Modelling Tails for Collinear Data with Outliers in the English Longitudinal Study of Ageing: Quantile Profile Regression

Xi Liu[1], Silvia Liverani[2,3], Kimberley J. Smith[4] and Keming Yu[1]

[1] Department of Mathematics, Brunel University London, Uxbridge UB8 3PH, UK

[2] School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, UK

[3] The Alan Turing Institute, The British Library, London NW12DB, UK

[4] Department of Psychological Sciences, School of Psychology, University of Surrey, Guildford GU2 7XH, UK

Abstract

Research has shown that high blood glucose levels are important predictors of incident diabetes. However, they are also strongly associated with other cardiometabolic risk factors such as high blood pressure, adiposity and cholesterol which are also highly correlated with one another. The aim of this analysis was to ascertain how these highly correlated cardiometabolic risk factors might be associated with high levels of blood glucose in older adults aged 50 or older from wave 2 of the English Longitudinal Study of Ageing. Due to the high collinearity of predictor variables and our interest in extreme values of blood glucose we proposed a new method, called quantile profile regression, to answer this question.

Profile regression, a Bayesian non-parametric model for clustering responses and covariates simultaneously, is a powerful tool to model the relationship between a response variable and covariates, but the standard approach of using a mixture of Gaussian distributions for the response model will not identify the underlying clusters correctly, particularly with outliers in the data or heavy tail distribution of the response. Therefore, we propose quantile profile regression to model the response variable with an asymmetric Laplace distribution, allowing us to model more accurately clusters which are asymmetric and predict more accurately for extreme values of the response variable and/or outliers.

Our new method performs more accurately in simulations when compared to Normal profile regression approach as well as robustly when outliers are present in the data. We conclude with an analysis of the English Longitudinal Study of Ageing.

Keywords: Asymmetric Laplace distribution, Bayesian inference, clustering, Dirichlet process mixture model, profile regression, quantile regression.

1 Introduction

The English Longitudinal Study of Ageing (ELSA) is a longitudinal cohort study of adults aged 50 or older which commenced in 1998, with data collection taking part every two years (Steptoe et al., 2012). The aim of the applied portion of our study is to ascertain how cardiometabolic risk factors might be associated with high blood glucose in people who do not currently have a diagnosis of diabetes. Research has shown that high blood glucose levels are an important predictor of incident diabetes (Tabák et al., 2012). However, it has been shown that only considering high blood glucose in prediction of diabetes risk may be overly simplistic as many other cardiometabolic risk factors that are highly correlated with high blood glucose (Haffner et al., 1990; Li et al., 2009) are also associated with diabetes risk (Ford, 2005; Kolberg et al., 2009). Thus we wanted to determine how cardiometabolic risk predictors may cluster together with an outcome of high blood glucose in people who do not have a current diagnosis of diabetes.

Because our interest is in modelling most accurately the patients with the highest levels of blood glucose, quantile regression models provide the appropriate framework to model the upper tail of the distribution of the response variable while also being robust to outliers. Quantile regression models were first introduced by Koenker and Bassett (1978) and have been applied to a wide range of applications in biostatistics, including survival analysis, ecology, earnings inequality and mobility, income and wealth distribution, value at risk and mutual fund investment styles (Knight and Ackerly, 2002; Geraci and Bottai, 2007). Quantile regression models aim at estimating either the conditional median or other quantiles of the response variable. Their main advantage over least-squares regression is their flexibility for modelling data with heterogeneous conditional distributions. Moreover, quantile regression models provide a richer characterization of the data, allowing us to consider the impact of predictors on quantiles of the response variable, not merely its conditional mean, and thus these models are robust to outliers. See Davino et al. (2013) for a more in-depth discussion of the advantages of quantile regression. However, collinearity (that is, high correlations among predictor variables) in quantile regression models leads to unreliable and unstable estimates of parameters.

Cardiometabolic risk predictors are usually highly correlated and therefore create collinearity problems when used in a standard multiple regression model, a well known issue in many statistical applications when trying to assess meaningful relationships between predictors and response variables. A common approach in this case is to examine each predictor separately, to avoid instability in the estimates, but compromising the possibility of learning about the complex relationships involving several predictors at the same time. An alternative approach is to combine the correlated variables into summary indexes and to assess the relationship of these with the outcome of interest, but this approach loses information on the single variables included in the summary. In this paper we discuss a third approach, which identifies subgroups of the observations such that the main outcome variable is related to (some of) the covariates (or profiles) that have been collected. More specifically, clinicians may be interested in identifying suitable subgroups of

patients presenting similar features; this categorisation can be used, for example, to suitably apply the optimal treatment for the (sub)population that will benefit the most. One class of suitable clustering model that has been proposed as an alternative to regression models when dealing with collinearity are Dirichlet process mixture models (Dunson et al., 2008). In particular we will consider profile regression: first proposed by Molitor et al. (2010), it is a semi-parametric Bayesian method where covariate profiles are allocated to clusters and associated via a regression model with a relevant outcome. This method was implemented by Liverani et al. (2015) in the R package PReMiuM and applied in a variety of areas, including, for example, epidemiology (Hastie et al., 2013; Molitor et al., 2014; Pirani et al., 2015; Mattei et al., 2016; Liverani et al., 2016; Coker et al., 2016, 2018) and genetics (Papathomas et al., 2012).

Therefore, in this paper we propose a new profile regression model with a quantile regression submodel to allow a careful modelling of the data when the interest is on lower or upper tails of the distribution of the response profiles rather than their mean. We name this new method 'quantile profile regression'. Quantile profile regression includes a mixture of asymmetric Laplace distributions (ALD), which were proposed by Yu and Moyeed (2001) for quantile regression in a Bayesian framework based on a 'working likelihood'. The closest work to ours is by Kottas and Kranjajić (2009) who developed a Dirichlet process mixture model of ALDs for the error distribution of a quantile regression. However their interest was in the errors and their mixture over the scale parameters, while we are interested in using a mixture of ALDs for the response which links covariate profiles to clusters (not a direct regression function of covariates) and other possible fixed factors via a regression model. Covariates and fixed effects are differentiated by their link to the response: covariates have cluster-specific parameters, while fixed effects have global (ie. non cluster-specific) parameters.

Another close proposal is by Franczak et al. (2014). They proposed the use of shifted asymmetric Laplace distributions for model-based clustering and provided an Expectation-Maximisation algorithm. Their mixture model was multivariate and aimed at classical classification problems, and for certain selected examples they outperform the Gaussian mixture models. They did not study the potential relationship between predictors and covariates, while we assume the distribution of the response variable and the covariates to be cluster dependent.

The inference for quantile profile regression is carried out by Markov chain Monte Carlo (MCMC). We have implemented three types of samplers for sampling of the Dirichlet process (DP) for profile regression: the truncated sampler by Jara et al. (2011), the slice independent by Kalli et al. (2011) and the slice dependent by Papaspiliopoulos (2008). We have added a Gibbs sampler to this implementation for the parameters of the quantile extension. Our novel mixture modelling approach is demonstrated on both simulated and real data. In these analyses, our mixture of asymmetric Laplace distributions performs favourably when compared to Gaussian profile regression.

The paper is organised as follows. Section 2 of the paper gives a brief overview of quantile regression and the asymmetric Laplace distribution. Section 3 describes the Dirichlet process mixture model for Bayesian clustering. Profile regression employing a likelihood function that is based on the asymmetric Laplace distribution is developed in Section 4. In Sections 5 and 6.1 simulated data and a well-known real dataset are used to validate our model, which is then applied to the analysis of the ELSA dataset in Section 6.2.

2 Quantile Regression and Asymmetric Laplace Distribution

Consider a standard linear quantile regression model

$$Y = \mathbf{U}^T\gamma + \varepsilon, \tag{1}$$

where Y is the response variable, $\mathbf{U}$ is a $d \times 1$ covariate vector, γ is a $d \times 1$ regression coefficient vector and ε is the error term whose pth $(0 < p < 1)$ quantile, denoted by $Q_p(\varepsilon|\mathbf{U})$, is zero. Then the pth conditional quantile of Y given $\mathbf{U}$ is given by

$$Q_p(Y|\mathbf{U}) = \mathbf{U}^T\gamma(p). \tag{2}$$

Given i.i.d. observations $\{U_i, Y_i\}_{i=1}^n$, the regression coefficient $\gamma(p)$ now depends on the quantile of interest and it can be estimated by minimising the following objective function,

$$\min_{\gamma} \sum_{i}^{n} \rho_p\left(Y_i - \mathbf{U}_i^T\gamma(p)\right), \tag{3}$$

where the loss function is the piecewise linear 'check function',

$$\rho_p(u) = u\left(p - I(u < 0)\right) \tag{4}$$

with the indicator function $I(\cdot)$.

Note that $\min_\gamma \sum_i^n \rho_p\left(Y_i - \mathbf{U}_i^T\gamma(p)\right) = \max_\gamma \prod_i^n \exp(-\rho_p\left(Y_i - \mathbf{U}_i^T\gamma(p)\right))$. Therefore, the link between the minimisation of the loss function (4) and the maximum an ALD-based likelihood function can be constructed with the $ALD(\mu, \sigma; p)$ as

$$f_p(v|\mu, \sigma) = \frac{p(1-p)}{\sigma} \exp\left\{-\rho_p\left(\frac{v-\mu}{\sigma}\right)\right\} \tag{5}$$

for $0 < p < 1$, where μ and σ are location and scale parameters respectively. Based on this ALD-based likelihood function, Yu and Moyeed (2001) and Yu and Stander (2007) and among others introduced Bayesian quantile regression. Sriram et al. (2013) provided posterior consistency of this Bayesian quantile regression method under misspecification. Hu et al. (2013) has also shown that Bayesian quantile regression methods are not sensitive to this ALD-based likelihood assumption.

3 Dirichlet Process Mixture Model

Dirichlet process mixture models are defined for data $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, regarded as exchangeable or as independently drawn from an unknown distribution. This distribution is modelled as a mixture of distributions of the form $F(\theta)$, with the mixing distribution over θ being G. The prior for this mixing distribution is a Dirichlet process with concentration parameter α and base distribution G_0 (Ferguson, 1973):

$$\begin{aligned} Y_i|\theta_i &\sim F(\theta) && (6)\\ \theta_i|G &\sim G && (7)\\ G &\sim DP(G_0, \alpha). && (8) \end{aligned}$$

An infinite mixture model will not face the misspecification of parameters in contrast to finite models, especially when using a model structure which is far from the real one, and hence will generate more stable solutions.

3.1 Profile regression

We will focus on the Dirichlet process mixture model described in Liverani et al. (2015). This model links a response vector $\mathbf{Y}$ with the covariate metrix $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_d)$ nonparametrically through their cluster membership. Also, the approach enables the potential supplemental fixed effects $\mathbf{W}$, which have an effect on the response which is the same for all clusters (also referred to as a global effect). It is worth noting that the allocated clusters are based on the joint effects of $\mathbf{X}$ and $\mathbf{Y}$, implicitly handling latent high dimensional interactions which would be quite challenging to capture via classical approaches.

Consider a response variable Y_i and a covariate profile $X_i = (x_{i,1}, ..., x_{i,d})$ for i in $1, 2, \ldots, n$. The observed data follows an infinite mixture distribution, where mixture component c has density conditional on some component specific parameters $\mathbf{\Theta}_c$ and global parameters $\mathbf{\Lambda}$. Therefore, the proposed model is given by a joint probability model for the outcome Y_i and profile X_i, where these probability models are conditionally independent within clusters:

$$f(Y_i, X_i|\mathbf{\Theta}, \mathbf{\Lambda}, \mathbf{W}_i) = \sum_{c=1}^{\infty} \psi_c f(Y_i|\mathbf{\Theta}_c, \mathbf{\Lambda}, \mathbf{W}_i) f(X_i|\mathbf{\Theta}_c, \mathbf{\Lambda}) \quad (9)$$

where $\mathbf{\Theta} = (\psi_1, \mathbf{\Theta}_1, \psi_2, \mathbf{\Theta}_2, \cdots)$, and the weight of mixture component c is given by ψ_c. The mixture weights $\boldsymbol{\psi} = \{\psi_c, c \geq 1\}$ follow a stick breaking distribution which is given by

$$\begin{aligned} \psi_c &= V_c \prod_{l<c} (1 - V_l) \quad \text{for } c \in \mathbb{Z}^+ \setminus \{1\} && (10)\\ \psi_1 &= V_1 && (11)\\ V_c &\sim \text{Beta}(1, \alpha) \quad \text{i.i.d. for } c \in \mathbb{Z}^+. && (12) \end{aligned}$$

In order to help identify the specific cluster a data point belongs to and simplify the likelihood, it is common and convenient to bring in a vector of latent allocation variables $\mathbf{Z} = (Z_1, \ldots, Z_n)$, such that $Z_i = c$ identifies the allocation of individual i to cluster c.

There is a wide range of choices for the response sub-model $f(Y_i|\mathbf{\Theta}_c, \mathbf{\Lambda}, \mathbf{W}_i)$ and the profile sub-model $f(X_i|\mathbf{\Theta}_c, \mathbf{\Lambda})$, including Gaussian, Bernoulli, Binomial, Poisson, Multinomial and Weibull distributions (Liverani et al., 2015). For example the case where for each individual i, $D_i = X_i$ is a vector of J locally independent discrete categorical random variables, where the number of categories for covariate $j = 1, 2, \ldots, J$ is K_j. Then we can write $\Theta_c = \Phi_c = (\Phi_{c,1}, \Phi_{c,2} \ldots, \Phi_{c,J})$ with $\Phi_{c,j} = (\phi_{c,j,1}, \phi_{c,j,2}, \ldots, \phi_{c,j,K_j})$ and

$$f(X_i|\mathbf{\Theta}_c, \mathbf{\Lambda}) = \prod_{j=1}^{J} \phi_{Z_i,j,X_{i,j}}. \tag{13}$$

In this case there are no global parameters Λ. We let $\Theta_0 = a = (a_1, a_2, \ldots, a_J)$, where for $j = 1, 2, \ldots, J$, $a_j = (a_{j,1}, a_{j,2}, \ldots, a_{j,K_j})$ and we adopt conjugate Dirichlet priors $\Phi_{c,j} \sim \text{Dirichlet}(a_j)$.

As another example, continuous response data is modelled by a Gaussian distribution. The parameter Θ_c is extended to contain θ_c for each cluster c. As before Λ contains β, but also σ_Y^2. These parameters allow us to write the response model as:

$$f(Y_i|\mathbf{\Theta}_c, \mathbf{\Lambda}, \mathbf{W}_i) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left\{-\frac{1}{2\sigma_Y^2}(Y_i - \lambda_i)^2\right\},$$

where $\lambda_i = \theta_{Z_i} + \beta^\top W_i$. We impose the same prior settings as for the discrete response models, with the additional prior on $\tau_Y = 1/\sigma_Y^2$ being Gamma(s_{τ_Y}, r_{τ_Y}), where s_{τ_Y} and r_{τ_Y} are the shape and rate hyper parameters that extend Θ_0. We will refer to the profile regression model with the Gaussian distribution for the response sub-model as *Gaussian profile regression*.

As seen above, the parameters of the covariates are cluster specific, so each cluster will be differentiated by the values taken by the parameters of the covariates. Fixed effects have global parameters, so the effect of variables included in the model as fixed effects will be the same for all observations, regardless of which cluster they belong to. Therefore, covariates are the variables of interest for the inference, while fixed effects are variables which we are controlling for.

Posterior inference on $\mathbf{Z}$ offers us with information concerning the clustering of the observations. We carry out inference via Markov Chain Monte Carlo using the stick-breaking construction of the Dirichlet process and the slice independent sampler proposed by Kalli et al. (2011). As in this type of models the posterior distribution is invariant to switching component labels, our implementation includes also label switching moves, which greatly improve the convergence rate of the MCMC (Hastie et al., 2015). Moreover, we have implemented in the R package PReMiuM (Liverani et al., 2015) a range of post-processing functions which deduce the clustering structure from the rich MCMC output.

4 Quantile Profile Regression

We extend profile regression to allow for asymmetric Laplace distributions for the response variable. We name this model Bayesian profile quantile regression. Let the response sub-model be

$$f(Y_i|\mathbf{\Theta}_{Z_i}, \mathbf{\Lambda}, \mathbf{W}_i) = f(Y_i|\theta_{Z_i}, \beta, \sigma_Y, \mathbf{W}_i) = \frac{p(1-p)}{\sigma_Y} \exp\left\{-\rho_p\left(\frac{Y_i - \lambda_i}{\sigma_Y}\right)\right\} \tag{14}$$

that is, $Y_i|Z_i, \mathbf{\Theta}_{Z_i}, \mathbf{\Lambda}, \mathbf{W}_i \sim \text{ALD}(\lambda_i, \sigma_Y; p)$, where $\lambda_i = \theta_{Z_i} + \beta^T \mathbf{W}_i$ and $\mathbf{\Lambda} = (\beta, \sigma_Y)$ contains the global parameters, which take the same values for all clusters, and $\mathbf{\Theta}_{Z_i} = (\theta_{Z_i})$ contains the cluster-specific parameters. The parameter p refers to the quantile of interest and it is set, not estimated from the model, depending on the aims of the analysis. For example, if a population of males has the 90% quantile of the weight distribution corresponding to obesity and we aim to investigate how some correlated predictors are related to obesity, then we could set $p = 0.9$.

For each cluster c, we adopt a t location-scale distribution for θ_c, with hyperparameters μ_θ and σ_θ with 7 degrees of freedom, as discussed by Molitor et al. (2010) and Gelman et al. (2008). For each fixed effect l, which is an element of the vector β, we adopt the same prior for β_l, but with hyperparameters μ_β and σ_θ. Additionally, we set the prior distribution of σ_Y to be an inverse Gamma with parameters s_{σ_Y} and r_{σ_Y}, which are respectively the shape and scale parameters. Adopting this conjugate prior, updates for σ_Y are simple Gibbs updates, as shown in the next section. See Liverani et al. (2015) for details on the prior distributions for the other parameters.

4.1 Inference for quantile profile regression

We discuss here the details of the sampling from the posterior distribution of the new parameter σ_Y introduced by the ALD. See Liverani et al. (2015) for details of the samplers for all other parameters of the model.

We can derive a Gibbs sampler for the global parameter σ_Y. The posterior distribution of $f(\sigma_Y|\mathbf{D})$, with $\mathbf{D} = (\mathbf{Y}, \mathbf{X})$, is given by:

$$\begin{aligned}
f(\sigma_Y|\mathbf{D}) &\propto \left\{\prod_{i=1}^{n} f(Y_i|\theta_{Z_i}, \beta, \sigma_Y, \mathbf{W}_i)\right\} p(\sigma_Y) \\
&\propto \left\{\prod_{i=1}^{n} \frac{1}{\sigma_Y} \exp\left\{-\frac{1}{\sigma_Y}(Y_i - \lambda_i)\left(p - I(Y_i < \lambda_i)\right)\right\}\right\} \frac{1}{\sigma_Y^{s_{\sigma_Y}+1}} \exp\left\{-\frac{r_{\sigma_Y}}{\sigma_Y}\right\} \\
&\propto \frac{1}{\sigma_Y^{s_{\sigma_Y}+n+1}} \exp\left\{-\frac{1}{\sigma_Y}\left(r_{\sigma_Y} + \sum_{i=1}^{n}(Y_i - \lambda_i)(p - I(Y_i < \lambda_i))\right)\right\}
\end{aligned}$$

so we have that

$$\sigma_Y|\mathbf{D} \sim \text{IG}\left(s_{\sigma_Y} + n, r_{\sigma_Y} + \sum_{i=1}^{n}(Y_i - \lambda_i)\left(p - I\left(Y_i \leq \lambda_i\right)\right)\right) \tag{15}$$

4.2 Posterior Predictive Distribution

As for profile regression, we can sample the posterior predictive distribution of pseudo-profiles. The pseudo-profiles are predictive scenarios determined by the covariates. At each iteration the predictive subjects are allocated to one of the existing clusters in accordance with their own covariate profiles. We can then derive the posterior predictive distribution of the response variable for each pseudo-profile. We compute the full posterior predictive distribution of the pseudo-profiles and then compute the quantile of interest. p.

Let X_{n+1} be a covariate profile for which we are interested in predicting the response variable, and let $\tilde{Z}^r_{n+1} = c$ if the pseudo-profile $n+1$ is allocated to cluster c at the sweep r of the MCMC sampler. We can then compute the posterior probability $p(\tilde{Z}^r_{n+1} = c|X_s, \tilde{\boldsymbol{\Theta}}^r, \tilde{\boldsymbol{\Lambda}}^r, \mathbf{D}, \mathbf{W})$ for each pseudo-profile, where $\tilde{\boldsymbol{\Theta}}^r$ are the cluster specific parameters and $\tilde{\boldsymbol{\Lambda}}^r$ the global parameters sampled at the sweep r of the MCMC sampler. We can compute $p(\tilde{Z}^r_{n+1} = c|X_s, \tilde{\boldsymbol{\Theta}}^r, \tilde{\boldsymbol{\Lambda}}^r, \mathbf{D}, \mathbf{W})$ using the slice sampler based on the covariate profile X_{n+1}. Essentially, the pseudo-profile is assigned to a cluster at each iteration of the MCMC based on the similarity to the covariate profiles in the existing clusters. Unlike the training data $\mathbf{D}$, the pseudo-profiles are never allocated to empty clusters. This is because we have no data for the empty clusters, and their parameters are draws from their prior distributions. However, this can lead to a spurious allocation if the covariate profile is significantly different from the covariate profiles of the training data.

Once the covariate profile is associated to a cluster c at sweep r, we can estimate the predictive density of a future observation as follows,

$$\begin{aligned}
&f\left(Y_{n+1}|X_{n+1} = \mathbf{X}, \boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}, \mathbf{W}, \mathbf{V}, \mathbf{D}\right) \\
&= f\left(Y_{n+1}|Z_{n+1} = c, \boldsymbol{\Theta}, \boldsymbol{\Lambda}, \mathbf{W}, \mathbf{V}, \mathbf{D}\right) f\left(Z_{n+1} = c|X_{n+1} = \mathbf{X}, \boldsymbol{\phi}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}, \mathbf{W}, \mathbf{V}, \mathbf{D}\right) \\
&= \psi_c(\mathbf{X}) f\left(Y_{n+1}|Z_{n+1} = c, \boldsymbol{\Theta}, \boldsymbol{\Lambda}, \mathbf{W}, \mathbf{V}, \mathbf{D}\right) \\
&= \psi_c(\mathbf{X}) \text{ALD}(Y_{n+1}; Z_{n+1} = c).
\end{aligned} \tag{16}$$

We obtain a posterior predictive distribution based on the allocation of the covariate profile of interest at each sweep.

5 Simulation Study

We have implemented and released quantile profile regression in the R package PReMiuM. Here, we provide the results of the application of quantile profile regression on simulated data. First we simulate data from the ALD and show that the method proposed is more effective than Gaussian profile regression to retrieve generating parameters for skewed data. Then we show that if we are interested in a specific quantile of the distribution, even when the generating mechanism is Gaussian, quantile profile regression makes more accurate predictions than Gaussian profile regression.

We simulate data from an ALD using the methods in Yu and Zhang (2005): if ξ and η are independent and identical standard exponential distributions, then

$$\frac{\xi}{p} - \frac{\eta}{1-p} \sim \text{ALD}(0, 1; p)$$

and if $A \sim \text{ALD}(0, 1; p)$ then $B \sim \text{ALD}(\mu, \sigma; p)$ when $B = \mu + \sigma A$. Five clusters were generated by drawing independent samples from the ALD for the outcome Y and the Gaussian distribution for the covariate X as follows.

$$Y_i \sim \text{ALD}(\theta_{Z_i} + \beta^T \mathbf{W}_i, \sigma_Y; q) \tag{17}$$

$$X_i \sim \text{Normal}(\mu_{Z_i}, \tau^2_{Z_i}) \tag{18}$$

with $q = 0.05$ and $i = 1, 2, \ldots, 300$. As the profile sub-model $f(X_i|\mathbf{\Theta}_{Z_i}, \mathbf{\Lambda})$ is Gaussian with parameters μ_{Z_i} and $\tau^2_{Z_i}$, the cluster-specific parameters contained in $\mathbf{\Theta}$ are $(\theta_1, \theta_2, \ldots, \theta_5) = (-200, 0, 3, 40, 150)$, $(\mu_1, \mu_2, \ldots, \mu_5) = (0, 6, -8, -3, 5)$ and $(\tau^2_1, \tau^2_2, \ldots, \tau^2_5) = (6, 7, 4, 10, 17)$. When the observation i belongs to cluster c, the allocation variable $Z_i = c$. The sizes of the five simulated clusters were 600, 200, 400, 300 and 800 observations respectively. The coefficients β were set equal to 0, therefore omitting the fixed effects. We set $\sigma_Y = 1$. Figure 1 shows the first set of simulated data.

We use quantile profile regression and Gaussian profile regression, both available in the R package PReMiuM. We set the same priors for both models and keep the hyperparameters constant. The parameters θ_c have a t-distribution with 7 degrees of freedom, mean 0 and scale 2.5. The shape and scale of σ_Y and τ^2 are 2.5 and 2.5 respectively. The prior on the mean vector for μ_c has the empirical covariate means as mean and the inverse of the diagonal matrix with elements equal to square of empirical range for each covariate, multiplied by the number of covariates, as precision matrix. The Gamma prior on the Dirichlet parameter α has a shape parameter of 2 and rate of 1. For all simulations below, we ran 20,000 iterations of burn-in and 20,000 iterations after that. We obtain good convergence diagnostics on the trace, density and autocorrelation for various parameters (not shown). See Hastie et al. (2015) for more details on convergence for this type of model.

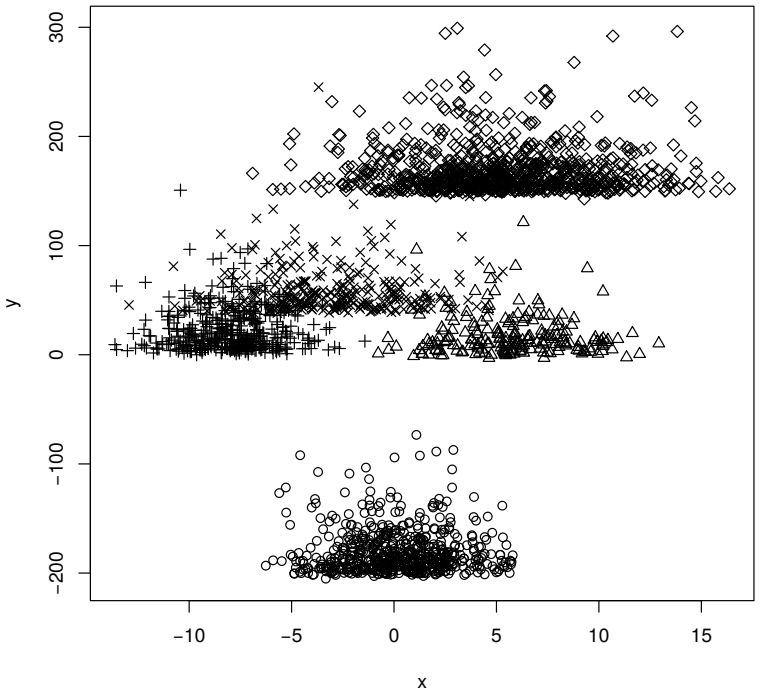

Figure 1: Data simulated as by Equations (17) and (18). The five generating clusters can be identified by the different symbols used for the data points.

We initially simulated the ALD data with $q = 0.05$ and run the algorithm using different settings of the parameter p. Table 1 shows the mean of the posterior distributions of θ applying the proposed quantile profile regression (with $p = 0.05$) and Gaussian profile regression, averaged over 100 repetitions. Quantile profile regression provides more accurate estimations of the generating parameters. On the other hand, when the data is generated with $q = 0.95$ (simulations and results not shown), the accuracy is highest for $p = 0.95$. As the choice of p is driven by the application and chosen at priori, we only show results for $q = 0.05$ below, without loss of generality.

Table 1: Posterior means of θ, averaged over 100 repetitions. The first row gives the generating values of the parameter θ for the five clusters. The second row gives the posterior means for the clusters obtained applying quantile profile regression with parameter $p = 0.05$ and the last row applying Gaussian profile regression.

	1	2	3	4	5
θ	-200.00	0.00	3.00	40.00	150.00
quantile p=0.05	-200.01	0.52	5.00	35.39	149.92
Gaussian	-181.40	16.52	21.72	51.50	167.89

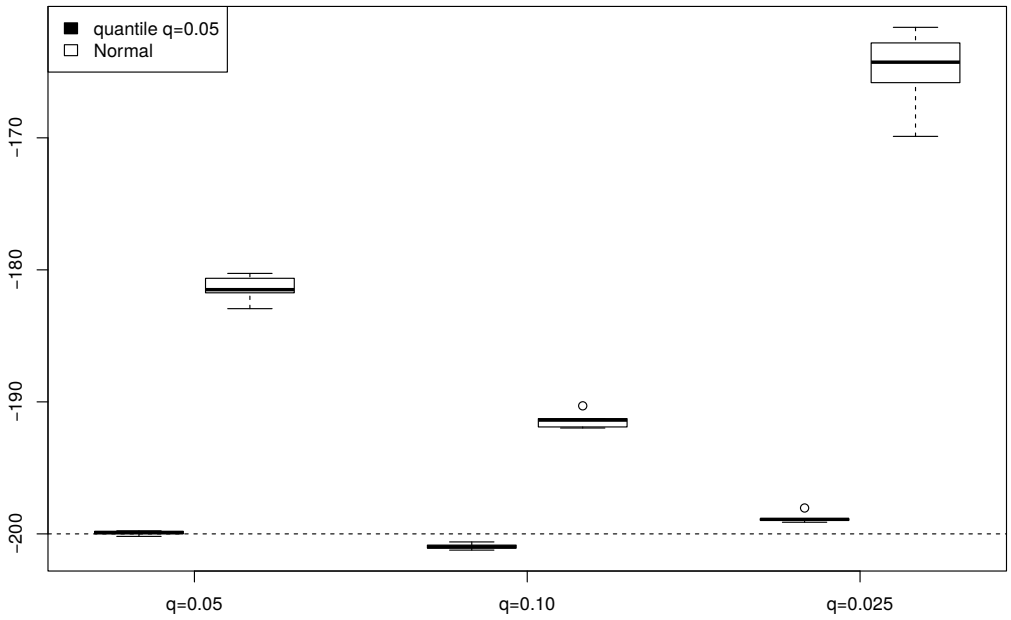

Figure 2: Boxplots of the posterior mean of θ_1 over 100 runs for quantile profile regression with $p = 0.05$ and Gaussian profile regression, repeated for different generating values of $q = 0.05, 0.10, 0.025$. The horizontal dashed line marks the generating value $\theta_1 = -200$.

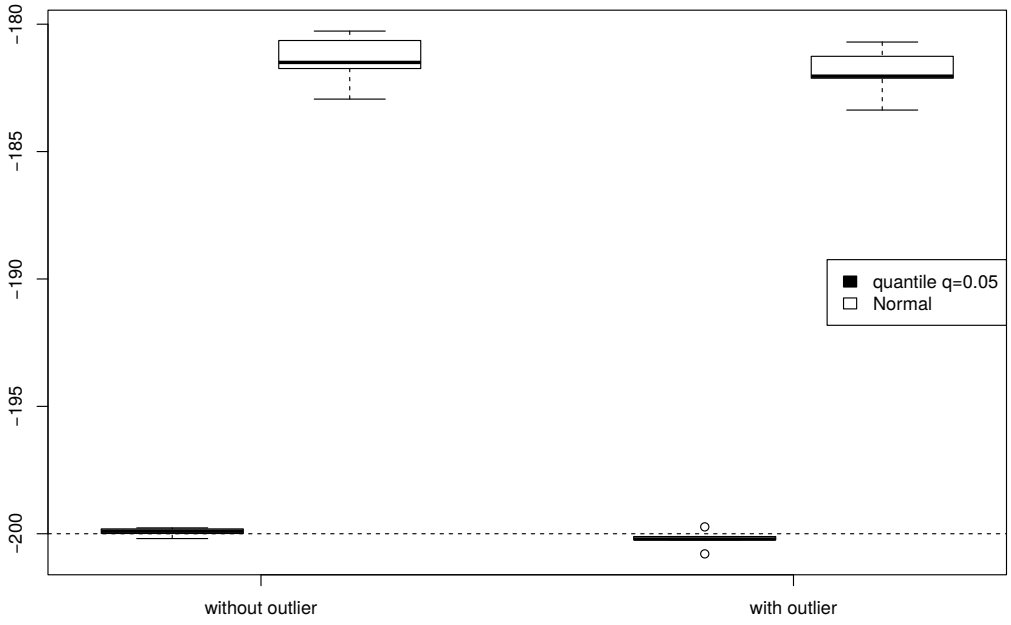

Figure 3: Boxplots of the posterior mean of θ_1 over 100 runs for quantile profile regression with $p = 0.05$ and Gaussian profile regression, comparing the results on the original data and adding an outlier at $x = 15$ and $y = -320$. The horizontal dashed line marks the generating value $\theta_1 = -200$.

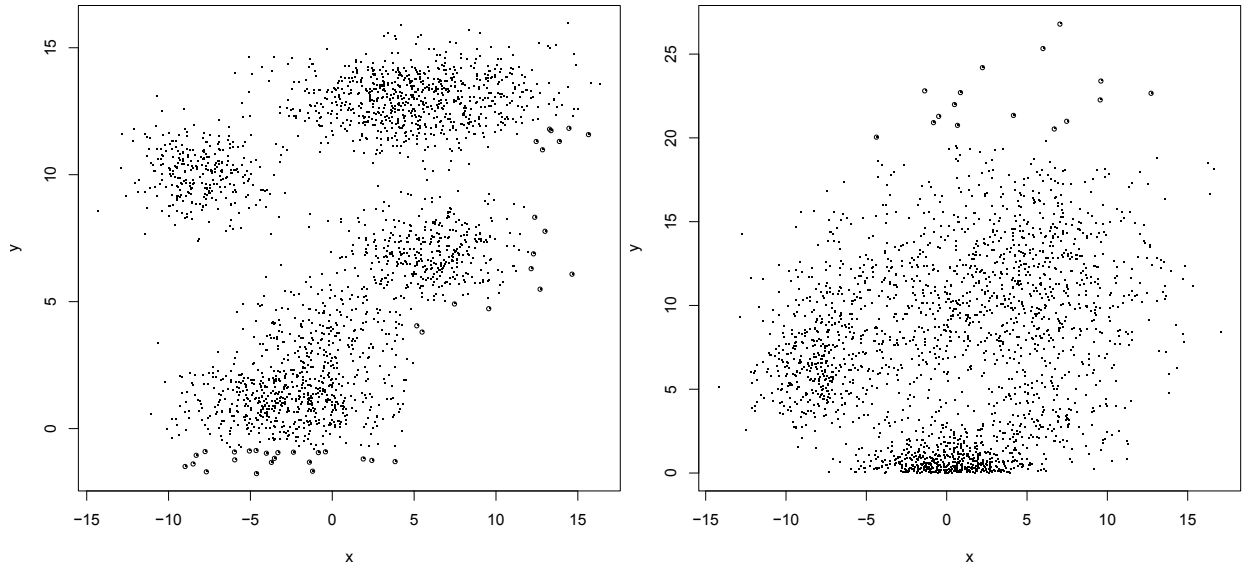

Figure 4: The two dataset with the response variable Y simulated from a mixture of Gaussian distributions (left hand side) and from a mixture of Gamma distributions (right hand side). The circled observations are the ones of interest for prediction.

Therefore, as we are in a setting where our interest is in the lowest quantiles of the data, we concentrate on the estimation of θ_1, the parameter of the cluster corresponding to the lowest values of the outcome Y. Figure 2 shows the boxplots of the posterior mean of θ_1 over 100 runs of quantile profile regression and Gaussian profile regression against its generating value of -200. Quantile profile regression consistently performs more accurately than the alternative method. This is not due to the unfair advantage of knowing the generating value $q = 0.05$, as the same plot shows also that quantile profile regression outperforms Gaussian profile regression also for $q = 0.1$ and $q = 0.025$.

Moreover, the results were also robust to the addition of an outlying observation which took the values $x = 15$ and $y = -320$. The results are shown in Figure 3.

To test whether our proposed method works well when the data generating mechanism is unknown, we simulated two additional datasets. In the first simulated dataset Y is simulated from Gaussian distributions as follows

$$Y_i \sim \text{Normal}(\theta_{Z_i} + \beta^T W_i, \sigma_Y^2) \tag{19}$$

with $i = 1, 2, \ldots, 300$ with $\sigma_Y^2 = 1$. In the second simulated dataset Y is simulated from (skewed) Gamma distributions as follows

$$Y_i \sim \text{Gamma}(\theta_{Z_i} + \beta^T W_i, \alpha) \tag{20}$$

with $i = 1, 2, \ldots, 300$ and the rate $\alpha = 1$. The cluster-specific parameters are $(\theta_1, \theta_2, \ldots, \theta_5) = (-6, -2, 0, 3, 6)$ for both simulated datasets. The sizes of the five simulated clusters were 300, 600, 200, 400 and 800 obser-

vations respectively for both simulated datasets, as for the previous simulation. The coefficients β were set equal to 0, therefore omitting the fixed effects. For both simulated datasets, the covariates are simulated from $X_i \sim \text{Normal}(\mu_{Z_i}, \tau^2_{Z_i})$ with $(\mu_1, \mu_2, \ldots, \mu_5) = (-3, 0, 6, -8, 5)$ and $(\tau^2_1, \tau^2_2, \ldots, \tau^2_5) = (10, 6, 7, 4, 17)$. The data is shown in Figure 4.

As expected, for data simulated from Gaussian distributions, Gaussian profile regression outperforms quantile profile regression when estimating the posterior distribution of θ_c and predicting the outcome Y (not shown). However, the main aim of quantile profile regression is to predict extreme values of the response variable Y. Therefore we will compare the prediction accuracy of Gaussian profile regression and quantile profile regression when the prediction concerns the lowest values of Y for the datasets generated from Gaussian and Gamma distributions. We do this by separating the entire sample into training data and validation data. The latter dataset is formed by the points circled in Figure 4 (the extreme observations of Y for different values of X). We then compute two measures of prediction accuracy for the the validation dataset.

The first measure of prediction accuracy that we use to compare the predictive power of quantile profile regression against Gaussian profile regression is the root mean square error (RMSE) of the predicted values with respect to the observed outcome. This measure of goodness of fit is given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n}} \tag{21}$$

where $\hat{Y}_i$ denotes the mean of the posterior predictive distribution for Y_i. As well as measuring predictive accuracy for point predictions using a scoring function like the RMSE, we also use a measure of predictive accuracy for probabilistic prediction based on proper scoring rules, the continuous ranked probability score (CRPS). We used the implementation by Jordan et al. (2018). For both these methods a lower score indicates a better forecast. We find that quantile profile regression outperforms Gaussian profile regression.

The mean and standard deviation results for 100 repetitions of the methods are shown in Table 2. Profile regression is clearly outperforming Gaussian profile regression in terms of prediction accuracy for both simulated datasets. We also illustrate the performance of quantile profile regression with respect to standard ordinary least squares regression analysis and CART. Classification and regression tree methods (CART) are one of the most commonly used non-parametric methods that require no distributional assumptions. CART uses tree building methods, a form of binary recursive partitioning, and classifies subjects or predicts the outcome by selecting the most important risk factors available from the study population. Ordinary least square (OLS) regression is the standard approach for estimating associations in epidemiology. Predictive scoring rules could not be used for these comparisons because these methods do not produce probabilistic forecasts. The results for the RMSE are given in Table 2. We conclude that quantile profile regression outperforms Gaussian profile regression, CART and standard OLS regression significantly and consistently

when predicting in the tails of the distribution whether the data is symmetric (as for the dataset simulated from a mixture of Gaussian distributions) or asymmetric (as for the dataset simulated from a mixture of Gamma distributions).

Table 2: The top table reports the results of the RMSE and the CRPS over 100 repetitions for data simulated from a mixture of Gaussian distributions, while the bottom table reports the results of the RMSE and the CRPS over 100 repetitions for data simulated from a mixture of Gamma distributions. In each table, each row corresponds to one of the methods used for the predictions of observations marked by circles in Figure 4: quantile profile regression (with $p = 0.05$ and $p = 0.95$ respectively), Gaussian profile regression, OLS and CART. Note that the first two methods are stochastic, and therefore we quote the mean and standard deviation of 100 repetitions, while the latter two are deterministic.

Data: simulated mixture of Gaussian distributions				
	mean(RMSE)	sd(RMSE)	mean(CRPS)	sd(CRPS)
quantile 0.05	3.66	0.11	1.68	0.07
Normal	6.06	0.00	4.64	0.01
OLS	6.11			
CART	5.86			
Data: simulated mixture of Gamma distributions				
	mean(RMSE)	sd(RMSE)	mean(CRPS)	sd(CRPS)
quantile 0.95	8.77	0.35	6.20	0.04
Normal	13.77	0.26	12.86	0.02
OLS	14.07			
CART	14.22			

Source code to reproduce the results is available as Supporting Information on the journal's web page.

6 Applications to real datasets

6.1 Pima Diabetes Data Analysis

We present here an analysis of the Pima Diabetes dataset (Smith et al., 1988). This is a sample of native American women of the Pima heritage, aged 21 or over. We aim to investigate whether the women show signs of diabetes, which, according to the criteria of the World Health Organization, corresponds to checking whether the 2-hour post-load plasma glucose was at least 200 mg/dl at any survey examination. We removed missing values, remaining with data on 393 women. The other variables collected are highly correlated, so

we propose to use profile regression. These variables are: diastolic blood pressure ('pres'), triceps skin fold thickness ('skin'); 2-hour serum insulin ('insu'); body mass index ('mass') and diabetes pedigree ('pedi'). The summary statistics of all these variables are provided in Table 3. We also adjust for age by including it as a fixed effect. We will focus on the 95% quantile because we aim to determine the variables which may link to hyperglycemia.

Table 3: Summary statistics for the variables included in the analysis of the Pima dataset: minimum, first quartile, mean, third quartile, maximum and standard deviation.

	Min	Q_1	Mean	Q_3	Max
pres	24.00	62.00	70.67	78.00	110.00
skin	7.00	21.00	29.12	37.00	63.00
insu	14.00	76.00	155.72	190.00	846.00
mass	18.20	28.40	33.07	37.10	67.10
pedi	0.09	0.27	0.52	0.69	2.42
age	21.00	23.00	30.84	36.00	81.00

We carry out the analysis using quantile profile regression and Gaussian profile regression. The response submodel is given by $p(Y_i|\mathbf{\Theta}_{Z_i}, \mathbf{\Lambda}, \mathbf{W}_i) \equiv \text{ALD}(\theta_{Z_i} + \beta^T W_i, \sigma_Y; p)$ for quantile profile regression and by $p(Y_i|\mathbf{\Theta}_{Z_i}, \mathbf{\Lambda}, \mathbf{W}_i) \equiv \text{Normal}(\theta_{Z_i} + \beta^T W_i, \sigma_Y^2)$ for Gaussian profile regression. In both cases the covariate submodel is given by $p(X_i|\mathbf{\Theta}_{Z_i}, \mathbf{\Lambda}) \equiv \text{Normal}(\mu_{Z_i}, \tau_{Z_i}^2)$. We used the same priors as for the simulated data above and our results were not sensitive to the choice of prior. We ran 20,000 iterations of burn-in and 20,000 iterations after that. We obtain good convergence diagnostics on the trace, density and autocorrelation for various parameters. We obtained posterior distributions for all parameters and the clusters obtained provide an informative description of the dataset. The main results are included in Figure 5.

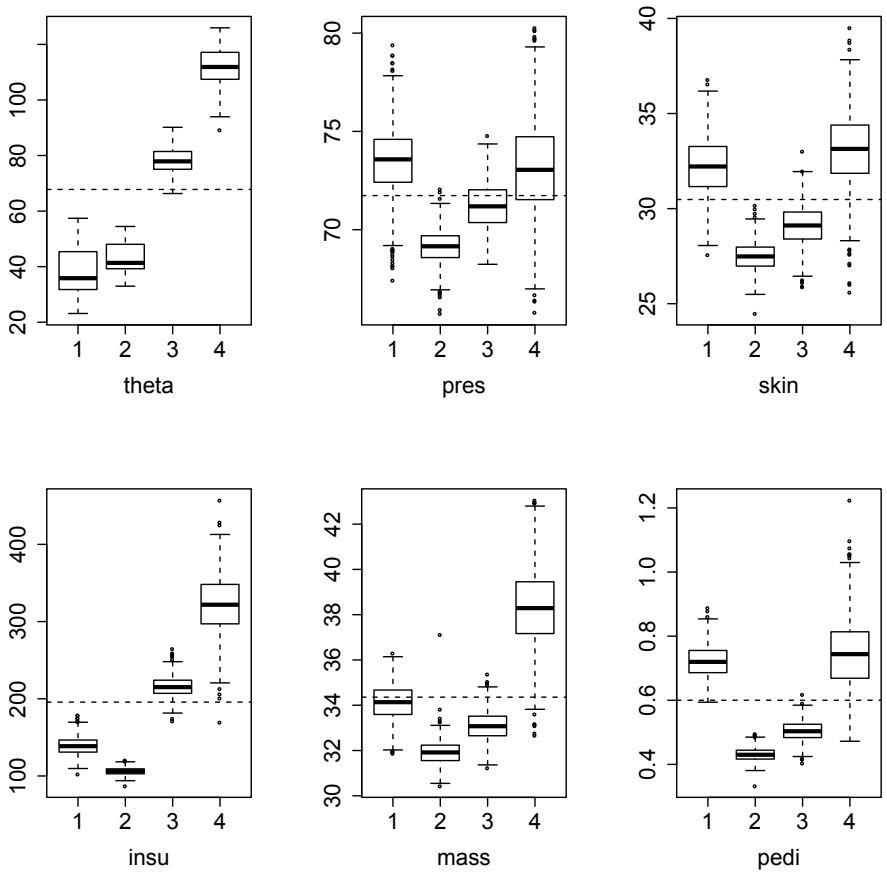

Figure 5: Boxplots for the posterior distributions of the parameter θ_c and the posterior distribution of the mean of each covariate for each cluster c.

We are interested in high values of Y because they correspond to hyperglycemia. We can show the higher accuracy of quantile profile regression over Gaussian profile regression when predicting values of Y for the observations such that $Y > 180$. The RMSE and the CRPS obtained comparing these predictions to the observed values are given in Table 4. As for simulated data, quantile profile regression proves to be more accurate when predicting values around the quantiles of interest.

Table 4: Mean and standard deviation of the RMSE and the CRPS over 100 repetitions of the predictions of the Y values such that $Y > 180$ applying quantile profile regression and Gaussian profile regression.

	RMSE	sd(RMSE)	CRPS	sd(CRPS)
quantile 0.95	4.01	1.27	3.11	0.97
Gaussian	4.30	0.07	3.89	0.15

Source code to reproduce the results is available as Supporting Information on the journal's web page.

6.2 Analysis of the English Longitudinal Study of Ageing (ELSA) dataset

We also conducted an analysis with data from the English Longitudinal Study of Ageing (ELSA). ELSA is a longitudinal cohort study of adults aged 50 or older which commenced in 1998, with data collection taking part every two years (Steptoe et al., 2012). The data used in our study are from the nurse visit conducted at Wave 2 of ELSA (2004-2005). A total of 7,666 people took part in this visit where biological data were collected for the first time. The data are available for download from the UK Data Service at `http://dx.doi.org/10.5255/UKDA-SN-5050-9`.

The aim of this study is to ascertain how cardiometabolic risk factors might be associated with high blood glucose in people who do not currently have a diagnosis of diabetes. Research has shown that high blood glucose levels are important predictors of incident diabetes (Tabák et al., 2012). However, it has been shown that only considering high blood glucose in prediction of diabetes risk may be overly simplistic as many other cardiometabolic risk factors that are highly correlated with high blood glucose (Haffner et al., 1990; Li et al., 2009) are also associated with diabetes risk (Ford, 2005; Kolberg et al., 2009). Thus we wanted to determine how cardiometabolic risk predictors may cluster together with an outcome of high blood glucose in people who do not have a current diagnosis of diabetes using quantile profile regression modelling. In theory we would expect to see higher levels of blood glucose associated with higher cardiometabolic risk factors.

We removed erroneous data (such as values outside their expected range) and missing values as we needed all relevant information to conduct our analysis. We also removed data for people with a current diagnosis of diabetes ($n = 333$), leaving data for 2,859 participants. We included the following variables as covariates: mean systolic blood pressure ('SYSVAL'), mean diastolic blood pressure ('DIAVAL'), mean arterial pressure ('MAPVAL'), cholesterol level ('CHOL'), high-density lipoprotein level ('HDL'), triglycerides level ('TRIG'), low-density lipoprotein level ('LDL'), C-reactive protein level (CRP: 'HSCRP'), mean waist ('WSTVAL'), mean waist/hip ratio ('WHVAL') and valid BMI ('BMIVAL'). See Table 5 for summary statistics of these variables.

Variables used in this analysis are highly correlated, as shown in Table 6, thus providing a strong rationale for the use of quantile profile regression. These variables are all quantitative and continuous and they will be included in our model as covariates. Within our model we adjusted for gender by including it as a fixed effect in the model. We opted to focus on the 95% quantile because we wanted to determine how cardiometabolic risk factors might link with high blood glucose. We are interested in high values of Y because they correspond to high blood glucose.

Table 5: Summary statistics for the variables included in the analysis of the ELSA dataset: minimum, first quartile, mean, third quartile, maximum and standard deviation.

	Min	Q_1	Mean	Q_3	Max	SD
SYSVAL	80.00	121.00	133.74	144.50	214.00	18.19
DIAVAL	36.50	69.50	76.32	83.00	118.00	10.41
MAPVAL	51.50	87.50	95.46	103.00	144.00	11.77
CHOL	2.10	5.30	6.04	6.80	12.30	1.15
HDL	0.50	1.30	1.57	1.80	3.40	0.38
TRIG	0.40	1.00	1.52	1.90	4.50	0.71
LDL	0.70	3.10	3.78	4.40	9.20	0.97
HSCRP	0.20	0.80	3.51	3.70	151.00	6.43
WSTVAL	61.25	85.20	94.15	102.32	171.60	12.65
WHVAL	0.64	0.82	0.88	0.95	1.26	0.08
BMIVAL	16.02	24.45	27.52	30.05	55.97	4.54

Table 6: Correlation matrix for the covariates.

	SYSVAL	DIAVAL	MAPVAL	CHOL	HDL	TRIG	LDL	HSCRP	WSTVAL	WHVAL	
SYSVAL	1										
DIAVAL	**0.64**	1									
MAPVAL	**0.89**	**0.92**	1								
CHOL	0.08	0.12	0.11	1							
HDL	0.00	0.02	0.02	**0.42**	1						
TRIG	0.10	0.12	0.12	0.25	-0.34	1					
LDL	0.06	0.09	0.08	**0.94**	0.22	0.10	1				
HSCRP	0.06	0.02	0.04	-0.04	-0.11	0.05	-0.02	1			
WSTVAL	0.17	0.20	0.21	-0.10	**-0.42**	0.31	-0.06	0.16	1		
WHVAL	0.16	0.14	0.17	-0.13	**-0.42**	0.27	-0.08	0.11	**0.78**	1	
BMIVAL	0.16	0.22	0.21	-0.01	-0.28	0.27	0.00	0.16	**0.79**	0.35	1

We carry out the analysis using quantile profile regression and Gaussian profile regression, with the same response and covariates models and priors above as for the simulated data. We ran 30,000 iterations of burn-in and 30,000 iterations after that. We obtain good convergence diagnostics on the trace, density and autocorrelation for various parameters.

Quantile profile regression identified four clusters of 1,432, 436, 760 and 231 observations respectively. Figure 6 shows boxplots of the posterior distribution for θ_c.

We are interested in high values of Y because they correspond to high blood glucose. When examining values of Y, values of blood glucose ≥ 5.6 mmol/L are considered as high risk for the development of diabetes (Centers for Disease Control and Prevention, 2011). The blood glucose levels of clusters two, three and four (with credible intervals) were all higher than 5.6 mmol/L indicating these three groups could be considered high risk for diabetes. The 95% credible interval for β, which quantifies the linear relationship between gender and the response, is $(-0.39, 0.10)$.

In assessing the cardiometabolic profile of each cluster we can look to Figure 7, showing boxplots of the posterior distributions of μ_c, so we can now examine in detail the relationship between the response variable and the covariates. Cluster 1 (lowest blood glucose levels) generally showed low cardiometabolic risk. Cluster 2 (who had moderately high levels of blood glucose) showed a relatively high risk profile for other cardiometabolic risk factors, with high levels of cholesterol, triglycerides, CRP and blood pressure with borderline/high anthropometric indicators. Cluster 3 (who showed a high blood glucose profile) had borderline blood pressure and anthropometric indicators but a high risk cholesterol and triglyceride profile. Cluster 4 (the highest blood glucose levels) showed very high levels of all cardiometabolic indicators. These results indicate that quantile profile regression modelling is able to discriminate between different levels of blood glucose based on the presence of cardiometabolic risk factors in a way that is theoretically sound (i.e., low blood glucose levels are associated with low cardiometabolic risk and high blood glucose levels are associated with higher cardiometabolic risk) while also being sensitive enough the reveal different groups that could be of clinical interest (e.g., cluster 4 indicated levels of extremely high inflammation through raised CRP which could be of interest to clinicians).

These data indicate that quantile profile regression could be a useful tool for identifying clusters of people based on shared cardiometabolic risk factors. As this analysis was cross sectional we cannot infer whether these clusters would predict incidence of type 2 diabetes, however this modelling tool shows promise for application in the context of illness risk.

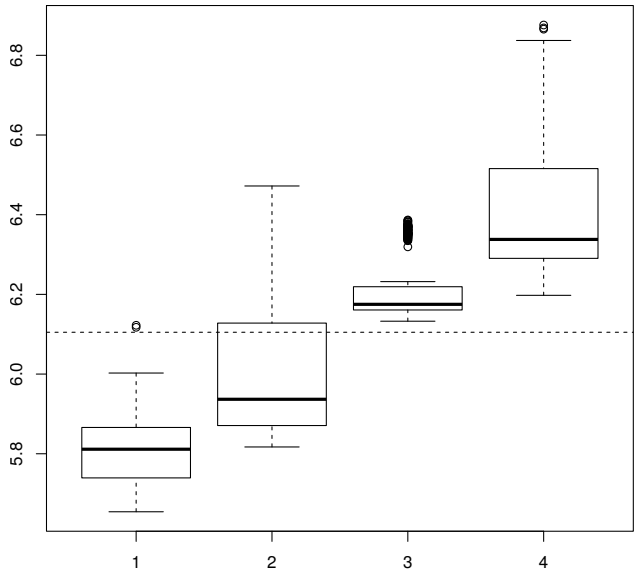

Figure 6: Boxplots of the posterior distribution of θ_c for the 4 clusters identified by quantile profile regression. The horizontal dashed line is the overall posterior mean.

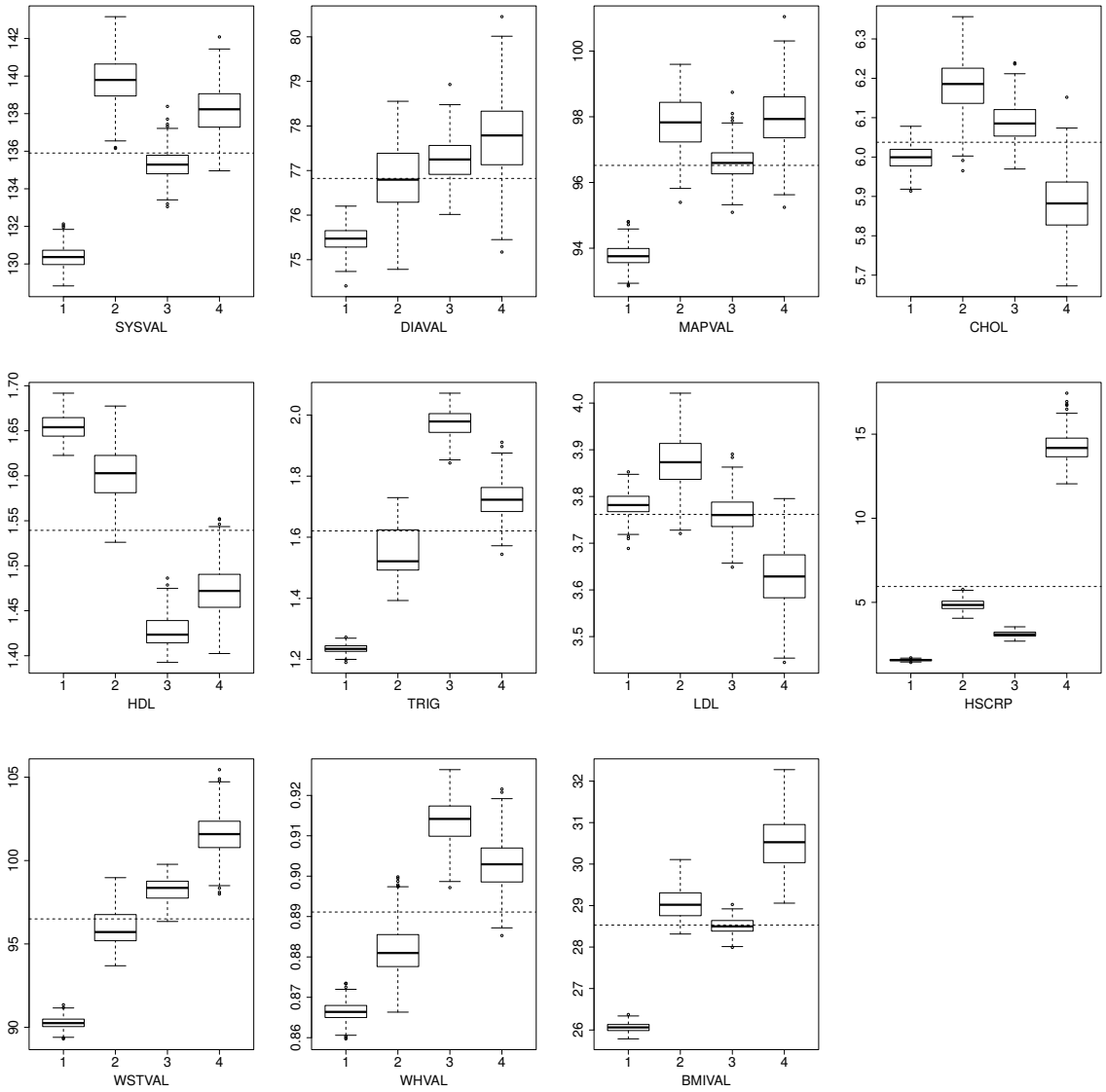

Figure 7: Boxplots of the posterior distribution of μ_c for the four clusters identified by quantile profile regression. The horizontal dashed line is the overall posterior mean for each covariate.

Finally, we can show the higher accuracy of quantile profile regression over Gaussian profile regression when predicting extreme values of Y. We do this by including the observations with $Y >= 6$ in a validation dataset. The RMSE and the CRPS obtained comparing these predictions to the observed values are given in Table 7. Quantile profile regression proves to be more accurate when predicting values around the quantiles of interest.

Table 7: Mean and standard deviation of the RMSE and the CRPS over 100 repetitions of the predictions of the Y values such that $Y >= 6$ applying quantile profile regression and Gaussian profile regression.

	RMSE	sd(RMSE)	CRPS	sd(CRPS)
quantile 0.95	25.56	3.75	16.86	2.50
Gaussian	47.20	0.11	40.61	0.21

7 Conclusions and future work

We have proposed a new method for collinear data which is robust when outliers are present and more accurate than existing methods when the modelling interest is in the tails of the distribution. The method is an extension of profile regression, a Bayesian clustering model, and it was applied to simulated and real data and it provided a significant increase in accuracy with considerable reduction in the residuals, especially under extreme quantiles, compared to an estimation with a Gaussian mixture model.

This method allows to explain the complex relationships between predictors and the response variables, as demonstrated in Sections 5 and 6. Profile regression is able to disentangle the complex relationships between predictors and response variables and can be used to evaluate how changes in the predictors might affect the response variable.

One limitation of the model proposed in its present form is that the asymmetric Laplace distribution is included for the response variable but not for the predictors, so it does not account for interest in the tails of the distribution of the covariates. This is the topic of future work.

The code is available from the authors upon request.

Acknowledgments

This work has been supported in part by the National Institute for Health Research Method Grant (NIHR-RMOFS-2013-03-09) and the National Natural Science Foundation of China (Grant No. 71490725, 11261048, 11371322). Conflict of Interest: None declared.

References

Coker, E., S. Liverani, J. K. Ghosh, M. Jerrett, B. Beckerman, A. Li, B. Ritz, and J. Molitor (2016). Multi-pollutant exposure profiles associated with term low birth weight in Los Angeles County. *Environment International 91*, 1–13.

Coker, E., S. Liverani, J. G. Su, and J. Molitor (2018). Multi-pollutant modeling through examination of susceptible subpopulations using profile regression. *Current Environmental Health Reports 5*(1), 59–69.

Davino, C., M. Furno, and D. Vistocco (2013). *Quantile regression: theory and applications*, Volume 988. John Wiley & Sons.

Dunson, D. B., A. B. Herring, and A. M. Siega-Riz (2008). Bayesian inference on changes in response densities over predictor clusters. *Journal of the American Statistical Association 103*(484), 1508–1517.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics 1*(2), 209–230.

Ford, E. S. (2005). Risks for all-cause mortality, cardiovascular disease, and diabetes associated with the metabolic syndrome. *Diabetes Care 28*(7), 1769–1778.

Franczak, B. C., R. P. Browne, and P. D. McNicholas (2014). Mixtures of shifted Asymmetric Laplace Distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 36*(6), 1149–1157.

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics 2*(4), 1360–1383.

Geraci, M. and M. Bottai (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics 8*(1), 140–154.

Haffner, S. M., M. P. Stern, H. P. Hazuda, B. D. Mitchell, and J. K. Patterson (1990). Cardiovascular risk factors in confirmed prediabetic individuals: does the clock for coronary heart disease start ticking before the onset of clinical diabetes? *Journal of the American Medical Association 263*(21), 2893–2898.

Hastie, D. I., S. Liverani, L. Azizi, S. Richardson, and I. Stücker (2013). A semi-parametric approach to estimate risk functions associated with multi-dimensional exposure profiles: application to smoking and lung cancer. *BMC Medical Research Methodology 13*(1), 129.

Hastie, D. I., S. Liverani, and S. Richardson (2015). Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and Computing 25*(5), 1023–1037.

Hu, Y., R. B. Gramacy, and H. Lian (2013). Bayesian quantile regression for single-index models. *Statistics and Computing 23*(4), 437–454.

Jara, A., T. E. Hanson, F. A. Quintana, P. Müller, and G. L. Rosner (2011). DPpackage: Bayesian semi-and nonparametric modeling in R. *Journal of Statistical Software 40*(5), 1.

Jordan, A., F. Krüger, and S. Lerch (2018). Evaluating probabilistic forecasts with the R package scoringRules. *arXiv preprint arXiv:1709.04743*.

Kalli, M., J. E. Griffin, and S. G. Walker (2011). Slice sampling mixture models. *Statistics and Computing 21*(1), 93–105.

Knight, C. A. and D. D. Ackerly (2002). Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. *Ecology Letters 5*(1), 66–76.

Koenker, R. and G. J. Bassett (1978). Regression quantiles. *Econometrica 46*(1), pp. 33–50.

Kolberg, J. A., T. Jørgensen, R. W. Gerwien, S. Hamren, M. P. McKenna, E. Moler, M. W. Rowe, M. S. Urdea, X. M. Xu, T. Hansen, et al. (2009). Development of a type 2 diabetes risk model from a panel of serum biomarkers from the inter99 cohort. *Diabetes care 32*(7), 1207–1212.

Kottas, A. and M. Kranjajić (2009). Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics 36*(2), 297–319.

Li, C., E. S. Ford, G. Zhao, and A. H. Mokdad (2009). Prevalence of pre-diabetes and its association with clustering of cardiometabolic risk factors and hyperinsulinemia among us adolescents. *Diabetes care 32*(2), 342–347.

Liverani, S., D. I. Hastie, L. Azizi, M. Papathomas, and S. Richardson (2015). PReMiuM: An R package for Profile Regression Mixture Models using Dirichlet Processes. *Journal of Statistical Software 64*(7), 1–30.

Liverani, S., A. Lavigne, and M. Blangiardo (2016). Modelling collinear and spatially correlated data. *Spatial and Spatio-temporal Epidemiology 18*, 63–73.

Mattei, F., S. Liverani, F. Guida, M. Matrat, S. Cenée, L. Azizi, G. Menvielle, M. Sanchez, C. Pilorget, B. Lapôtre-Ledoux, et al. (2016). Multidimensional analysis of the effect of occupational exposure to organic solvents on lung cancer risk: the icare study. *Occupational and environmental medicine 73*(6), 368–377.

Molitor, J., I. J. Brown, Q. Chan, M. Papathomas, S. Liverani, N. Molitor, S. Richardson, L. Van Horn, M. L. Daviglus, A. Dyer, J. Stamler, P. Elliott, and I. R. Group (2014). Blood pressure differences associated With optimal macronutrient intake trial for heart health (OMNIHEART)–like diet compared with a typical American diet. *Hypertension 64*(6), 1198–1204.

Molitor, J., M. Papathomas, M. Jerrett, and S. Richardson (2010). Bayesian profile regression with an application to the National Survey of Children's Health. *Biostatistics 11*(3), 484–498.

Papaspiliopoulos, O. (2008). A note on posterior sampling from Dirichlet mixture models. Technical Report 8, CRISM Paper.

Papathomas, M., J. Molitor, C. Hoggart, D. Hastie, and S. Richardson (2012). Exploring data from genetic association studies using Bayesian variable selection and the Dirichlet process: application to searching for gene× gene patterns. *Genetic Epidemiology 36*(6), 663–674.

Pirani, M., N. Best, M. Blangiardo, S. Liverani, R. W. Atkinson, and G. W. Fuller (2015). Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. *Environment International 79*, 56–64.

Smith, J. W., J. Everhart, W. Dickson, W. Knowler, and R. Johannes (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus. *Johns Hopkins APL Technical Digest 10*, 262–266.

Sriram, K., R. V. Ramamoorthi, and G. Pulak (2013). Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Bayesian Analysis 8*(2), 479–504.

Steptoe, A., E. Breeze, J. Banks, and J. Nazroo (2012). Cohort profile: the English longitudinal study of ageing. *International Journal of Epidemiology 42*(6), 1640–1648.

Tabák, A. G., C. Herder, W. Rathmann, E. J. Brunner, and M. Kivimäki (2012). Prediabetes: a high-risk state for diabetes development. *The Lancet 379*(9833), 2279–2290.

Yu, K. and R. A. Moyeed (2001). Bayesian quantile regression. *Statistics & Probability Letters 54*(4), 437–447.

Yu, K. and J. Stander (2007). Bayesian analysis of a Tobit quantile regression model. *Journal of Econometrics 137*(1), 260–276.

Yu, K. and J. Zhang (2005). A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics - Theory and Methods 34*(9-10), 1867–1879.