criteria in $\mathbb{C}$ lead us to the solution given by (where T denotes transpose)

$$q^*(y, \mathbf{z}) = \frac{\exp\{\boldsymbol{\beta}^{*\mathrm{T}} \mathbf{z} y + h_1^*(y) + h_2^*(\mathbf{z})\}\, p_1(y)\, p_2(\mathbf{z})}{\displaystyle\int \exp\{\boldsymbol{\beta}^{*\mathrm{T}} \mathbf{v} u + h_1^*(u) + h_2^*(\mathbf{v})\}\, p_1(u)\, p_2(\mathbf{v}) \mathrm{d}u\, \mathrm{d}\mathbf{v}}. \tag{2.3}$$

Here the (adjustment) functions $h_1^*(y)$ and $h_2^*(\mathbf{z})$ are needed so that the solution maintains the Y- and $\mathbf{Z}$-marginals as specified in $\mathbb{C}$. Sometimes the exact expressions of $h_1^*(y)$ and $h_2^*(\mathbf{z})$ are available, as in example 1 below. In other cases they are approximated by using the algorithm given. Deriving the conditional distribution from the joint distribution (2.3), a model of Y, for given $\mathbf{Z}$, is

$$q^*(y|\mathbf{z}) = \frac{\exp\{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{z}y + h_1^*(y)\}\, p_1(y)}{\int \exp\{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{z}u + h_1(u)\}\, p_1(u)\,\mathrm{d}u}. \tag{2.4}$$

Since theorem 1 characterizes equation (2.3) as an optimum solution under $\mathbb{C}$, we shall say that equation (2.4) is the *optimum* model for Y using $\mathbf{Z}$ (meaning, derived from equation (2.3)). The following theorem proves the optimality of equation (2.4) by using the duality technique from Bhattacharya and Dykstra (1995, 1997). More details on duality are in the supplemental file that is available on line.

Theorem 1. Assume that $\mathbf{c}$, $p_1(y)$ and $p_2(\mathbf{z})$ in $\mathbb{C}$ are *known*. Then model (2.4) is the optimum model of Y for given $\mathbf{Z}$ that incorporates the correlation and marginal information in $\mathbb{C}$, in the sense that the corresponding joint distribution Q^* (in $\mathbb{C}$) is the closest to the independence model P in the minimum KL distance than any other Q in $\mathbb{C}$.

The following two examples use theorem 1 to prove a characterizing property of the multivariate normal distribution in the sense that the closest distribution from the product of (multivariate) normal distribution marginals onto the set $\mathcal{C}$ is also multivariate normal. Example 1 shows exact expressions for $h_1^*(y)$ and $h_2^*(\mathbf{z})$. Comparing these two examples, we can see the importance of the presence of marginal constraints.

2.1. Example 1

For simplicity, we consider the case $d = 2$, and one can extend it to any d. Consider the random vector (Y, Z_1, Z_2) such that the marginals are $Y \sim P_1 = N(0,1)$, and $(Z_1, Z_2) \sim P_2 = N_2\{\mathbf{0}, (1, \rho//\rho, 1)\}$, $-1 \leqslant \rho \leqslant 1$, for a *known* value of ρ. Consider the class $\mathbb{C}_1$ of probability distributions Q_1 where

$$\mathbb{C}_1 = \{Q_1 : E_{Q_1}(YZ_1) = \rho_1, E_{Q_1}(YZ_2) = \rho_2, Y \sim N(0,1), (Z_1, Z_2) \sim N\{\mathbf{0}, (1, \rho//\rho, 1)\}\},$$

and $-1 \leqslant \rho_1, \rho_2 \leqslant 1$ are *known*. Assume that the values of ρ, ρ_1 and ρ_2 are such that $\boldsymbol{\Sigma}_1$ derived below is positive definite (as the numerical example shows). We look for Q_1^* in $\mathbb{C}_1$ which is closest to $P = P_1P_2$ in the minimum KL divergence (P_1 and P_2 independent). From equation (2.3) and theorem 1, the solution is Q_1^*, where $q_1^*(y, \mathbf{z}) = \mathrm{d}Q_1^*/\mathrm{d}\mu$ (μ is Lebesgue measure), is given by ($\mathbf{z} = (z_1, z_2)$ and $\mathbf{v} = (v_1, v_2)$)

$$q_1^*(y, \mathbf{z}) = \frac{\exp\{\beta_{11}^* yz_1 + \beta_{12}^* yz_2 + h_1^*(y) + h_2^*(\mathbf{z})\}\, p_1(y)\, p_2(\mathbf{z})}{\int \exp\{\beta_{11}^* uv_1 + \beta_{12}^* uv_2 + h_1^*(u) + h_2^*(\mathbf{v})\}\, p_1(u)\, p_2(\mathbf{v})\,\mathrm{d}u\,\mathrm{d}\mathbf{v}}. \tag{2.5}$$

Consider the case when $(Y, Z_1, Z_2) \sim N_3(\mathbf{0}, \boldsymbol{\Sigma}_1)$, where $\boldsymbol{\Sigma}_1 = (\sigma_{ij})$, with $\sigma_{ii} = 1$, $\sigma_{12} = \rho_1$, $\sigma_{13} = \rho_2$, $\sigma_{23} = \rho$ and $\sigma_{ij} = \sigma_{ji}, \forall\ i, j$. Then from the expression for $\boldsymbol{\Sigma}_1^{-1} = (\sigma^{ij})$, the joint PDF of (Y, Z_1, Z_2) is

$$\begin{aligned} q(y, z_1, z_2) = {} & \frac{\exp\{-1/(2k_1)\}}{(2\pi)^{3/2} k_1^{1/2}} \{(1-\rho^2)y^2 + (1-\rho_2^2)z_1^2 + (1-\rho_1^2)z_2^2 - 2(\rho\rho_2 - \rho_1)yz_1 \\ & -2(\rho\rho_1 - \rho_2)yz_2 - 2(\rho_1\rho_2 - \rho)z_1z_2\}, \end{aligned} \tag{2.6}$$

where $k_1 = \det(\boldsymbol{\Sigma}_1) = 1 - \rho^2 - \rho_1^2 - \rho_2^2 + 2\rho\rho_1\rho_2$.

The distribution $N_3(\mathbf{0}, \mathbf{\Sigma}_1)$ is in $\mathbb{C}_1$. Using the normal PDF expressions for $p_1(y)$ and $p_2(\mathbf{z})$ in equation (2.5) and comparing the resulting exponent with that in equation (2.6), we can see using algebra that these two exponents would be the same (except for the additive constants) if we set

$$\begin{aligned} h_1^*(y) &= \frac{-\rho_1^2-\rho_2^2+2\rho\rho_1\rho_2}{2k_1}(y^2-1), \\ h_2^*(z_1,z_2) &= \frac{-\rho^2-\rho_1^2+2\rho\rho_1\rho_2}{2k_1(1-\rho^2)}(z_1^2-1)+\frac{-\rho^2-\rho_2^2+2\rho\rho_1\rho_2}{2k_1(1-\rho^2)}(z_2^2-1) \\ &\quad +\frac{-2\rho+2\rho^3+\rho_1\rho_2+\rho\rho_1^2+\rho\rho_2^2-3\rho^2\rho_1\rho_2}{k_1(1-\rho^2)}(z_1z_2-\rho) \end{aligned}$$

and $\beta_{11}^*=(\rho\rho_2-\rho_1)/k_1$ and $\beta_{12}^*=(\rho\rho_1-\rho_2)/k_1$. Also, $\int h_1^*(y)\,\mathrm{d}P_1(y) = \int h_2^*(\mathbf{z})\,\mathrm{d}P_2(\mathbf{z}) = 0$ is satisfied (see the proof of theorem 1). Since the solution of equation (1.2) is unique, $Q_1^*=N_3(\mathbf{0},\mathbf{\Sigma}_1)[\in\mathbb{C}]$ must be the solution to our problem. As a numerical example, let $\rho=0.7, \rho_1=0.3$ and $\rho_2=0.5$; then q_1^* has $\beta_{11}^*=-0.132$ and $\beta_{12}^*=0.763$.

2.2. Example 2

Consider the same setting as in example 1. Removing the marginal information from $\mathbb{C}_1$, consider the class $\mathbb{C}_2$ of probability distributions Q_2 where

$$\mathbb{C}_2=\{Q_2: E_{Q_2}(YZ_1)=\rho_1, E_{Q_2}(YZ_2)=\rho_2\},$$

where $-1\leqslant\rho_1,\rho_2\leqslant 1$ are *known*. We look for Q_2^* in $\mathbb{C}_2$ which is closest to $P=P_1P_2$ (with the same P as in example 1) in the minimum KL divergence. Using a similar technique to those in theorem 1 and its proof, the solution is Q_2^*, where $q_2^*(y,\mathbf{z})=\mathrm{d}Q_2^*/\mathrm{d}\mu$ (μ is Lebesgue measure) is given by

$$\begin{aligned} q_2^*(y,\mathbf{z}) &= \frac{\exp(\beta_{21}^*yz_1+\beta_{22}^*yz_2)\,p_1(y)\,p_2(z_1,z_2)}{\int\exp(\beta_{21}^*uv_1+\beta_{22}^*uv_2)\,p_1(u)\,p_2(v_1,v_2)\,\mathrm{d}u\,\mathrm{d}\mathbf{v}} \\ &= \frac{\exp(\beta_{21}^*yz_1+\beta_{22}^*yz_2-y^2/2-[1/\{2(1-\rho^2)\}](z_1^2-2\rho z_1z_2+z_2^2))}{\int\exp(\beta_{21}^*uv_1+\beta_{22}^*uv_2-u^2/2-[1/\{2(1-\rho^2)\}](v_1^2-2\rho v_1v_2+v_2^2))\mathrm{d}u\,\mathrm{d}\mathbf{v}}. \qquad (2.7) \end{aligned}$$

Write the exponent of the numerator of the rightmost term of equation (2.7) as $-\frac{1}{2}(y,z_1,z_2)\mathbf{\Sigma}_2^{-1}(y,z_1,z_2)^{\mathrm{T}}$, where

$$\begin{aligned} \mathbf{\Sigma}_2^{-1} &= \begin{pmatrix} 1 & -\beta_{21}^* & -\beta_{22}^* \\ -\beta_{21}^* & 1/(1-\rho^2) & -\rho/(1-\rho^2) \\ -\beta_{22}^* & -\rho/(1-\rho^2) & 1(1-\rho^2) \end{pmatrix}, \\ \mathbf{\Sigma}_2 &= \frac{1}{k_2}\begin{pmatrix} 1 & \beta_{21}^*+\beta_{22}^*\rho & \beta_{21}^*\rho+\beta_{22}^* \\ \beta_{21}^*+\beta_{22}^*\rho & 1-(1-\rho^2)\beta_{22}^* & \rho+(1-\rho^2)\beta_{21}^*\beta_{22}^* \\ \beta_{21}^*\rho+\beta_{22}^* & \rho+(1-\rho^2)\beta_{21}^*\beta_{22}^* & 1-(1-\rho^2)\beta_{21}^* \end{pmatrix} \end{aligned}$$

and $k_2=1-\beta_{21}^*-\beta_{22}^*-2\beta_{21}^*\beta_{22}^*\rho$. Setting the $(1,2)$th and $(1,3)$th elements of $\mathbf{\Sigma}_2$ equal to ρ_1 and ρ_2 respectively, we obtain $\beta_{21}^*=k_2(\rho_1-\rho\rho_2)/(1-\rho^2)$ and $\beta_{22}^*=k_2(\rho_2-\rho\rho_1)/(1-\rho^2)$.

Using these values of β^*_{21} and β^*_{22} in the expression for k_2 above, we obtain the quadratic $k_2^2(\rho_1^2+\rho_2^2-2\rho\rho_1\rho_2)+k_2-1=0$, which is easily solved for k_2.

Since $Q_2^*=N(\mathbf{0},\mathbf{\Sigma}_2)\in\mathbb{C}_2$, and the solution of equation (1.2) is unique, Q_2^* must be the solution to our problem. Note from $\mathbf{\Sigma}_2$ above that the marginal distributions of Y and $\mathbf{Z}$ are now *different* from those of the solution in example 1 ($\mathbf{\Sigma}_1$). As a numerical example, using the same ρ, ρ_1 and ρ_2 as in example 1, we obtain that q_2^* has $\beta^*_{21}=-0.076$, $\beta^*_{22}=0.442$ and $k_2=0.777$ (discarding the negative value).

Comparing these two examples, $\mathbb{C}_1\subset\mathbb{C}_2$ and $(\beta^*_{11},\beta^*_{12})\neq(\beta^*_{21},\beta^*_{22})$. Thus, fixing the marginals in equation (2.2) makes a difference in the solution of equation (1.2). It may seem difficult to motivate equation (2.7) with regard to $P=P_1P_2$ when the marginal distributions are not specified in $\mathbb{C}_2$; however, as shown in Section 3.2, this is the scenario that produces the proportional likelihood ratio model.

Analytical solutions for $h_1^*(y)$ and $h_2^*(\mathbf{z})$ can be obtained by using a general algorithm given in Bhattacharya (2006). For a given sample, the variables are discrete, and we provide an algorithm to find $h_1^*(y)$ and $h_2^*(\mathbf{z})$ in this paper. In constructing a model, we would benefit from utilizing as much dependence between Y and $\mathbf{Z}$ as possible. So, if $\mathrm{corr}(Y,\mathbf{Z})$ is low, then we could employ some functions, e.g. $g_1(Y)$ and $g_2(\mathbf{Z})$, provided that $\mathrm{corr}\{g_1(Y),g_2(\mathbf{Z})\}$ is higher. Similar approaches have been used by Fokianos and Kaimi (2006) and Guerrero and Johnson (1982). We discuss this further in Sections 3, 7 and 8.

For an unknown data-generating process, however, the true values of $\mathbf{c}$, $p_1(y)$ and $p_2(\mathbf{z})$ will *not* be known. Then we can neither define $\mathbb{C}$ as in expression (2.1); nor does the optimality of equation (2.4) hold as above. However, for large sample size, the marginal distributions of Y and $\mathbf{Z}$ can be estimated by using their empirical distributions along with the population correlations between Y and Z_i by the corresponding sample values. Using these observed values of constraints from the sample, we define a discrete version of $\mathbb{C}$ (namely, $\mathbb{K}$ in expression (4.2)) in the same way as in equation (2.1) in Section 4. Assuming that these sample (moment) constraints represent the corresponding population (moment) constraints efficiently and consistently, the resulting model (though not necessarily optimum) is expected to perform well. Model (2.4) now becomes *semiparametric* because, along with the parametric component $\boldsymbol{\beta}$, there is also the non-parametric component $p_1(y)$, where no distributional assumption about $p_1(y)$ is made.

We would like to begin with as much information as possible about the response and covariates that is available but with independence between them, i.e. first find $\hat{P}=\hat{P}_1\hat{P}_2$. This choice of $\hat{P}$ also gives us all known models as shown in Section 3. If $\hat{P}$ were to be a uniform distribution (ME formulation), that would mean throwing away some of the observed information about the response and covariates.

Next we discuss some properties of equation (2.4) with $p_1(y)$ *unknown*. The likelihood ratio is

$$\log\left\{\frac{q^*(y_2|\mathbf{z})}{q^*(y_1|\mathbf{z})}\right\}=\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}(y_2-y_1)+h_1(y_2)-h_1(y_1)+\log\left\{\frac{p_1(y_2)}{p_1(y_1)}\right\},\qquad(2.8)$$

for any $\mathbf{z}$, y_1 and y_2 such that $p_1(y_1)>0$ and $p_1(y_2)>0$. If $\mathbf{z}$ is one dimensional, then

$$\beta=\log\left\{\frac{q(y+1|z+1)/q(y|z+1)}{q(y+1|z)/q(y|z)}\right\},$$

i.e. $\exp(\beta)$ measures the likelihood ratio that the response increases by 1 unit when the covariate increases by 1 unit. Hence equation (2.4) can be called a (generalized) proportional likelihood ratio model as the likelihood ratio (2.8) depends on $\mathbf{z}$ through $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}$ (we call it ‘generalized’ because of the extra term $h_1(y_2)-h_1(y_1)$, compared with Luo and Tsai (2012)).

Let $\theta=\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}$. If $\mu(\theta)=E(Y|\mathbf{Z}=\mathbf{z})$, then $\mathrm{var}(Y|\mathbf{Z}=\mathbf{z})=\mu'(\theta)$, which is a property of the sufficient statistics in exponential families (which model (2.4) is not when $p_1(y)$ is unknown). Also, it can be easily seen that $\boldsymbol{\beta}$ is invariant to monotone increasing transformations on Y and location shifts of $\mathbf{Z}$.

3. Relationship to other models

3.1. Logistic regression model

We show below that model (2.3) reduces to the logistic regression model when Y is binary. To see this, let Y take values $y_1=0$ or $y_2=1$ and $\mathbf{Z}$ be fixed at a value $\mathbf{z}_{\mathbf{j}}$, where $\mathbf{j}=(j_1,\ldots,j_d)$. Also, let $\{r_{i\mathbf{j}}\}$ be given constants, where $\Sigma_{i,\mathbf{j}}\, r_{i\mathbf{j}}=1$. Model (2.4) is given by

$$P(Y=y_i|\mathbf{Z}=\mathbf{z}_{\mathbf{j}})=\frac{r_{i\mathbf{j}}\exp(\boldsymbol{\beta}^{\mathrm{T}}y_i\mathbf{z}_{\mathbf{j}}+\beta_{k+1,i})}{\sum\limits_{w=1}^{2} r_{w\mathbf{j}}\exp(\boldsymbol{\beta}^{\mathrm{T}}y_w\mathbf{z}_{\mathbf{j}}+\beta_{k+1,w})}$$

so defining $\pi(\mathbf{z}_{\mathbf{j}})=P(Y=1|\mathbf{Z}=\mathbf{z}_{\mathbf{j}})$ we obtain

$$\pi(\mathbf{z}_{\mathbf{j}})=\frac{r_{2\mathbf{j}}\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_{\mathbf{j}}+\beta_{k+1,2})}{r_{1\mathbf{j}}\exp(\beta_{k+1,1})+r_{2\mathbf{j}}\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_{\mathbf{j}}+\beta_{k+1,2})}.$$

Then,

$$\frac{\pi(\mathbf{z}_{\mathbf{j}})}{1-\pi(\mathbf{z}_{\mathbf{j}})}=\frac{r_{2\mathbf{j}}\exp(\beta_{k+1,2}-\beta_{k+1,1})}{r_{1\mathbf{j}}}\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_{\mathbf{j}})$$

so

$$\mathrm{logit}\{\pi(\mathbf{z}_{\mathbf{j}})\}=\beta_0+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_{\mathbf{j}},$$

which is the logistic regression model, where

$$\beta_0=\ln\left(\frac{r_{2\mathbf{j}}}{r_{1\mathbf{j}}}\right)+\beta_{k+1,2}-\beta_{k+1,1}=\ln\left\{\frac{P(Y=1)}{P(Y=0)}\right\}+\beta_{k+1,2}-\beta_{k+1,1}$$

depends on the (unknown) marginal distribution of Y. By theorem 1, if c_is and the marginal distributions of Y and $\mathbf{Z}$ as in expression (2.2) were specified, then the logistic regression model (is the conditional distribution of Y for given $\mathbf{Z}=\mathbf{z}$ of a joint distribution in $\mathbb{C}$, which) is the closest to the independence model subject to the criteria in $\mathbb{C}$.

3.2. Proportional likelihood ratio model

The semiparametric proportional likelihood ratio model that was proposed by Luo and Tsai (2012) (see also Chan (2013)) is given by

$$q^*_{\mathrm{PLR}}(y|\mathbf{z})=\frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}y)p_1(y)}{\int \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}u)p_1(u)\mathrm{d}u},\tag{3.1}$$

where the Y-marginal $p_1(y)$ is not specified. Note that equation (3.1) does not carry the term $h_1(y)$ in the exponent like in equation (2.4). Hence the marginal PDF of Y under model (3.1) is

$$\frac{\int \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}y)\,p_1(y)p_2(\mathbf{z})\mathrm{d}\mathbf{z}}{\int \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}u)p_1(u)\mathrm{du}},$$

which is not necessarily equal to $p_1(y)$. The estimation of parameters of model (3.1) depends on an algorithm; however, the right condition to guarantee the convergence of the algorithm is not tractable. Also, the covariates Z_i need to be bounded.

In fact, following Section 2 and as in example 2, the I-projection of P onto $\mathbb{C}'' = \{Q : \mathrm{corr}(Y, Z_i) = c_i, 1 \leqslant i \leqslant d\}$ is Q'^*, where

$$\frac{\mathrm{d}Q'^*}{\mathrm{d}\mu} = q'^*(y, \mathbf{z}) = \frac{\exp(\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{z}y)\, p_1(y)\, p_2(\mathbf{z})}{\int \exp(\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{v}u)\, p_1(u)\, p_2(\mathbf{v})\mathrm{d}u\mathrm{d}\mathbf{v}},$$

from which equation (3.1) follows. Thus if $p_1(y)$ and $p_2(\mathbf{z})$ were known, then model (3.1) is *optimum* in the KL sense subject to $\mathbb{C}''$. Leaving $p_1(y)$ and $p_2(\mathbf{z})$ unknown, model (3.1) is a special case of model (2.4) setting $h_1^*(y) = h_2^*(\mathbf{Z}) = 0$. Note that there can be many candidate distributions of Y and $\mathbf{Z}$ that can produce the same level of dependence $\mathbf{c}$. It is shown in examples 1 and 2 that absence of the marginals in the constraints may lead to a different model. It is *our point* that by setting the restrictions as $\mathbb{C}$ instead of $\mathbb{C}''$ we would capture the marginal information of Y and $\mathbf{Z}$ in the model, which might provide more accurate inference about Y by using $\mathbf{Z}$.

3.3. Generalized linear model

Consider a generalized linear model that specifies the conditional distribution of Y given $\mathbf{Z}$ as

$$f_{\mathrm{glm}}(y|\mathbf{z}) = \exp\left\{\frac{\theta y - b(\theta)}{a(\tau)} + c(y, \tau)\right\},$$

where a, b and c are given functions so that $f_{\mathrm{glm}}(y|\mathbf{z})$ is a PDF. Now let $\theta = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}$ and set $c^\dagger(y, \tau) = h_1(y) + \log\{p_1(y)\}$, $\boldsymbol{\beta}^\dagger = \boldsymbol{\beta}/a(\tau)$ and

$$b^\dagger(\theta) = a(\tau) \ln\left[\int \exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}u + h_1(u)\} p_1(u)\, \mathrm{d}u\right].$$

Then

$$f_{\mathrm{glm}}^\dagger(y|\mathbf{z}) = \exp\left\{\boldsymbol{\beta}^{\dagger\mathrm{T}}\mathbf{z}y - \frac{b^\dagger(\theta)}{a(\tau)} + c^\dagger(y, \tau)\right\}$$

is of the form (2.4). Thus, $f_{\mathrm{glm}}^\dagger(y|\mathbf{z})$ extends the generalized linear model by allowing the part $p_1(y)$ in $c^\dagger(y, \tau)$ to be unspecified. Since $p_1(y)$ is unspecified, the standard methods of statistical inference for exponential families do not apply.

3.4. Density ratio models

In the presence of case–control data, the two-sample density ratio model (Qin, 1998) is $f(\mathbf{z}|y = 1) = \exp\{\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{h}(\mathbf{z})\} f(\mathbf{z}|y = 0)$, for some functions $\mathbf{h}$. Using the Bayes formula, we can write this as

$$q^*(y|\mathbf{z}) = \frac{\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{h}(\mathbf{z})y\} p_1(y)}{\int \exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{h}(\mathbf{z})u\} p_1(u)\mathrm{d}u}.$$

Thus using theorem 1 when the distributions of Y and $\mathbf{Z}$ are *known*, we can characterize it as the *optimum* in the sense that it is the conditional distribution derived from a joint distribution in the set $\mathbb{C} = \{Q : \mathrm{corr}\{Y, \mathbf{h}(\mathbf{Z})\} = \mathbf{c}\}$ which is *closest* (KL) to the independence model P. When the distributions of Y and $\mathbf{Z}$ are *unknown*, it is a special case of model (2.4) with no marginal information preserved. Kay and Little (1987) and Guerrero and Johnson (1982) have considered transformations of the explanatory variables in logistic regression models for binary data.

Also, the m-sample density ratio model of Fokianos and Kaimi (2006) can be characterized as the *best* in the sense that it is the conditional distribution derived from a joint distribution in the set $\mathbb{C}=\{Q:\mathrm{corr}\{k_i(Y),h_j(\mathbf{Z})\}=c_{ij},\forall\, i,j\}$ which is closest (KL) to the independence model, assuming that the distributions of Y and $\mathbf{Z}$ are *known*. When they are unknown, the m-sample density ratio model is a special case of model (2.4) with no marginal information preserved.

3.5. Dependence measures

Joe (1989) and Good (1976) suggested the use of $\delta_{\mathbf{X}_1,\ldots,\mathbf{X}_m}=I(f_{\mathbf{X}_1,\ldots,\mathbf{X}_m}|\Pi_{i=1}^{m}f_{\mathbf{X}_j})$ as a measure of multivariate dependence for random vectors $\mathbf{X}_1,\ldots,\mathbf{X}_m$. Joe (1989) showed that, for multivariate normal distributions with $m=2$, $\delta_{\mathbf{X}_1,\mathbf{X}_2}$ is equivalent to $|\rho|$, where ρ is the correlation coefficient. Thus, in general, $\delta_{\mathbf{X}_1,\ldots,\mathbf{X}_m}$ can be useful as a measure of functional dependence between random vectors $\mathbf{X}_1,\ldots,\mathbf{X}_m$. In this paper we *identify* the joint distribution $f_{Y,Z_1,\ldots,Z_d}$ that is closest to the independence model $f_Y f_{Z_1,\ldots,Z_d}$ by minimizing $I(f_{Y,Z_1,\ldots,Z_d}|f_Y f_{Z_1,\ldots,Z_d})$ under dependence and marginality restrictions.

Interpretations similar to those in Sections 3.1–3.4 can also be obtained for other models such as the Gilbert–Lele–Vardi biased sampling model (Gilbert *et al.*, 1999) and the semiparametric single-index model (Ichimura, 1993). In this context, note that Rathouz and Gao (2009) and Huang and Rathouz (2012) modelled the mean directly, where the parameter $\boldsymbol{\beta}$ is essentially a contrast in the mean response.

4. Estimation

Suppose that a random sample $(Y_i,\mathbf{Z}_i)$, $i=1,\ldots,n$, is available from Q, the joint distribution of $(Y,\mathbf{Z})$. Let $y_1,\ldots,y_k$ be the distinct observed values of Y, where y_i occurs n_i times, $1\leqslant i\leqslant k$, $\Sigma_{i=1}^{k}n_i=n$. Similarly, for $\mathbf{Z}=(Z_1,\ldots,Z_d)$, let $\{(z_{s1},\ldots,z_{sk_s})\}$ be the distinct observed values of Z_s, $s=1,\ldots,d$. The index $\mathbf{j}=(j_1,\ldots,j_d)$ runs lexicographically from $(1,\ldots,1)$ to $(k_1,\ldots,k_d)$. Also, let $(z_{1j_1},\ldots,z_{dj_d})$ occur $m_{\mathbf{j}}$ times and, jointly, $(y_i,z_{1j_1},\ldots,z_{dj_d})$ occur $n_{i\mathbf{j}}$ times, $1\leqslant i\leqslant k$, $1\leqslant j_t\leqslant k_t$, $1\leqslant t\leqslant d$, $\Sigma_{\mathbf{j}}m_{\mathbf{j}}=\Sigma_{i,\mathbf{j}}n_{i\mathbf{j}}=n$.

For discrete Y and $\mathbf{Z}$, we write Q as $\mathbf{q}=(q_{i\mathbf{j}})$ and P as $\mathbf{p}=(p_{i\mathbf{j}})$. The observed CDFs are $\hat{P}=(\hat{p}_{i\mathbf{j}})$ and $\hat{Q}=(\hat{q}_{i\mathbf{j}})$. Here $q_{i\mathbf{j}}=P(Y=y_i,\mathbf{Z}=\mathbf{z}_{\mathbf{j}})=q(y_i,\mathbf{z}_{\mathbf{j}})$. Then the observed joint probability mass function is $\hat{q}_{i\mathbf{j}}=P(Y=y_i,Z_1=z_{1j_1},\ldots,Z_d=z_{dj_d})=n_{ij}/n$, which is used to calculate the observed value of $c_s=\Sigma_{i\mathbf{j}}y_i z_{s\mathbf{j}_s}n_{i\mathbf{j}}/n$, $s=1,\ldots,d$, and the marginal distributions of Y and $\mathbf{Z}$ as $\Sigma_{\mathbf{j}}q_{i\mathbf{j}}=n_i/n$, $i=1,\ldots,k$, $\Sigma_i q_{i\mathbf{j}}=m_{\mathbf{j}}/n$, $\mathbf{j}=(j_1\ldots j_d)=(1,\ldots,1),\ldots,(k_1\ldots k_d)$ respectively.

When the CDFs P and Q in equation (1.2) are replaced by the corresponding empirical CDFs $\hat{P}=(\hat{p}_{i\mathbf{j}})$ and $\hat{Q}=(\hat{q}_{i\mathbf{j}})$ respectively, the problem reduces to the discrete case and can be expressed as

$$\inf_{\hat{Q}\in\mathbb{K}} I(\hat{Q}|\hat{P})=\inf_{(\hat{q}_{i\mathbf{j}})\in\mathbb{K}}\sum_{i\mathbf{j}}\hat{q}_{i\mathbf{j}}\ln\left(\frac{\hat{q}_{i\mathbf{j}}}{\hat{p}_{i\mathbf{j}}}\right), \tag{4.1}$$

where the constraints in $\mathbb{C}$ in expression (2.2) are replaced by their discrete versions in $\mathbb{K}$ in equation (4.1), which is defined as

$$\mathbb{K}=\Big\{\mathbf{q}=(q_{i\mathbf{j}}):\sum_{i\mathbf{j}}y_i z_{sj_s}q_{i\mathbf{j}}=c_s,\ s=1,\ldots,d,\ \sum_{\mathbf{j}}q_{i\mathbf{j}}=\frac{n_i}{n},\ i=1,\ldots,k,\ \sum_{i}q_{i\mathbf{j}}=\frac{m_{\mathbf{j}}}{n},$$
$$\mathbf{j}=(j_1\ldots j_d)=(1,\ldots,1),\ldots,(k_1\ldots k_d)\Big\}. \tag{4.2}$$

For easier exposition, we number $k_1 \ldots k_d$ of $\mathbf{j}$-values using $t = 1, \ldots, k_1 \ldots k_d$ (as in expression (4.3) below), so $t = \mathbf{j}$ for some $\mathbf{j}$, by slight abuse of notation. In equation (4.1), $\hat{P} = (\hat{p}_{i\mathbf{j}})$ is the sample estimate of $P = P_1 P_2$, so $\hat{p}_{i\mathbf{j}} = (n_i/n) m_{\mathbf{j}}/n$, and the estimate to be found is $\hat{Q} = (\hat{q}_{i\mathbf{j}})$.

Although $\hat{P}$ satisfies the row and column constraints, it may not satisfy constraints $\Sigma_{i\mathbf{j}}\, y_i z_{sj_s} q_{i\mathbf{j}} = c_s$. For solving problems such as equation (4.1) for multiway contingency tables, Bhattacharya and Dykstra (1997) have shown that a Fenchel duality theorem can be used to identify a dual optimization problem, which might be easier to solve. When the constraint region is an intersection of convex cones, they proposed an algorithm which is guaranteed to converge. To adopt this strategy in our case, for a given n, consider problem (4.1) as the *primal* problem. For simplicity of notation, we suppress the dependence on n (for sets and dual solutions $h_{i\mathbf{j}}$) up to expression (4.10) below.

First we express equation (4.2) as an intersection of convex sets as follows:

$$\mathbb{K} = (\cap_{s=1}^{d} \mathbb{K}_s) \cap (\cap_{r=1}^{k} \mathbb{M}_r) \cap (\cap_{t=1}^{k_1 \ldots k_d} \mathbb{N}_t),$$

where

$$\left.\begin{aligned} \mathbb{K}_s &= \Big\{\mathbf{q} = (q_{i\mathbf{j}}) : \sum_{i\mathbf{j}} (y_i z_{sj_s} - c_s) q_{i\mathbf{j}} = 0\Big\}, && 1 \leqslant s \leqslant d, \\ \mathbb{M}_r &= \Big\{\mathbf{q} = (q_{i\mathbf{j}}) : \sum_{i\mathbf{j}} (I_{i\mathbf{j}}^r - n_r/n) q_{i\mathbf{j}} = 0\Big\}, && r = 1, \ldots, k, \\ \mathbb{N}_t &= \Big\{\mathbf{q} = (q_{i\mathbf{j}}) : \sum_{i\mathbf{j}} (J_{i\mathbf{j}}^t - m_t/n) q_{i\mathbf{j}} = 0\Big\}, && t = 1, \ldots, k_1 \ldots k_d, \end{aligned}\right\} \tag{4.3}$$

where $I_{i\mathbf{j}}^r = 1$ if $r = i$, and $I_{i\mathbf{j}}^r = 0$ otherwise, $J_{i\mathbf{j}}^t = 1$ if $t = \mathbf{j}$, and $J_{i\mathbf{j}}^t = 0$ otherwise. These will be referred to as $\mathbb{K}$-constraints, $\mathbb{M}$-constraints and $\mathbb{N}$-constraints respectively. The dual cone $\mathbb{K}^*$ of $\mathbb{K}$ is given by (Rockafellar, 1970)

$$\mathbb{K}^* = (\oplus_{s=1}^{d} \mathbb{K}_s^*) \oplus (\oplus_{r=1}^{k} \mathbb{M}_r^*) \oplus (\oplus_{t=1}^{k_1 \ldots k_d} \mathbb{N}_t^*), \tag{4.4}$$

where '$\oplus$' indicates that the direct sum of vectors $\mathbb{K}_s^*$, $\mathbb{M}_r^*$ and $\mathbb{N}_t^*$ are the dual cones of $\mathbb{K}_s$, $\mathbb{M}_r$ and $\mathbb{N}_t$ respectively, each consisting of vectors of length $kk_1 \ldots k_d$:

$$\begin{aligned} \mathbb{K}_s^* &= \{\mathbf{h}_{1s} = (h_{1si\mathbf{j}}) = (\gamma_{1s}(y_i z_{sj_s} - c_s)), \gamma_{1s} \in \mathbb{R}, \forall\, i, \mathbf{j}\}, \\ \mathbb{M}_r^* &= \{\mathbf{h}_{2r} = (h_{2ri\mathbf{j}}) = (\gamma_{2r}(I_{i\mathbf{j}}^r - n_r/n)), \gamma_{2r} \in \mathbb{R}, \forall\, i, \mathbf{j}\}, \\ \mathbb{N}_t^* &= \{\mathbf{h}_{3t} = (h_{3ti\mathbf{j}}) = (\gamma_{3t}(J_{i\mathbf{j}}^t - m_t/n)), \gamma_{3t} \in \mathbb{R}, \forall\, i, \mathbf{j}\}. \end{aligned}$$

Then $\mathbf{h} = (h_{i\mathbf{j}}) \in \mathbb{K}^*$ is a vector of length $kk_1 \ldots k_d$, where each of its elements is a sum of three components, and can be written as (here we suppress n and v as $h_{i\mathbf{j}} = h_{nvi\mathbf{j}}, h_{1si\mathbf{j}} = h_{nv1si\mathbf{j}}$, etc.)

$$h_{i\mathbf{j}} = \sum_{s=1}^{d} h_{1si\mathbf{j}} + \sum_{r=1}^{k} h_{2ri\mathbf{j}} + \sum_{t=1}^{k_1 \ldots k_d} h_{3ti\mathbf{j}}, \tag{4.5}$$

and the *dual* problem to problem (4.1) is given by

$$\inf_{\mathbf{h} \in \mathbb{K}^*} \sum_{i\mathbf{j}} \hat{p}_{i\mathbf{j}} \exp(h_{i\mathbf{j}}). \tag{4.6}$$

For fixed n, problem (4.1) is solved by iteratively finding the I-projections onto the $\mathbb{K}$-, $\mathbb{M}$- and $\mathbb{N}$-constraints. Each cycle has $d + k + k_1 \ldots k_d$ steps, which are the same as the number of constraints. First for the $\mathbb{K}$-constraints, for $1 \leqslant s \leqslant d$, if, at the vth cycle, the $(s-1)$th step (referred to as the $(v, s-1)$th step), the solution is $\hat{\mathbf{q}}(n, v, s-1) = (\hat{q}_{i\mathbf{j}}(n, v, s-1))$, then, at the (v, s)th step, we solve for $(\mathbf{q} = \hat{\mathbf{q}}(n, v, s))$ in

$$\inf_{\mathbf{q} \in \mathbb{K}_s} I\{\mathbf{q} | \hat{\mathbf{q}}(n, v, s-1)\}. \tag{4.7}$$

The dual objective function at the (v, s)th step is obtained from the convex conjugate (see the on-line supplement, theorem S1) of the function in expression (4.7). The corresponding dual problem solves

$$\inf_{h_{1si\mathbf{j}}\in\mathbb{K}_s^*} \sum_{i\mathbf{j}} \hat{q}_{i\mathbf{j}}(n, v, s-1)\exp(h_{i\mathbf{j}}) = \inf_{\gamma_{1s}\in\mathbb{R}} \sum_{i\mathbf{j}} \hat{q}_{i\mathbf{j}}(n, v, s-1)\exp(h_{i\mathbf{j}}). \tag{4.8}$$

Note that the dual problem (4.8) amounts to solving for the scalar γ_{1s} only, which is considerably *easier* than solving for $\mathbf{q}$ in expression (4.7). The solution to the right-hand side of equation (4.8), say, at $\gamma_{1s} = \gamma_{1nvs}$ is obtained by differentiating it with respect to γ_{1s} and setting it equal to 0, i.e.

$$\sum_{i\mathbf{j}} \hat{q}_{i\mathbf{j}}(n, v, s-1)(y_i z_{sj_s} - c_s)\exp(h_{i\mathbf{j}}) = 0, \tag{4.9}$$

which can be solved easily by the Newton-Raphson method. Then, using theorem S1, we obtain the solution to expression (4.7) as

$$\hat{q}_{i\mathbf{j}}(n, v, s) = \frac{\hat{q}_{i\mathbf{j}}(n, v, s-1)\exp(h_{1si\mathbf{j}})}{\sum_{a\mathbf{b}} \hat{q}_{a\mathbf{b}}(n, v, s-1)\exp(h_{a\mathbf{b}})}, \tag{4.10}$$

where $\mathbf{b} = (b_1, \ldots, b_d)$.

Next, for the $\mathbb{M}$-constraints, we must match the y-marginals of the current table to the observed values n_r/n. If, at the $(v, d+r-1)$th step, the solution is $\hat{\mathbf{q}}(n, v, d+r-1)$, then the *exact* dual solution at the $(v, d+r)$th step is $\gamma_{2nvr} = \log[n_r/\{n\hat{q}_{r-1,+}(n, v, d+r-1)\}]$, where $\hat{q}_{r-1,+}(n, v, d+r-1) = \Sigma_{\mathbf{j}}\,\hat{q}_{r-1,\mathbf{j}}(n, v, d+r-1)$. This amounts to rescaling the Y-marginals of the current table as in step 3 of the algorithm (see below).

Finally, considering the $\mathbb{N}$-constraints, we must match the $\mathbf{z}$-marginals of the current table to the observed values m_j/n; the *exact* dual solution at the $(v, d+k+t)$th step is $\gamma_{3nvt} = \log[m_{\mathbf{j}}/\{n\hat{q}_{+,t-1}(n, v, d+k+t-1)\}]$, (recall that $t = \mathbf{j}$ for some $\mathbf{j}$) where $\hat{q}_{+,t-1}(n, v, d+k+t-1) = \Sigma_i\,\hat{q}_{i,t-1}(n, v, d+k+t-1)$. This amounts to rescaling the $\mathbf{Z}$-marginals of the current table as in step 4 of the algorithm (see below).

Now we shall write $h_{i\mathbf{j}}$ as $h_{nvi\mathbf{j}}$ to denote the dual solutions when the sample size is n and the algorithm has run for v cycles. Similar updates will apply to the subscripts of β and γ. Incorporating the above steps repeatedly over v *complete* cycles, and letting $\beta_{1nvs} = \Sigma_{e=1}^{v}\gamma_{1nes}$, $\beta_{2nvr} = \Sigma_{e=1}^{v}\gamma_{2ner}$ and $\beta_{3nvt} = \Sigma_{e=1}^{v}\gamma_{3net}$, we obtain

$$\hat{q}_{i\mathbf{j}}(n, v, d+k+k_1\ldots k_d) = \frac{q_{i\mathbf{j}}\exp(h_{nvi\mathbf{j}})}{\sum_{a\mathbf{b}} q_{a\mathbf{b}}\exp(h_{nva\mathbf{b}})}, \tag{4.11}$$

where $q_{i\mathbf{j}} = \hat{q}_{i\mathbf{j}}(n, 0, 0) = (n_i/n)m_{\mathbf{j}}/n = \hat{p}_{i\mathbf{j}}$, $\mathbf{h}_{nv} = (h_{nvi\mathbf{j}})$ and

$$h_{nvi\mathbf{j}} = \sum_{s=1}^{d} \beta_{1nvs}(y_i z_{sj_s} - c_s) + \sum_{r=1}^{k} \beta_{2nvr}\left(I_{i\mathbf{j}}^{r} - \frac{n_r}{n}\right) + \sum_{t=1}^{k_1\ldots k_d} \beta_{3nvt}\left(J_{i\mathbf{j}}^{t} - \frac{m_t}{n}\right). \tag{4.12}$$

Thus we propose the following algorithm for solving problem (4.1).

4.1. Algorithm

Step 1: calculate the constraint values $c_1, \ldots, c_d$ and marginal totals n_i/n and $m_{\mathbf{j}}/n$ from the observed table of $(Y, \mathbf{Z})$ values. Initialize as $\gamma_{n00} = 0$ and $q_{i\mathbf{j}} = \hat{q}_{i\mathbf{j}}(n, 0, 0) = (n_i/n)m_{\mathbf{j}}/n = \hat{p}_{i\mathbf{j}}$. Let $v = 1$.

Step 2: begin with $u=1$. At the vth cycle, uth step, let $\gamma_{1nvu}=\gamma^*$ solve equation (4.9). Use equation (4.10) to form $\hat{q}_{i\mathbf{j}}(n,v,u)$. Do for $u=1,\dots,d$.
Step 3: replace $\hat{q}_{i\mathbf{j}}(n,v,u)$ by $\hat{q}_{i\mathbf{j}}(n,v,u)n_i/\{n\hat{q}_{i+}(n,v,u)\}$, where $\hat{q}_{i+}(n,v,u)=\Sigma_{\mathbf{j}}\,\hat{q}_{i\mathbf{j}}(n,v,u)$, $1\leqslant i\leqslant k$, $d+1\leqslant u\leqslant d+k$.
Step 4: replace $\hat{q}_{i\mathbf{j}}(n,v,u)$ by $\hat{q}_{i\mathbf{j}}(n,v,u)m_{\mathbf{j}}/\{n\hat{q}_{+\mathbf{j}}(n,v,u)\}$, where $\hat{q}_{+\mathbf{j}}(n,v,u)=\Sigma_i\,\hat{q}_{i\mathbf{j}}(n,v,u)$, $\forall\,\mathbf{j}$, $d+k+1\leqslant u\leqslant d+k+k_1\dots k_d$.
Step 5: Let $\hat{q}_{i\mathbf{j}}(n,v,d+k+k_1\dots k_d)=\hat{q}_{i\mathbf{j}}^{nv}$. Stop if $\max\{|\Sigma_{i\mathbf{j}}\,y_iz_{s\mathbf{j}_s}\,\hat{q}_{i\mathbf{j}}^{nv}-c_s|,\forall\ s,|\hat{q}_{i+}^{nv}-n_i/n|, \forall i, |\hat{q}_{+\mathbf{j}}^{nv}-m_{\mathbf{j}}/n|,\forall\,\mathbf{j}\}<\epsilon$, for some prespecified $\epsilon>0$; otherwise, replace v by $v+1$, and go to step 2.

We have used Fortran to implement the algorithm. This algorithm is for a *fixed* sample size n. Using results from Bhattacharya and Dykstra (1997), it follows that the above algorithm converges as $v\to\infty$. Convergence is in seconds for all the cases that we considered. Thus $\mathbf{h}_{nv}=(h_{nvij})$ converges to $\mathbf{h}_n=(h_{ni\mathbf{j}})$ as $v\to\infty$, where

$$\begin{aligned}h_{nij}&=\sum_{s=1}^{d}\beta_{1ns}(y_iz_{sj_s}-c_s)+\sum_{r=1}^{k}\beta_{2nr}\left(I_{i\mathbf{j}}^r-\frac{n_r}{n}\right)+\sum_{t=1}^{k_1\dots k_d}\beta_{3nt}\left(J_{ij}^t-\frac{m_t}{n}\right),\\&=\sum_{s=1}^{d}\beta_{1ns}(y_iz_{sj_s}-c_s)+\beta_{2ni}\left(1-\frac{n_r}{n}\right)+\beta_{3n\mathbf{j}}\left(1-\frac{m_t}{n}\right)\end{aligned}$$

for some β_{1ns}, β_{2nr} and β_{3nt}. Comparing with equation (4.12), it must be that $\beta_{nv1s}\to\beta_{1ns}$, $\beta_{nv2r}\to\beta_{2nr}$, $\beta_{nv3t}\to\beta_{3nt}$, $\forall\,s,r,t$ as $v\to\infty$. The asymptotic distribution properties of β_{1ns} are studied in Section 5.

5. Asymptotic properties

For sample size n, at cycle v, define the vector $\boldsymbol{\beta}_{1nv}=(\beta_{1nv1},\dots,\beta_{1nvd})^{\mathrm{T}}$. Next we show that the sequence of vectors $\{\boldsymbol{\beta}_{1nv},v\geqslant1\}$ which is generated by the algorithm converges to the true parameter $\boldsymbol{\beta}_1=\boldsymbol{\beta}$ in model (2.4) as $v\to\infty$ and $n\to\infty$.

Theorem 2. For a fixed n, the algorithm in Section 4.1 yields the sequence of solutions $\{\boldsymbol{\beta}_{1nv},v\geqslant1,n\geqslant1\}$ at the vth cycle. Then $\boldsymbol{\beta}_{1nv}$ converges to $\boldsymbol{\beta}_{1n}$ as $v\to\infty$, and $\boldsymbol{\beta}_{1n}$ converges to the true parameter $\boldsymbol{\beta}_1=\boldsymbol{\beta}$ in model (2.4), with probability 1, as $n\to\infty$.

Writing $\boldsymbol{\beta}_{1n}=\boldsymbol{\beta}_n$, in theorem 3 we consider the asymptotic normality properties of $\boldsymbol{\beta}_n$. Recall that $\mathrm{d}Q^*/\mathrm{d}P=q^*(y,\mathbf{z})$ is defined in equation (2.3).

Theorem 3. Assume that Q^* exists and $\int\exp(h)\mathrm{d}Q^*<\infty$ for $\boldsymbol{\beta}$ in an open neighbourhood of zero. Assume that $\int\|y\mathbf{z}\|^2(\mathrm{d}Q^*/\mathrm{d}P)\mathrm{d}Q^*<\infty$. Then

$$\surd n(\boldsymbol{\beta}_n-\boldsymbol{\beta})\overset{\mathrm{D}}{\to}N(\mathbf{0},\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{-1}),$$

where

$$\begin{aligned}\boldsymbol{\Sigma}^*&=\frac{\displaystyle\int(y\mathbf{z}-\mathbf{c})(y\mathbf{z}-\mathbf{c})^{\mathrm{T}}\exp(h)\mathrm{d}Q^*}{\displaystyle\int\mathrm{d}Q^*},\\\boldsymbol{\Sigma}&=\frac{\displaystyle\int(y\mathbf{z}-\mathbf{c})(y\mathbf{z}-\mathbf{c})^{\mathrm{T}}\exp(h)\mathrm{d}P}{\displaystyle\int\exp(h)\mathrm{d}P}.\end{aligned}\qquad(5.1)$$

To estimate the asymptotic variance of $\boldsymbol{\beta}_n$, we replace the integrals in expression (5.1) with the discrete sums, use midpoints of the discretized intervals as the values of y_i and z_{ij}s, find inverses of the resulting matrices and use matrix multiplication.

6. Simulation

6.1. Continuous case

Since the algorithm of Section 4.1 works for any distribution of Y and Z_is, we would like to compare its performance with some known procedure (e.g. maximum likelihood estimates (MLEs)) in cases when the correct model is known and when it is misspecified. We consider samples of sizes $n=50, 100, 150$ for the random vector $(Y, Z_1, Z_2) \sim \mathbf{N}_3(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (2, 0.5, 0.5//0.5, 1.5, 0.5//0.5, 0.5, 2)$. True β_1- and β_2-values are obtained by comparing the PDF of (Y, Z_1, Z_2) with equations (2.3) and (2.4) and computing $\boldsymbol{\Sigma}^{-1}$; here, true $(\beta_1, \beta_2) = (0.1538, 0.1026)$.

We need to categorize the observed values of Y and $\mathbf{Z}$ to construct their observed marginal distributions. These are used to form the constraints in $\mathbb{K}$. To determine the number of classes for each variable, first the optimal histogram bin width is calculated from Scott (1979), which asymptotically minimizes the integrated mean-squared error. He suggested the use of $h_n = 3.49\sigma n^{-1/3}$, where σ is the standard deviation, when the sample is taken from a normal distribution. As the standard deviations of three variables are chosen to be close $(\surd 2, \surd 1.5, \surd 2)$, we decided to use the same k for all three variables. We did some experimentation with different k for three variables, but this did not produce any significantly different results.

Using the values of y, z_1 and z_2 as the midpoints of those categories and frequency distributions, the observed marginal distributions of Y and (Z_1, Z_2) and the observed values of $E(YZ_1)$ and $E(YZ_2)$ are found. These values are used to define the constraints in $\mathbb{K}$. Multiplying the observed marginal distributions of Y and (Z_1, Z_2), a three-way contingency table is formed, which is the starting point of the algorithm. The algorithm stops when all constraints are satisfied subject to a prespecified value. The algorithm converged in all cases. All simulations are replicated 1000 times.

The results are in Table 1 using $k=8$. A more detailed Table 1 is in the on-line supplement to show the effect of the choice of k on the simulation results, where we varied k for $n=50, 100$. For instance, for $k=4$, we considered intervals $(-\infty, -h_n)$, $(h_n, 0)$, $(0, h_n)$ and (h_n, ∞), and similarly for other values of even k. The results are mostly invariant for $k>4$. The results are similar if the intervals are not centred near 0, or considering an odd number of intervals (those results are not presented here).

The estimates that are produced by the algorithm are compared with the MLEs of β_1 and β_2. Comparing expression (2.3) with the expression of the normal PDF exponent $(-\frac{1}{2}\Sigma_{ij}x_i\sigma^{ij}x_j)$, we find that the MLEs of β_1 and β_2 are given by $\hat{\beta}_1 = -\hat{\sigma}^{12}$ and $\hat{\beta}_2 = -\hat{\sigma}^{13}$, where $\hat{\boldsymbol{\Sigma}}^{-1} = (\hat{\sigma}^{ij})$ is the inverse of the sample covariance matrix.

The *misspecified model* assumes that $Y|\mathbf{Z} \sim \exp(\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}) + N(0, 2)$, $\mathbf{Z} \sim \mathbf{N}_2(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{Z}})$, where $\boldsymbol{\Sigma}_{\mathbf{Z}} = (1.5, 0.5//0.5, 2)$ (the same as the covariance matrix of $\mathbf{Z}$ specified). This model is outside model (2.4) because the mean is exponential in $\boldsymbol{\beta}$. But the procedure proposed is not affected as it does not depend on $\boldsymbol{\Sigma}$. See the on-line supplement for additional results from a wrongly specified covariance matrix.

For each of β_1 and β_2, we calculate the bias, standard error values and 95% coverage probability for β_1 and β_2. Note that the MLE depends on the normal PDF, whereas our procedure does not. Simulations in Table 1 indicates that the procedure performs favourably with the MLE when the correct model is used, and better than the MLE when the model is misspecified. The performance of the procedure improves as the sample size increases.

Table 1. Comparing the procedure with MLEs under correct and misspecified normal models

Method	*n*	*bias*(β_{1n})	*se*(β_{1n})	*bias*(β_{2n})	*se*(β_{2n})	*CP*(β_{1n})†	*CP*(β_{2n})†
Procedure	50	−0.0065	0.0987	−0.0091	0.0843	945	952
	100	−0.0106	0.0644	−0.0167	0.0553	946	947
	150	−0.0099	0.0521	−0.0137	0.0456	946	939
MLE (correct)	50	0.0169	0.1116	0.0090	0.0963	939	949
	100	0.0097	0.0739	0.0031	0.0611	945	950
	150	0.0057	0.0585	0.0023	0.0503	953	948
MLE (misspecified)	50	−0.2908	0.1083	−0.1896	0.0935	951	941
	100	−0.2898	0.0703	−0.1879	0.0612	940	945
	150	−0.2890	0.0574	−0.1899	0.0504	937	941

†95% coverage probability.

6.2. *Discrete case*

Suppose that Y takes values 0 and 1, with probabilities 0.4 and 0.6 respectively, and $(Z_1, Z_2) = (0,0), (0,1), (1,0), (1,1)$ with probabilities 0.3, 0.25, 0.22 and 0.23 respectively. Assuming that $(\beta_1, \beta_2) = (1,1)$, an eight-cell multinomial probability vector is generated, where $p_{i,j_1,j_2} \propto r_{i,j_1,j_2} \times \exp(\beta_1 y_i z_{1j_1} + \beta_2 y_i z_{2j_2} + \beta_{3i} + \beta_{4j_1j_2})$, $i, j_1, j_2 = 1, 2$, where the β_{3i}- and $\beta_{4j_1j_2}$-values are such that p_{i,j_1,j_2} satisfies the Y- and (Z_1, Z_2) marginal probabilities and r_{i,j_1,j_2} are obtained by multiplying the marginal probabilities of Y and (Z_1, Z_2); finally, the true p_{i,j_1,j_2}-values are found by numerical methods to be 0.1847, 0.0927, 0.0816, 0.0410, 0.1153, 0.1573, 0.1384 and 0.1890. The constraints $E(YZ_1) = c_1$ and $E(YZ_2) = c_2$ simplify nicely to $p_{221} + p_{222} = c_1$ and $p_{212} + p_{222} = c_2$ respectively. We take random samples of size n from the eight-cell multinomial, for $n = 50, 100, 150$. From the observed table we calculate the marginal distributions of Y and (Z_1, Z_2), and also find the observed values of c_1 and c_2. In this case, exact dual solutions are available not only for the marginal constraints but also for the correlation constraints; for any $\mathbf{r} = (r_{ijk})$, when solving $\min_{\mathbf{p}: p_{221}+p_{222}=c_1} I(\mathbf{p}|\mathbf{r})$, the dual problem is solved by

$$\hat{\gamma}_1 = \ln\left\{\frac{c_1(r_{111}+r_{211}+r_{121}+r_{112}+r_{212}+r_{122})}{(1-c_1)(r_{221}+r_{222})}\right\},$$

and similarly for constraint $p_{212} + p_{222} = c_2$ the dual problem is solved by

$$\hat{\gamma}_2 = \ln\left\{\frac{c_2(r_{111}+r_{211}+r_{121}+r_{112}+r_{221}+r_{122})}{(1-c_2)(r_{212}+r_{222})}\right\}.$$

As stated in Section 3.1, in the context of this simulation, we essentially have a logistic regression model. So we compared our procedure with the logistic regression model

$$P(Y=1) = \frac{\exp(\beta_0 + \beta_1 yz_1 + \beta_2 yz_2)}{1 + \exp(\beta_0 + \beta_1 yz_1 + \beta_2 yz_2)},$$

where β_0, β_1 and β_2 are solved by using the Newton–Raphson (NR) method. As seen from Table 2, the NR numbers are slightly lower than those from the procedure for $n = 50$; however, it is the other way around for $n = 100, 150$. It is possible that this is due to the simple rescaling nature of the algorithm that is proposed here; the NR method involves many matrix operations including inversions, all of which contribute to the rounding errors, for example. The *misspecified model* is taken to be a linear model $p_{i,j_1,j_2} \propto r_{i,\,j_1,\,j_2}(\beta_1 y_i z_{1j_1} + \beta_2 y_i z_{2j_2} + \beta_{3i} + \beta_{4j_1j_2})$. Here the

Table 2. Comparing the procedure under correct and misspecified multinomial models

n	*Method*	*bias*(β_{1n})	*se*(β_{1n})	*bias*(β_{2n})	*se*(β_{2n})	*CP*(β_{1n})†	*CP*(β_{2n})†
50	Procedure	0.1120	0.7285	0.0732	0.6842	954	950
	NR (correct)	0.0820	0.6773	0.0653	0.6554	946	946
	Misspecified	3.1789	1.9909	2.4155	1.8946	696	783
100	Procedure	0.0398	0.4483	0.0075	0.4538	953	951
	NR (correct)	0.0396	0.4500	0.0219	0.4553	953	950
	Misspecified	2.9301	1.2994	2.2793	1.2098	411	551
150	Procedure	0.0354	0.3691	−0.0006	0.3463	951	955
	NR (correct)	0.0398	0.3738	0.0154	0.3473	952	957
	Misspecified	2.9103	1.0241	2.2066	0.9828	195	391

†95% coverage probability.

probabilities are linear in $\boldsymbol{\beta}$, so it is outside model (2.4). Here the procedure performs favourably with the NR method when the model is correct, and it performs better when the model is misspecified. See the on-line supplement for additional results when the model is misspecified with $\beta_0 = 0$ in the logistic regression model.

7. Real data examples

We considered two examples with different correlation levels. Most of the covariates in the first and second example have respectively relatively low and high correlations with the response. We fit and compare different models in each case. The first example has two continuous and one discrete predictors, and it shows that the power transformation of the variables is useful. The second example has four continuous predictors. In each case, the fitted non-linear model seems to be a better fit to the data than the multiple linear regression model.

7.1. Trail making data

We consider a data set from Luo and Tsai (2012), which gives the scores of 334 patients on part A of the trail making test with covariates education, age and diagnosis. The continuous response variable Y is the test score, with range 0–150 s; if a patient cannot complete the test in 150 s, a score of 150 is given. The continuous covariates are Z_1, years of education (3–21 years) and Z_2, age (53–108 years), and the discrete covariate is Z_3, diagnosis (0 or 1). From the data, we find that $\text{corr}(Y, Z_1) = -0.40$ $\text{corr}(Y, Z_2) = 0.18$ and $\text{corr}(Y, Z_3) = 0.21$.

To determine k, the number of classes, we find, for each variable Z, the quantities sd, standard deviation, and h_n, optimal bin width (Scott, 1979), and then k is calculated as $k = Z_{(334)} + \epsilon - (Z_{(1)} - \epsilon)/h_n$, where $Z_{(334)}$ and $Z_{(1)}$ are the maximum and minimum values of Z respectively and $\epsilon > 0$ is a very small number; of course, k is rounded to the nearest integer.

Using these classes, a four-way contingency table of sample proportions is formed from the observed data. Using this table and the midpoint of each class for each variable are used to find the observed values of $E(YZ_1)$, $E(YZ_2)$ and $E(YZ_3)$, which define three dependence constraints. In addition, observed Y-marginal totals produce 12 constraints and observed $\mathbf{Z}$-marginal totals altogether produce $9 \times 13 \times 2 = 234$ constraints, which make a total of 249 constraints. Multiplying the observed marginal distributions of Y and $\mathbf{Z}$, a four-way contingency table is formed, which is the starting point of the algorithm. The algorithm stops when all constraints are satisfied subject to a prespecified value.

After the algorithm has converged, we find *model 1*, $\beta_1^* = -4.76 \times 10^{-3}$, $\beta_2^* = 1.01 \times 10^{-3}$ and $\beta_3^* = 16.45 \times 10^{-3}$, with $\mathrm{sd}(\beta_1^*) = 0.51 \times 10^{-3}$, $\mathrm{sd}(\beta_2^*) = 0.09 \times 10^{-3}$ and $\mathrm{sd}(\beta_3^*) = 2.60 \times 10^{-3}$. Thus all three coefficients are statistically significant. Although these values are on a par with those in Luo and Tsai (2012), model 1 *maintains* the observed Y and $\mathbf{Z}$ marginals. Luo and Tsai (2012) found the interaction terms to be statistically insignificant, so we did not fit them either. If we keep Z_2 and Z_3 at a fixed level, then the likelihood ratio that a patient's score increases by 10 s will increase by a factor of $\exp\{(0.00476)10\} = 1.05$, should the patient have 1 year less of education, for all values of z_1.

The multiple linear regression line is obtained by using all data values and is given by $\hat{y} = 42.17 - 2.25z_1 + 0.43z_2 + 6.91z_3$, with standard deviations of coefficients 12.08, 0.32, 0.14 and 2.41 respectively; thus all are significant. Figs 1 and 2 show the plot of the fitted model with *not transformed* variables Z_1 and Z_2 ($E(Y|Z_i)$ *versus* Z_i for $i = 1, 2$). The regression line is drawn by fixing the values of other predictors at their means. Also, 95% bootstrap bounds are shown in each plot. Since our goal is to fit a model based on dependence between Y and the covariates, it would be of interest if the linear relationship between Y and some function of Z_is is any stronger than that between Y and Z_is. Box and Tidwell (1962) suggested appropriate power transformations of independent covariates that may be useful for our purpose. A plot of Z_1 *versus* Z_2 reveals that independence may be assumed (see the on-line supplement). Considering $(Y, Z_1^{\alpha_1}, Z_2^{\alpha_2})$, the Box–Tidwell method suggests the estimates $\hat{\alpha}_1 = -0.52$ and $\hat{\alpha}_2 = 0$. Hence, we consider the transformations $Z_1^{-0.52}$ and $\log(Z_2)$; note that $\mathrm{corr}(Y, Z_1^{-0.52}) = -0.45$ and $\mathrm{corr}\{Y, \log(Z_2)\} = 0.22$ are slightly stronger than those without transformations. So we fit the

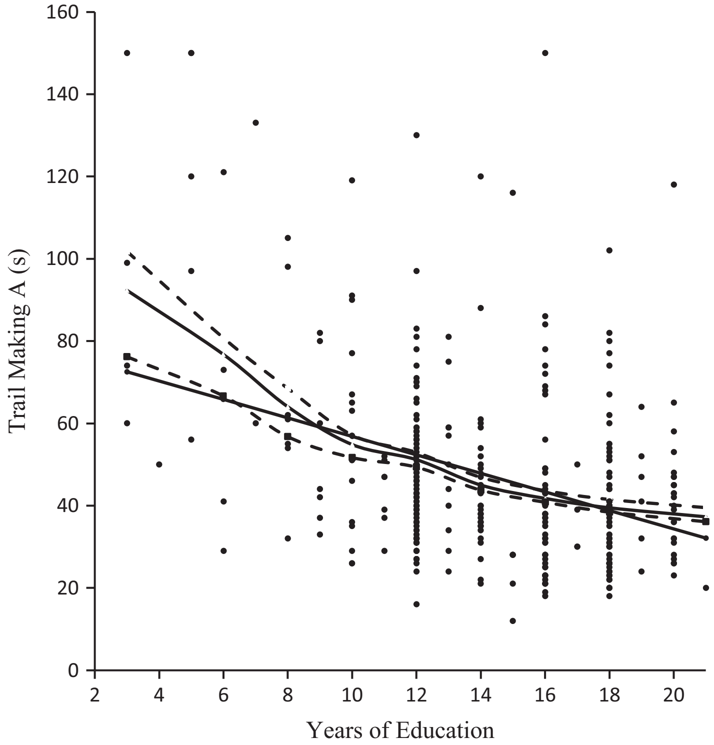

Fig. 1. Fitted trail making A (s) scores *versus* years of education using *not transformed* variables; ——, method proposed; – – –, 95% bootstrap bounds; ╲, multiple linear regression

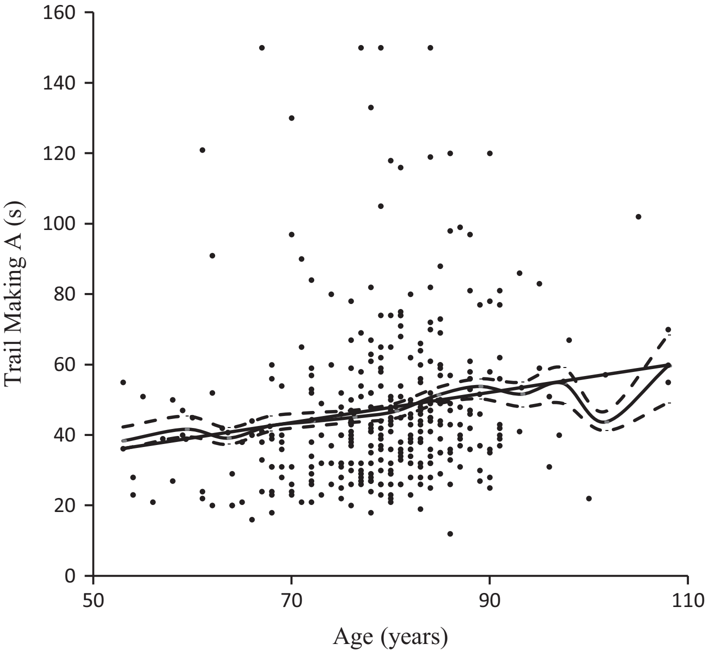

Fig. 2. Fitted trail making A (s) scores *versus* age (years) using *not transformed* variables; ———, method proposed; – – –, 95% bootstrap bounds; ⁄, multiple linear regression

model for Y by using the transformed covariates $(Z_1^{-0.52}, \log(Z_2), Z_3)$. As before we choose k for each variable in Table 4 (in the on-line supplement). A new four-way contingency table is formed by using the *transformed* variables, and similar steps are followed to those before. Then we apply the algorithm for a total of $3+12+13\times6\times2=171$ constraints, and find *model 2*: $\beta_1^{**}=0.25, \beta_2^{**}=0.07$ and $\beta_3^{**}=0.02$, with $\text{sd}(\beta_1^{**})=45.32\times10^{-3}$, $\text{sd}(\beta_2^{**})=3.02\times10^{-3}$ and $\text{sd}(\beta_3^{**})=3.18\times10^{-3}$. It is seen that, for model 2, all three coefficients are statistically significant like in model 1. Next we also fit a model using three interactions between the transformed variables; however, the interactions turned out to be insignificant. So we stayed with the model with no interactions like in model 2.

Figs 3 and 4 show the plot of the fitted model with *transformed* variables Z_1 and Z_2. The same regression line is drawn as before. Also, 95% bootstrap bounds are shown in each plot. From these graphs, it is clear that model 2 attends to the non-linear nature of the data more than model 1 does. From equation (2.8), if we keep Z_2 and Z_3 at a fixed level in model 2, then the likelihood ratio that a patient's score increases by 10 s will increase by a factor of $\exp[0.25\{(z_1-1)^{-0.5}-z_1^{-0.5}\}10]$, should the patient have 1 year less education. Clearly this depends on the value of z_1 (as opposed to model 1), such as it equals 1.38, 1.06 and 1.01 when $z_1=3,8,21$ respectively. The residual plots that were obtained from the two models are given in Figs 9–12 in the on-line supplement. From these plots, we observe that most residuals are around zero except a few which have very large positive values. Those individuals performed much better than expected.

7.2. Prostate cancer data

Stamey *et al.* (1989) examined the correlations between the level of prostate-specific antigen and several clinical measures in 97 men who were about to receive a radical prostatectomy. The goal

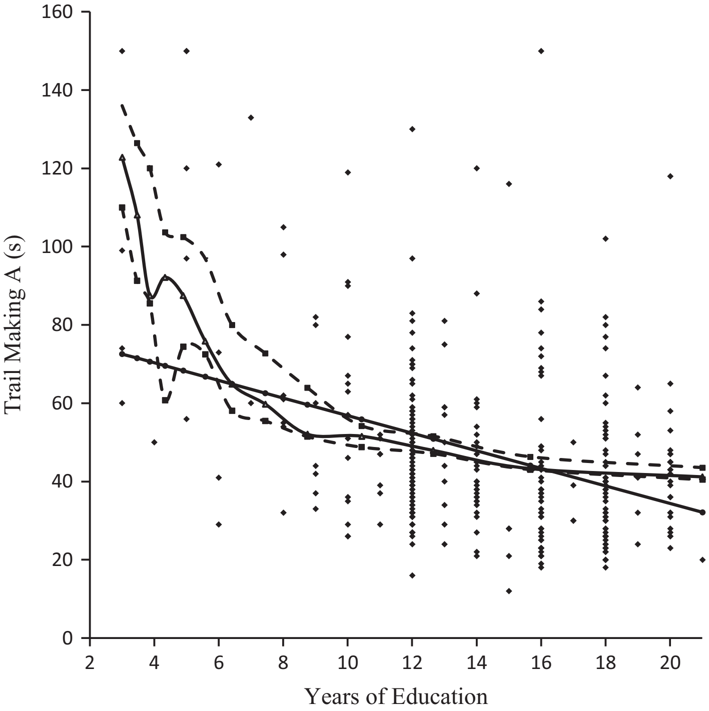

Fig. 3. Fitted trail making A (s) scores *versus* years of education using Box–Tidwell *transformed* variables; ———, method proposed; – – –, 95% bootstrap bounds; ╲, multiple linear regression

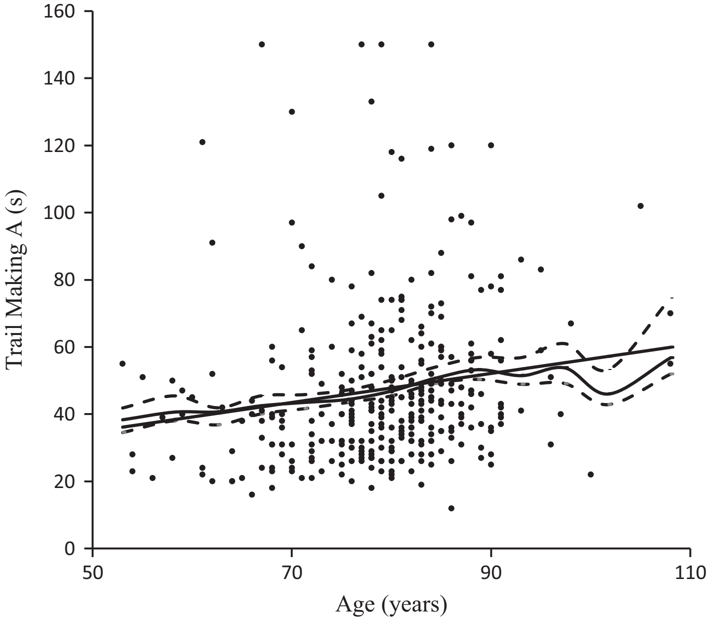

Fig. 4. Fitted trail making A (s) scores *versus* age (years) using Box–Tidwell *transformed* variables; ———, method proposed; – – –, 95% bootstrap bounds; ╱, multiple linear regression

(Hastie *et al.*, 2009) is to predict Y, the logarithm of prostate-specific antigen level, lpsa, using a number of measurements: Z_1, the logarithm of cancer volume, lcavol, Z_2, the logarithm of prostate weight, lweight, Z_3, age, and Z_4, the logarithm of capsular penetration, lcp; all variables are continuous. We find that $\mathrm{corr}(Y, Z_1) = 0.73$, $\mathrm{corr}(Y, Z_2) = 0.43$, $\mathrm{corr}(Y, Z_3) = 0.17$ and $\mathrm{corr}(Y, Z_4) = 0.55$. Further, possible interactions between these covariates are also of interest, such as, we find that $\mathrm{corr}(Y, Z_1 {*} Z_2) = 0.74$, $\mathrm{corr}(Y, Z_1 {*} Z_3) = 0.71$ and $\mathrm{corr}(Y, Z_1 {*} Z_4) = 0.36$. In this example, power functions of predictors did not significantly increase correlations with Y.

The number of classes for the variables are determined as described earlier (see Table 4 in the on-line supplement). These classes are used to form a five-way contingency table of sample proportions from the observed data. Next, this table and the midpoints of each class for each variable are used to find the observed values of $E(YZ_1)$, $E(YZ_2)$, $E(YZ_3)$ and $E(YZ_4)$, which define four dependence constraints. In addition, the observed Y-marginal totals produce seven constraints and observed $\mathbf{Z}$-marginal totals altogether produce $6 \times 4 \times 7 \times 7 = 1176$ constraints, yielding a total of 1187 constraints. The same steps are followed as in the previous example to find the estimates of β_is. To decide on an appropriate model, first we fit the model with all four predictors.

(a) *Model 1*: predictors (lcavol, lweight, age, lcp), $\hat{\beta} = (0.95, 1.40, -0.02, 0.37)$ and $\mathrm{sd}(\hat{\beta}) = (0.08, 0.17, 0.01, 0.06)$. Here 'age' is not a significant predictor.

Next we consider the remaining three variables, with their interactions in model 2. The same process as described above is followed.

(b) *Model 2*: predictors (lcavol, lweight, lcp, lcavol $*$ lweight, lcavol $*$ lcp, lweight $*$ lcp), $\hat{\beta} = (1.02, 1.41, 0.40, –1.48, –0.57, –0.16)$ and $\mathrm{sd}(\hat{\beta}) = (0.08, 0.18, 0.06, 0.87, 0.55, 0.38)$. Here none of the interaction terms is significant. So we refit the model without any of the interaction terms with three predictors.

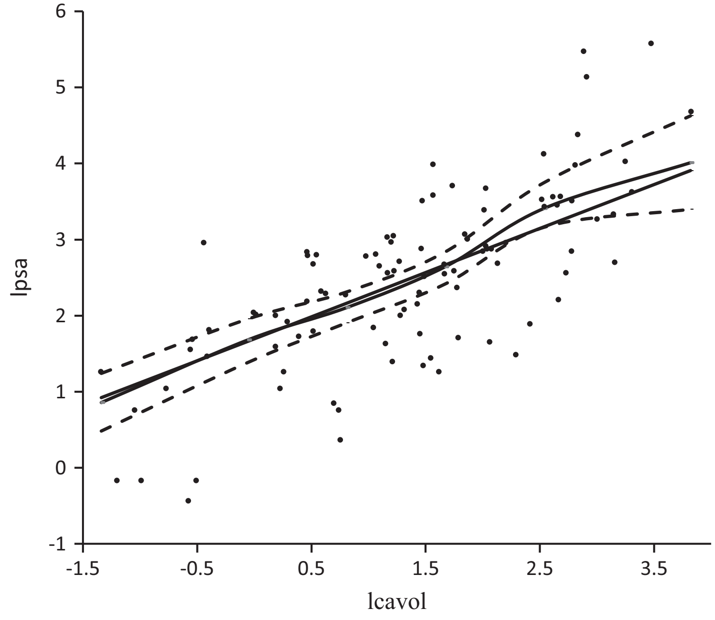

Fig. 5. Fitted prostate lpsa-values *versus* lcavol from model 3; ——, method proposed; – – –, 95% bootstrap bounds; /, multiple linear regression

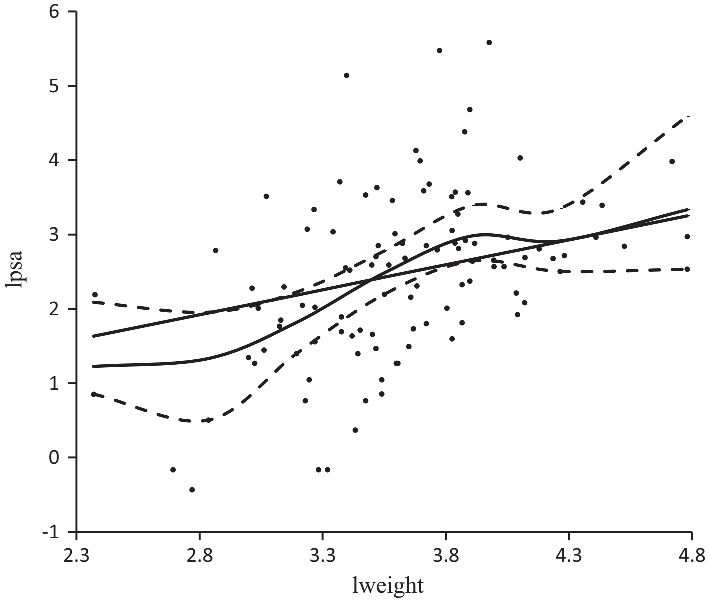

Fig. 6. Fitted prostate lpsa-values *versus* lweight from model 3; ——, method proposed; — — —, 95% bootstrap bounds; ⟋, multiple linear regression

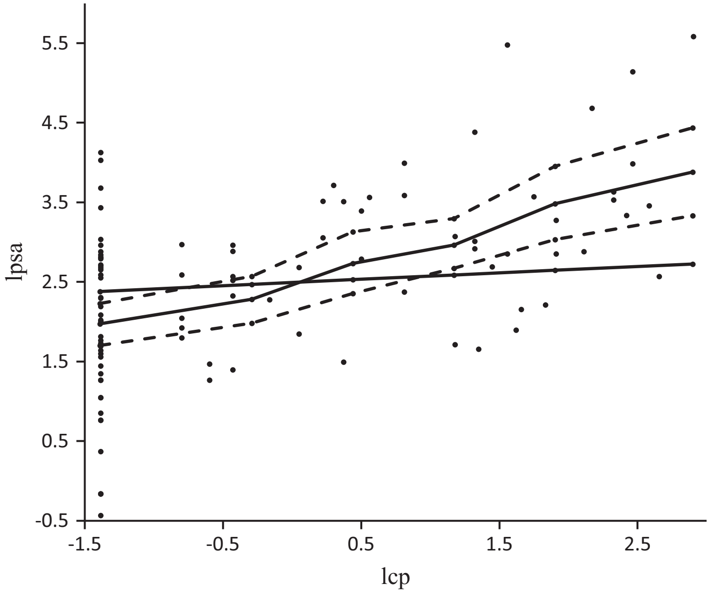

Fig. 7. Fitted prostate lpsa-values *versus* lcp from model 3; ——, method proposed; — — —, 95% bootstrap bounds; ⟋, multiple linear regression (where lcp is not significant)

(c) *Model3*: predictors (lcavol, lweight, lcp), $\hat{\beta} = (0.91, 1.27, 0.38)$ and $\mathrm{sd}(\hat{\beta}) = (0.07, 0.06, 0.05)$.

In Figs 5–7, we have plotted the conditional mean of lpsa given lcavol, lweight and lcp respectively, for model 3, and the 95% bootstrap bounds. The multiple linear regression curve is given by lpsa $= -0.73 + 0.58\,\mathrm{lcavol} + 0.67\,\mathrm{lweight} + 0.09\,\mathrm{lcp}$, with standard deviations of the three predictors $0.09, 0.18$ and 0.07 respectively. Hence lcp is not a significant predictor in the linear regression model. In each figure, the regression line is drawn by fixing the other predictors at their mean values.

The residual plots that were obtained from model 3 are given in Figs 13–15 in the on-line supplement. Since the residual plot *versus* lcp was deemed unsatisfactory, we have experimented with values other than $k=4$ for lcp and decided to go with $k=6$ for lcp; however, this did not cause a big change in the values of $\hat{\beta}_i$, and $\mathrm{sd}(\hat{\beta}_i)$, for any i. Thus all the results and figures that are presented here use $k=6$ for lcp. Scott (1979) mentioned that a value that is close to his suggested value would be acceptable for data analysis.

8. Discussion

This paper proposes a model that is characterized by the marginal distributions of the univariate response, joint distribution of covariates and the correlations between the response and each covariate. A key feature of the method proposed is that no functional form on the conditional distribution of the response given the covariates is assumed.

This paper uses the minimum relative entropy principle. An important point to note is that using this principle we find a $Q \in \mathbb{C}$ which is closest to P (in KL distance measure), and safeguards *against* using any other information in the process. Thus it may be that there are some very specific deterministic relationships between Y and $\mathbf{Z}$, but following this principle we might or might not be able to detect it exactly because that specific relationship is not described in $\mathbb{C}$. Thus the *optimality interpretation* of theorem 1 is that the procedure proposed chooses the solution from a pool ($\mathbb{C}$) of all distributions that shares the same properties of correlation and marginal distributions as those of the true $(Y, \mathbf{Z})$ (but the exact relationship between Y and $\mathbf{Z}$ is unspecified) and is closest to the joint distribution of independent Y and $\mathbf{Z}$.

To investigate the nature of the fitted models for different possible relationships between Y and Z further, we consider three cases below. First we consider a straight line relationship between Y and Z as in $Y=3+2Z$, where $Z \sim \mathrm{unif}(-1,1)$ and $Y \sim \mathrm{unif}(3,5)$. Here $\mathrm{corr}(Y,Z)=1$. We define the set $\mathbb{C}$ as

$$\mathbb{C}=\{Q: E(YZ)=c, Y \sim P_1, Z \sim P_2\},$$

where $P_1=\mathrm{unif}(3,5)$ and $P_2=\mathrm{unif}(-1,1)$, for some constant c. Let $P=P_1P_2$, and P_1 and P_2 be independent. Then $P \notin \mathbb{C}$. Although the exact I-projection of P onto $\mathbb{C}$ can be found as in equation (2.3), here we construct an estimate. On the basis of an observed sample of size 100 from the joint distribution of (Y,Z), we form $\mathbb{K}$ as in equation (4.2); using the algorithm a fitted model is found, and this process is repeated 1000 times. Fig. 8(a) shows the actual model and the average of fitted models from 1000 simulations. Here the averaged fitted model matches exactly the actual model.

Next we let Y and Z have a monotone relationship, but not a straight line. Let $Y=\exp(Z)/\{1+\exp(Z)\}$, where $Z \sim \mathrm{unif}(-5,5)$. Here $\mathrm{corr}(Y,Z) \approx 0.97$, and $\mathrm{corr}(Y,Z^3) \approx 0.8$. Here $Y \sim f_Y(y) = 1/\{2y(1-y)\}, 0.007<y<0.993$, and $f_Y(y)=0$ otherwise. Define $\mathbb{C}=\{Q: E(YZ)=c_1, E(YZ^3) = c_2, Y \sim P_1, Z \sim P_2\}$, where P_1 corresponds to $f_Y(y)$ and $P_2=\mathrm{unif}(-5,5)$, for some constants c_1 and c_2. Let $P=P_1P_2$, and P_1 and P_2 be independent. Then the I-projection of P onto $\mathbb{C}$ is the solution to the problem as in equation (2.3). To estimate, on the basis of an observed sample of size 100, we form $\mathbb{K}$. A fitted model is found, and this process is repeated 1000 times. Fig. 8(b) shows the actual model and the average of fitted models from 1000 simulations. Our procedure cannot locate the specific relationship $Y=\exp(Z)/\{1+\exp(Z)\}$ exactly, but it can find a model which satisfies the properties described in $\mathbb{C}$, and is closest to P. The fitted model would be closer to the actual model if more correlation constraints are considered, e.g. $E(YZ^5)=c_3$, in $\mathbb{C}$.

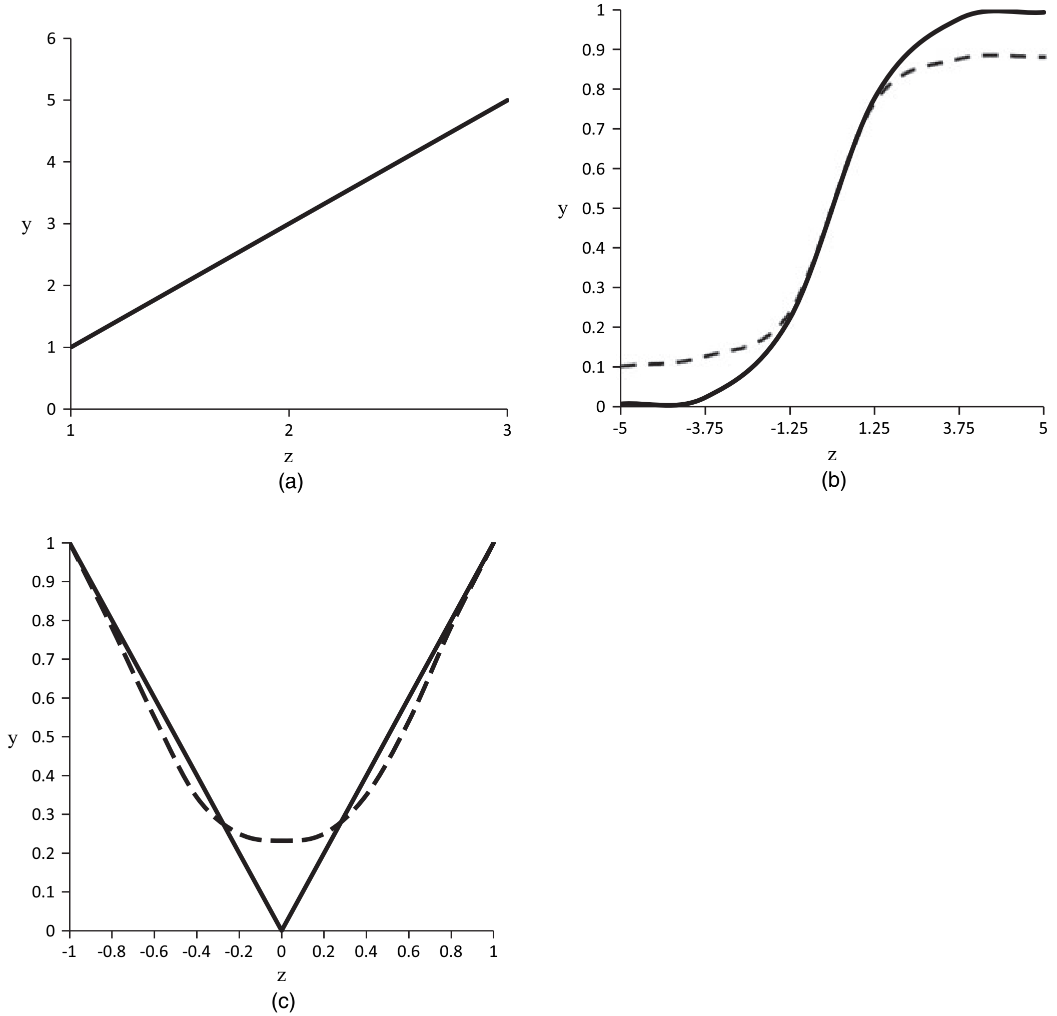

Fig. 8. (a) True and fitted models coincide, (b) monotone relationship (———, true model; – – –, fitted model) and (c) non-monotone relationship (———, true model; – – –, fitted model)

Next suppose that the relationship between Y and Z is *not* monotone. Let $Y=|Z|$ where $Z\sim\text{unif}(-1,1)$. Here $\text{corr}(Y,Z)=0$. Let

$$\mathbb{C}=\{Q:E(YZ)=0, Y\sim P_1, Z\sim P_2\},$$

where $P_1=\text{unif}(0,1)$ and $P_2=\text{unif}(-1,1)$. Let $P=P_1P_2$ and P_1 and P_2 be independent. Then $P\in\mathbb{C}$, and the I-projection of P onto $\mathbb{C}$ is itself, or $\beta=0$ in the expression (2.4) of the joint distribution of the solution. Thus we cannot locate the true model.

However, note that $\text{corr}(Y,Z^2)\approx 0.97$. Incorporating this information, define

$$\mathbb{C}'=\{Q:E(YZ)=0, E(YZ^2)=c, Y\sim P_1, Z\sim P_2\},$$

for some constant c. Then $P\notin\mathbb{C}'$ ($\mathbb{C}'\subset\mathbb{C}$). Again, the I-projection of P onto $\mathbb{C}'$ is given by equation (2.3). To estimate, on the basis of an observed sample of size 100, we form $\mathbb{K}$ and a fitted model is found. This is repeated 1000 times. Fig. 8(c) shows the actual model and the average of fitted models from 1000 simulations. Here the averaged fitted model is close to the actual model.

It may be concluded that correlations alone may not be sufficient to characterize the underlying dependence structure completely. However, the last example shows that including the additional constraint corr(Y, Z^2) in $\mathbb{C}$ is helpful in finding a suitable fitted model. Thus we recommend power transformations of Y and/or $\mathbf{Z}$ in constructing $\mathbb{C}$. See also Section 7.

The sensitivity of the choice of transformations can be seen in the trail making data example where we presented results with and without transformations. We have proposed the use of the Box and Tidwell (1962) method to find a suitable transformation, which requires independent covariates. For real data with dependent covariates, orthogonality of covariate vectors may be obtained by using $\mathbf{S}_\mathbf{z}^{-1/2}\mathbf{Z}$ as covariates instead of $\mathbf{Z}$, where $\mathbf{S}_\mathbf{z} = \text{cov}(\mathbf{Z})$.

The characterization of known models like those described in Section 3 through dependence using correlations between the response and each covariate appears to be new. This observation has led to the algorithm that is given in the paper. The *practical advantage* is that the algorithm proposed always converges, is simple to use (e.g. no matrix inversion as in the NR method and is applicable for any response and covariates. Robustness of the method to model misclassification is demonstrated in simulations for both continuous and discrete cases. The reason for this benefit is that it does not depend on any likelihood, as opposed to maximum likelihood estimators.

Acknowledgements

The authors thank the Joint Editor, the Associate Editor and the referees for their detailed reviews and encouraging comments. They helped to improve the manuscript greatly.

References

Agresti, A. (2013) *Categorical Data Analysis*. New York: Wiley.

Bhattacharya, B. (2006) An iterative procedure for general probability measures to obtain I-projection onto intersection of convex sets. *Ann. Statist.*, **34**, 878–902.

Bhattacharya, B. and Dykstra, R. (1995) A general duality approach to I-projections. *J. Statist. Planng Inf.*, **47**, 203–216.

Bhattacharya, B. and Dykstra, R. (1997) A Fenchel duality aspect of iterative I-projection procedures. *Ann. Inst. Statist. Math.*, **49**, 435–446.

Box, G. E. P. and Tidwell, P. W. (1962) Transformation of the independent variables. *Technometrics*, **4**, 531–542.

Chan, K. C. G. (2013) Nuisance parameter elimination for proportional likelihood ratio models with nonignorable missingness and random truncation. *Biometrika*, **100**, 269–276.

Csiszár, I. (1975) I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, **3**, 146–158.

Fokianos, K. and Kaimi, I. (2006) On the effect of misspecifying the density ratio model. *Ann. Inst. Statist. Math.*, **58**, 475–497.

Gilbert, P. B., Lele, S. R. and Vardi, Y. (1999) Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*, **86**, 27–43.

Good, I. J. (1976) On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.*, **4**, 1159–1189.

Guererro, V. M. and Johnson, R. A. (1982) Use of the Box-Cox transformation with binary response models. *Biometrika*, **69**, 309–314.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. New York: Springer.

Huang, A. and Rathouz, P. J. (2012) Proportional likelihood ratio models for mean regression. *Biometrika*, **99**, 223–229.

Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econmetr.*, **58**, 71–120.

Jaynes, E. T. (1957) Information theory and statistical mechanics. *Phys. Rev.*, **106**, 620–630.

Joe, H. (1989) Relative entropy measures of multivariate dependence. *J. Am. Statist. Ass.*, **84**, 157–164.

Kay, R. and Little, S. (1987) Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, **74**, 495–501.

Luo, X. and Tsai, W. Y. (2012) A proportional likelihood ratio model. *Biometrika*, **99**, 211–222.

Qin, J. (1998) Inferences for case-control data and semiparametric two-sample density ratio models. *Biometrika*, **85**, 619–630.

Rathouz, P. J. and Gao, L. P. (2009) Generalized linear models with unspecified reference distribution. *Biostatistics*, **10**, 205–218.

Reshef, D. N., Reshef, Y. A., Finucane, H. G., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. and Sabeti, P. C. (2011) Detecting novel associations in large data sets. *Science*, **334**, 1518–1522.

Rockafellar, R. T. (1970) *Convex Analysis*. Princeton: Princeton University Press.

Scott, W. D. (1979) On optimal and data-based histograms. *Biometrika*, **66**, 605–610.

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A. and Yang, N. (1989) Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II, Radical prostatectomy treated patients. *J. Urol.*, **141**, 1076–1083.

Szekely, G. J. and Rizzo, M. L. (2009) Brownian distance covariance. *Ann. Appl. Statist.*, **3**, 1236–1265.

Szekely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007) Measuring and testing dependence by correlation of distances. *Ann. Statist.*, **35**, 2769–2794.

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplement to: A minimum relative entropy based correlation model between the response and covariates’.