Moment-based Estimation of Mixtures of Regression Models

Ekstrøm, Claus Thorn; Pipper, Christian Bressen

Published in:
arXiv

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Moment-based Estimation of Mixtures of Regression Models

Claus Thorn Ekstrøm[1] and Christian Bressen Pipper[1]

[1]Section of Biostatistics, Department of Public Health,, University of Copenhagen, Øster Farimagsgade 5 B, DK-1014 Copenhagen K, Denmark

May 17, 2019

Abstract

Finite mixtures of regression models provide a flexible modeling framework for many phenomena. Using moment-based estimation of the regression parameters, we develop unbiased estimators with a minimum of assumptions on the mixture components. In particular, only the average regression model for one of the components in the mixture model is needed and no requirements on the distributions. The consistency and asymptotic distribution of the estimators is derived and the proposed method is validated through a series of simulation studies and is shown to be highly accurate. We illustrate the use of the moment-based mixture of regression models with an application to wine quality data. mixture regression model; moment estimation; estimating equations; robust inference

1 Introduction

The class of finite mixtures of regression models provides a flexible approach to model a wide range of phenomena and to handle non-standard data analysis problems such as excess zeros or heterogeneity (McLachlan and Peel, 2000). Mixture models can also be used to accommodate data that is ”contaminated” due to poor quality data, laboratory errors, or situations where data originates from multiple sources. In particular, zero-inflated regression models, and hurdle models can be considered special cases of the class of finite mixture of regression models with two components and both of these types of models see frequent use (Lambert, 1992; Kwagyan and Apprey, 2016).

Finite mixture regression models are relevant for analyzing many problems such as sudden-infant-death-syndrome (Dalrymple *and others*, 2003), HIV-risk reduction trials (Hu *and others*, 2011), medical care (Deb and Trivedi, 1997). However, problems relevant for finite mixtures are especially common in genomics where for example population admixture (certain subgroups of the population do not segregate a phenotype-influencing mutation), or in the presence of gene-environment interactions, where the effect of some genes are never triggered because the individual is living under specific environmental conditions.

Other recent approaches of using finite mixture models in genetics include the paper by Xu *and others* (2015) where they compare different model types (standard parametric, non-parametric, zero-inflated and hurdle models) to microbiome data in order to assess and infer the best model for these types of data. They find that mixture models in general fit the data best but that the choice of parametric model (Poisson or negative-binomial) can have substantial impact on the results and on the convergence on the estimation algorithm.

Typically, parameters in mixture regression models are estimated by specifying a fully parametric model for each of the components. This leads to efficient estimates provided that the parametric models are specified correctly. In this article we will consider the class of finite mixture regressions where the predictors are influencing exactly one of the regression model components. Within this framework we wish to estimate the association between the predictors and the outcome. We place no restriction on the distribution of the components but only require that the relationship between the predictors and the outcome can be modeled as a linear model. Moment-based estimators allow us to obtain unbiased estimates of the regression parameters with a minimum of assumptions. In particular, we do not require a specification of the distribution, and we only need to specify the average regression model for one of the components in the mixture model.

The recent paper by Kong *and others* (2017) uses an approach for identifying interaction that most closely resembles the approach in this manuscript. They also model the first and second moments but they are focused on interaction in particular and do not use a moment-based method for estimation.

The manuscript is organized as follows: The next section explains the problem, derives the moment-based estimators, and proves the large-sample properties relevant for inference. In section 3, we present properties of the proposed method for moment-based mixture of regression models and use simulations to compare the proposed moment-based mixture regression model to the analogous Gaussian mixture model. Finally, we apply the moment-based mixture of regression models to a dataset of wine quality before we discuss the findings along with possible extensions. The approaches presented in this manuscript are available in the R package `mommix` which can be found on github at `www.github.com/ekstroem/mommix`.

2 Methods

Let Y be given as a mixture of two distributions, that is;

$$Y = PY_1 + (1-P)Y_2,$$

where $\Pr(P=1) = 1 - \Pr(P=0) = p$ for some $p \in (0,1)$ and Y_1 and Y_2 are independent stochastic variables from the two underlying distributions. We are interested in modeling the relationship between a set of m predictors represented by the design matrix X and Y_1 and can think of Y_2 as a contamination of the response.

For the distribution of interest we assume that $E(Y_1|X) = \mu_1 + \beta^T X$, and $V(Y_1|X) = \sigma_1^2$, but for the other component we only require that $Y_2 \perp X$ and that it has well-defined mean and variance $E(Y_2) = \mu_2$, and $V(Y_2) = \sigma_2^2$.

The mean of Y is

$$E(Y|X) = p \cdot (\mu_1 + \beta^T X) + (1-p) \cdot \mu_2, \tag{1}$$

where $\beta \in \mathbb{R}^m$ is the parameter vector of interest. Simple algebra yields that the second order moment is

$$E(Y^2|X) = p \cdot ((\mu_1 + \beta^T X)^2 + \sigma_1^2) + (1-p) \cdot (\mu_2^2 + \sigma_2^2). \tag{2}$$

Note that the formula for the second order moment has a form corresponding to a multiple linear regression model with a mean that can be reparameterized as $2 \cdot \mu_1 \cdot p \cdot (\beta^T X) + p \cdot (\beta^T X)^2 + \tilde{\alpha}$. Thus, if X is univariate we can obtain estimates of $2 \cdot \mu_1 \cdot p \cdot \beta$ and $p \cdot \beta^2$ directly from least-squares regression of the second order moment and estimate $p \cdot \beta$ from least-squares regression of the first moment. These estimates can be combined to extract individual estimates of μ_1, β, and p. In the following we formalize and extend this idea.

Let $\lambda_1 = p \cdot \beta$ and set $\eta_i = \lambda_1^T X_i$ (the linear predictor), $\lambda_2 = 2 \cdot \mu_1$, and $\lambda_3 = p^{-1}$. Then equation (2) may be rewritten as

$$E(Y_i^2|X_i) = \tilde{\alpha} + \lambda_2 \cdot \eta_i + \lambda_3 \cdot \eta_i^2,$$

where the intercept $\tilde{\alpha}$ contains contributions from both Y_1 and Y_2 but holds no direct information about X.

From this parameterization we have that $\beta = \lambda_1 \cdot \lambda_3$, $\mu_1 = \lambda_2/2$, and $p = \frac{1}{\lambda_3}$. Thus, if we can estimate λ_1, λ_2, and λ_3 then we can also estimate the mixture proportion and all mean parameters related to the distribution of interest.

2.1 Inference and large sample properties

Note that $\lambda_1 = p \cdot \beta$ corresponds to the regression parameter vector when we regress Y on X. Consequently by least squares regression we can obtain a closed form estimator $\hat{\lambda}_1$ of λ_1. Specifically with $X_i \in \mathbb{R}^m$ and $Y_i \in \mathbb{R}$ denoting the data from the ith individual and with $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_n)^T$, $\tilde{X}_i = (1, X_i^T)^T$, and $Y = (Y_1, \ldots, Y_n)$ we have:

$$\hat{\lambda}_1 = A(m)(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T Y, \tag{3}$$

where $A(m)$ is a $m \times (m+1)$ matrix projecting an $m+1$ vector to its last m components.

The residual $Y_i - E(Y_i|\tilde{X}_i)$ is not ensured to follow a Gaussian distribution — even if the original distributions of Y_1 and Y_2 were Gaussian — but we may still resort to a characterization of $\sqrt{n}(\hat{\lambda}_1 - \lambda_1)$ as a sum of i.i.d. zero mean terms ensuring consistency and asymptotic normality by means of the central limit theorem. Specifically the characterization is as follows:

$$\sqrt{n}(\hat{\lambda}_1 - \lambda_1) = A(m)(\frac{1}{n}\tilde{X}^T \tilde{X})^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{X}_i \{Y_i - E(Y_i|\tilde{X}_i)\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i + o_P(1) \tag{4}$$

with

$$\varepsilon_i = A(m) E(\tilde{X}_i \tilde{X}_i^T)^{-1} \tilde{X}_i \{Y_i - E(Y_i|\tilde{X}_i)\}.$$

Recall that equation (2) may be rewritten as

$$E(Y_i^2|X_i) = \tilde{\alpha} + \lambda_2 \cdot \eta_i + \lambda_3 \cdot \eta_i^2$$

with $\eta_i = \lambda_1^T X_i$ (the linear predictor), $\lambda_2 = 2 \cdot \mu_1$, and $\lambda_3 = p^{-1}$. From this observation and the fact that we have an estimator of λ_1 it seems only natural to estimate λ_2 and λ_3 by a multiple linear regression of Y_i^2 on $\hat{\eta}_i = \hat{\lambda}_1^T X_i$. However, since the variance of Y_i^2 given X_i depends on η_i and η_i^2 it is more appropriate to estimate λ_2 and λ_3 using a weighted regression, where the weight is a function of η_i.

Specifically let w be a weight function, put $w_i = w(\eta_i)$, $\hat{w}_i = w(\hat{\eta}_i)$, and let $\mathbb{P}_n^w$ denote the weighted empirical mean, that is for $f_i = f(Y_i, X_i)$

$$\mathbb{P}_n^w(f) = \frac{\sum_{i=1}^n f_i w_i}{\sum_{i=1}^n w_i}.$$

Also define

$$\begin{aligned}
a_1^n(\lambda_1) &= \mathbb{P}_n^w(\eta^4) - \mathbb{P}_n^w(\eta^2)^2,\\
a_2^n(\lambda_1) &= \mathbb{P}_n^w(\eta^3) - \mathbb{P}_n^w(\eta)\mathbb{P}_n^w(\eta^2),\\
a_3^n(\lambda_1) &= \mathbb{P}_n^w(\eta^2) - \mathbb{P}_n^w(\eta)^2,\\
b_1^n(\lambda_1) &= \mathbb{P}_n^w(\eta Y^2) - \mathbb{P}_n^w(\eta)\mathbb{P}_n^w(Y^2),\\
b_2^n(\lambda_1) &= \mathbb{P}_n^w(\eta^2 Y^2) - \mathbb{P}_n^w(\eta^2)\mathbb{P}_n^w(Y^2).
\end{aligned}$$

Finally define

$$\begin{aligned}
\lambda_2^n(\lambda_1) &= \frac{a_1^n(\lambda_1)\cdot b_1^n(\lambda_1) - a_2^n(\lambda_1)\cdot b_2^n(\lambda_1)}{a_1^n(\lambda_1)\cdot a_3^n(\lambda_1) - a_2^n(\lambda_1)^2},\\
\lambda_3^n(\lambda_1) &= \frac{a_3^n(\lambda_1)\cdot b_2^n(\lambda_1) - a_2^n(\lambda_1)\cdot b_1^n(\lambda_1)}{a_1^n(\lambda_1)\cdot a_3^n(\lambda_1) - a_2^n(\lambda_1)^2}.
\end{aligned}$$

Then our estimators are given as

$$\hat{\lambda}_2 = \lambda_2^n(\hat{\lambda}_1), \tag{5}$$

$$\hat{\lambda}_3 = \lambda_3^n(\hat{\lambda}_1). \tag{6}$$

We now proceed to derive large sample properties of the estimators. In what follows we adopt the notation $\mathbf{P}f = Ef(X_i, Y_i)$ for any function of f of our data samples (X_i, Y_i)

To derive large sample results we need to assume that the weight function w is twice continuously differentiable and that $\lambda_1 \to w(\lambda_1^T X)$ and the derivatives $\lambda_1 \to D_{\lambda_1} w(\lambda_1^T X)$ and $\lambda_1 \to D_{\lambda_1}^2 w(\lambda_1^T X)$ are bounded by functions with finite means in some open neighbourhood of the true value of λ_1.

We also need the following lemmas

Lemma 2.1. *Let $f : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be a two times continuously differentiable function in the first coordinate. Assume that $\lambda_1 \to f(\lambda_1^T X, Y)\cdot w(\lambda_1^T X)$ and the derivatives $\lambda_1 \to D_{\lambda_1}\{f(\lambda_1^T X, Y)\cdot w(\lambda_1^T X)\}$ and $\lambda_1 \to D_{\lambda_1}^2\{f(\lambda_1^T X, Y)\cdot w(\lambda_1^T X)\}$ are bounded by functions with finite mean in an open neighbourhood of the true value of λ_1. Also put $f_i = f(\eta_i, Y_i)$ and $\hat{f}_i = f(\hat{\eta}_i, Y_i)$. Then with*

$$g(\lambda_1) = \frac{\mathbf{P}\{f\cdot w\}}{\mathbf{P}w}$$

it holds that g is differentiable with derivative

$$D_{\lambda_1} g(\lambda_1) = (\mathbf{P}w)^{-1} P\{D_{\lambda_1}(f\cdot w)\} - (\mathbf{P}w)^{-2}\mathbf{P}\{f\cdot w\}\mathbf{P}(D_{\lambda_1} w).$$

Moreover it holds that

$$\sqrt{n}\{\mathbb{P}_n^{\hat{w}}(\hat{f}) - \mathbb{P}_n^{w}(f)\} = \sqrt{n}\{\hat{\lambda}_1 - \lambda_1\}^T D_{\lambda_1} g(\lambda_1) + o_P(1).$$

Lemma 2.2. *Let $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and put $f_i = f(\eta_i, Y_i)$. Assume that*

$$P\{f \cdot w\}^2 < \infty.$$

Then

$$\sqrt{n}\{\mathbb{P}_n^{w}(f) - \frac{\mathbf{P}\{f \cdot w\}}{\mathbf{P}w}\} = \{\mathbf{P}w\}^{-1}\mathbb{G}_n(\{f - \frac{\mathbf{P}(f \cdot w)}{\mathbf{P}w}\} \cdot w) + o_P(1),$$

where $\mathbb{G}_n(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (f_i - \mathbf{P}f)$.

Theorem 2.1. *The estimators $\hat{\lambda}_2$ and $\hat{\lambda}_3$ are consistent, asymptotically normal, and have the following asymptotic characterizations:*

$$\sqrt{n}(\hat{\lambda}_2 - \lambda_2) = \sqrt{n}\{\hat{\lambda}_1 - \lambda_1\}^T D_{\lambda_1}\lambda_2(\lambda_1) + \{g_3(\lambda_1)\mathbf{P}w\}^{-1}\mathbf{G_n}(w\{a_1(\lambda_1)f^1 - a_2(\lambda_1)f^2\}) + o_P(1),$$

$$\sqrt{n}(\hat{\lambda}_3 - \lambda_3) = \sqrt{n}\{\hat{\lambda}_1 - \lambda_1\}^T D_{\lambda_1}\lambda_3(\lambda_1) + \{g_3(\lambda_1)\mathbf{P}w\}^{-1}\mathbf{G_n}(w\{a_3(\lambda_1)f^2 - a_2(\lambda_1)f^1\}) + o_P(1),$$

where $a_1(\lambda_1)$, $a_2(\lambda_1)$, $a_3(\lambda_1)$, $g_3(\lambda_1)$, $\lambda_2(\lambda_1)$, and $\lambda_3(\lambda_1)$ are the limits in probability of $a_1^n(\lambda_1)$, $a_2^n(\lambda_1)$, $a_3^n(\lambda_1)$, $g_3^n(\lambda_1)$ $\lambda_2^n(\lambda_1)$, and $\lambda_3^n(\lambda_1)$, respectively, and

$$f_i^1 = (\eta_i - \frac{\mathbf{P}(\eta \cdot w)}{\mathbf{P}w})\{Y_i^2 - E(Y_i^2|X_i)\},$$
$$f_i^2 = (\eta_i^2 - \frac{\mathbf{P}(\eta^2 \cdot w)}{\mathbf{P}w})\{Y_i^2 - E(Y_i^2|X_i)\}.$$

As an immediate consequence of (4) and Theorem 1 we have the that the obvious estimator $\hat{\beta} = \hat{\lambda}_3\hat{\lambda}_1$ of β is both consistent and asymptotically normal. In particular, we have the following iid decomposition of $\sqrt{n}(\hat{\beta} - \beta)$:

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}}\sum_{i}^{n}\xi_i + o_P(1), \tag{7}$$

where

$$\xi_i = [\lambda_3 \cdot I_m + \lambda_1\{D_{\lambda_1}\lambda_3(\lambda_1)\}^T]\varepsilon_i + \lambda_1 \cdot \{g_3(\lambda_1\mathbf{P}w\}^{-1} \cdot w_i \cdot \{a_1(\lambda_1)f_i^1 - a_2(\lambda_1)f_i^2\}.$$

Notice that the iid decomposition above enables consistent estimation of standard errors based on the law of large numbers. Specifically

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i\xi_i^T \xrightarrow{P} \mathrm{var}(\xi_i) \tag{8}$$

and accordingly a consistent estimator of $\text{var}(\xi_i)$ can be obtained by inserting the empirical counterpart of ξ_i in the left hand side of (8).

2.2 Choice of weights

The optimal weights for weighted least squares regression are $w_i \propto \frac{1}{V(Y|X)}$. For the present we suggest using

$$v_i = v(\eta_i) = \frac{1}{(1+\eta_i^2)}$$

as weights for the first order moment regression, (1), and

$$w_i = w(\eta_i) = \frac{1}{(1+\eta_i^4)}$$

as weights for the second order moment regression (2). Both choices mimic the order of conditional variances as functions of η_i and fulfills the requirements that the weight functions are twice continuously differentiable and locally bounded by functions with finite means. In the rest of the following we will use these weights.

3 Simulation study

We illustrate the empirical properties of the proposed estimation procedure through a series of simulations. Consider the following four scenarios:

1. Simple Gaussian mixture: $Y_1 \sim N(1+X, 1^2), Y_2 \sim N(0, 1^2)$ and $X \sim N(0, 1^2)$.
2. Gaussian distribution with zero-inflation: $Y_1 \sim N(1+X, 1^2), Y_2 = 0$ and $X \sim N(0, 1^2)$.
3. Exponential-Gaussian mixture: $Y_1 \sim \exp(1) + X$ (shifted exponential distribution with rate 1), $Y_2 \sim N(0, 0.5^2)$ and $X \sim N(0, 1^2)$.
4. Exponential distribution with zero-inflation: $Y_1 \sim \exp(1) + X$, $Y_2 = 0$ and $X \sim N(0, 1^2)$.

For each scenario we simulate 100 datasets each with $N \in \{300, 500, 800, 100, 1500, 2000, 3000\}$ observations and with a constant mixture proportion of $p = 0.7$ (i.e., 30% contamination). The estimates $\hat{\beta}$ and $\hat{p}$ are computed using both the moment mixture estimation procedure as well as a standard two-component Gaussian regression mixture model where X can only influence one of the components.

Figure 1 shows the empirical estimates of $\hat{\beta}$ (and the corresponding 95% pointwise standard errors) for the four different scenarios while Figure 1 shows the corresponding plot for the mixing proportion $\hat{p}$. Not surprisingly, the Gaussian regression mixture model provides an unbiased estimate of both $\hat{\beta}$ and $\hat{p}$ with the smallest variance when the data follows a mixture of Gaussian distributions. However, it is also clear that the moment mixture model — which places fewer assumptions on the distributions involved — is also unbiased and has a variance that is only slightly larger than the Gaussian mixture model. Substantial differences are seen when the distribution of Y_1 is *not* Gaussian where the moment mixture model quickly converges to the true value while the Gaussian mixture model results in biased estimates of both β and p. The estimates from the moment mixture model are unbiased even when the distributions are non-normal, but Figure 2 suggests that large samples may be needed in order to have sufficient power to use the moment mixture model to infer whether there are two or only one component.

The variance of the mixing proportion estimate, $\text{var}(\hat{p})$, is larger for the moment mixture model than for the Gaussian regression mixture model which essentially is the "price that is paid" by having fewer assumptions about the distribution and estimating the parameters from the moments of the distribution.

Table 1 shows results from the same four scenarios (although with mixing proportion $p = 0.5$) to illustrate the precision of the estimates and to compare the estimated standard errors from (8) to the empirical standard errors. The estimates for both β and p are close to the true values even for smaller sample sizes and when the error distribution is non-normal. Table 1 also shows that the estimated standard errors are close to the empirical standard errors which suggests that the asymptotic estimate in (8) provides a useful measure of the standard errors of the regression and mixing proportion parameters even for smaller sample sizes. The 95% coverage probabilities for β are close to the true value even for $N = 300$ regardless of the underlying distributions while the coverage probabilities for p are quite unstable for smaller sample sizes while they achieve right right level for $N = 2000$. Note that the mixing proportion is 50% so there is a substantial amount of noise in the data.

To investigate the efficiency of the proposed estimates we consider the case where we know the individual allocations, P_i. If we make no assumptions on the distribution of Y_2 but assume that Y_1 follows a normal distribution with mean $\beta_0 + \beta_1 X$ and variance σ_1^2 conditional on $X \in \mathbb{R}$, then the most efficient way to estimate β_1 would be by simple linear regression based only on the pairs (X_i, Y_i) where $P_i = 1$. For this estimator the asymptotic variance is equal to

$$\text{Var}(\hat{\beta}) = \frac{\sigma_1^2}{(p \cdot N - 1) \cdot \text{Var}(X_i)} \tag{9}$$

Accordingly, the maximum likelihood estimator $\hat{\beta}_1$ of β_1 from any parametric mixture model of Y with the above specification of Y_1 will have an asymptotic variance that is at least (9). We compare the estimated standard error from the moment mixture model to this lower bound to see the efficiency loss incurred by removing the assumptions about the parametric distributions. The result is seen in the right-most column of Table 1, and we see that the efficiency loss from using the moment mixture model is around a factor 4-5.

4 Application: The effect of pH on wine volatile acidity

Volatile acidity (VA) refers to the process when lactic acid bacteria and acetic acids turns wine into vinegar, and the process takes place mainly due to growth of bacteria, the oxidation of ethanol, or the metabolism of acids/sugars. Wines with a high level of pH are supposedly more susceptible to oxidation and the antibacterial effects of sulfur dioxide and of fumaric acid are reduced rapidly as the pH level increases. Consequently, wines are thought to lose their quality as they become less acidic (increased pH) since the volatile acidity increases.

The paper by Cortez *and others* (2009) considers 11 physicochemical properties of a selected sample of Portuguese *vinho verde* wines. Samples from 1599 red wines and 4898 white wines are available and the relationship between volatile acidity and pH are shown separately for red and white wines in Figure 3. While it is apparent that the available red wines generally have slightly higher levels of pH, it also appears as if the impact of pH on volatile acidity is largest for the red wines in the sample: Individual regression lines for the two types of wine show an almost horizontal line for white wine while there is an effect of pH on VA for the red wines (slopes -0.022 and 0.272, respectively, and the slopes are significantly different, $p < 0.0001$).

If we did not know that the full 6497 samples were comprised of two different types of wine we might pursue

Table 1: Results from simulation study of the moment mixture model based on 1000 simulations. The model provides unbiased estimates and the estimted SE match the empirical SEs. 95% CP is the 95% coverage probability and the efficiency is the relative efficiency of the SE for the moment mixture model relative to the SE of the optimal model.

	N	True value	Estimate	Emp. SE	Est. SE	95% CP	Rel. efficiency
Gaussian mixture							
β	300	1.0	1.004	0.331	0.323	0.939	3.96
p	300	0.5	0.481	0.197	0.174	0.850	
β	2000	1.0	1.002	0.120	0.121	0.948	3.83
p	2000	0.5	0.497	0.066	0.065	0.948	
Zero-inflated Gaussian							
β	300	1.0	0.977	0.315	0.307	0.943	3.76
p	300	0.5	0.506	0.181	0.156	0.870	
β	2000	1.0	0.995	0.117	0.118	0.956	3.73
p	2000	0.5	0.502	0.059	0.056	0.957	
Exponential-Gaussian mixture							
β	300	1.0	0.998	0.441	0.394	0.929	4.83
p	300	0.5	0.476	0.231	0.204	0.712	
β	2000	1.0	1.006	0.153	0.154	0.953	4.87
p	2000	0.5	0.495	0.085	0.080	0.956	
Zero-inflated exponential							
β	300	1.0	0.996	0.404	0.375	0.942	4.59
p	300	0.5	0.490	0.212	0.185	0.773	
β	2000	1.0	0.999	0.154	0.151	0.946	4.78
p	2000	0.5	0.500	0.080	0.073	0.947	

regressing volatile acidity on pH for the full data. This gives the dashed regression shown in Figure 3 (slope $\hat{\beta}$ =0.27, 95% CI 0.24-0.29), which suggests an overall effect of pH on volatile acidity.

When we fit the moment-based mixture model then we see that the data are likely to consist of a mixture, $\hat{p} = 0.28$ (95% CI: 0.24-0.33), which suggests that only a proportion of the wines are influenced by pH. Also, the effect driving the association between the volatile acidity and pH appears to be much stronger than what was observed from analyzing the full data with a simple linear regression model, $\hat{\beta} = 0.94$ (95% CI 0.51-1.37). Consequently, the moment mixture model is able to identify that the wine data is likely to consist of two types of wines that respond differently to changes in levels of pH with only placing very few restrictions on the underlying distributions.

We estimate the mixing proportion to be 28% whereas the dataset contains 24.6% red wines. While the wine colour classification need not correspond to the separation we estimate from the moment mixture model it may indeed be the case that the different wine types contain a set of features that influence the impact of pH which is what we observe for these data.

5 Discussion

In this article we address the problem of estimating regression parameters for a two-component mixture regression model where one component is influenced by a set of predictors. Our proposed method imposes no assumptions on the distributions of the components and only a minimum of restrictions on the regression effects. The moment-based mixture of regression models estimators can be used to detect or account for unspecified mixtures in regression problems, and since estimation is fast it could be used for large-scale studies such as in genome-wise association studies.

As shown by the simulation study, the price for the flexibility and lack of assumptions comes with a larger variance of the estimators but the variance is not prohibitively larger and the proposed methods provides consistent estimates even when the mixture distributions are highly skewed and other models such as the Gaussian mixture model fails.

In conclusion, we have introduced an estimation technique for mixtures of regression models that can be applied to a large number of situations. The moment-based estimator is very versatile: it can be used not only to estimate the regression parameters, but for larger datasets it provides a foundation for detecting the number of mixture components. If $\hat{p}$ is different from both 0 and 1 then this suggests that there indeed are two components where only one of them is influenced by the predictors of interest.

6 Software

Software in the form of the `mommix` R package and code, together with a sample input data set and complete documentation is available on GitHub at `www.github.com/ekstroem/mommix`.

Conflict of Interest: None declared.

References

Cortez, Paulo, Cerdeira, Antonio, Almeida, Fernando, Matos, Telmo and Reis, Jose. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* **47**, 557–53.

Dalrymple, M.L., Hudson, I.L. and Ford, R.P.K. (2003). Finite mixture, zero-inflated poisson and hurdle models with application to sids. *Computational Statistics & Data Analysis* **41**(3), 491 – 504. Recent Developments in Mixture Model.

Deb, Partha and Trivedi, Pravin K. (1997). Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics* **12**, 313–336.

Hu, Mei-Chen, Pavlicova, Martina and Nunes, Edward V. (2011). Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial. *Am J Drug Alcohol Abuse* **37**, 367–375.

Kong, Yinfei, Li, Daoji, Fan, Yingying and Lv, Jinchi. (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *Annals of Statistics* **45**(2), 897–922.

Kwagyan, John and Apprey, Victor. (2016). Modeling clustered binary data with excess zero clusters. *Statistical Methods in Medical Research* **0**, 1–16.

Lambert, D. (1992). Zero-inflated poisson regression with an application to defects in manufacturing. *Technometrics* **34**, 1–14.

McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*. New York: Wiley Series in Probability and Statistics.

Xu, Lizhen, Paterson, Andrew D., Turpin, Williams and Xu, Wei. (2015, 07). Assessment and selection of competing models for zero-inflated microbiome data. *PLOS ONE* **10**(7), 1–30.

7 Appendix

7.1 Proof of Theorem 2.1

Denote

$$g_1^n(\lambda_1) = a_1^n(\lambda_1) \cdot b_1^n(\lambda_1) - a_2^n(\lambda_1) \cdot b_2^n(\lambda_1),$$
$$g_2^n(\lambda_1) = a_3^n(\lambda_1) \cdot b_2^n(\lambda_1) - a_2^n(\lambda_1) \cdot b_1^n(\lambda_1),$$
$$g_3^n(\lambda_1) = a_1^n(\lambda_1) \cdot a_3^n(\lambda_1) - a_2^n(\lambda_1)^2$$

with corresponding limits in probability

$$g_1(\lambda_1) = a_1(\lambda_1) \cdot b_1(\lambda_1) - a_2(\lambda_1) \cdot b_2(\lambda_1),$$
$$g_2(\lambda_1) = a_3(\lambda_1) \cdot b_2(\lambda_1) - a_2(\lambda_1) \cdot b_1(\lambda_1),$$
$$g_3(\lambda_1) = a_1(\lambda_1) \cdot a_3(\lambda_1) - a_2(\lambda_1)^2,$$

where

$$a_1(\lambda_1) = \frac{\mathbf{P}(\eta^4 \cdot w)}{\mathbf{P}w} - [\frac{\mathbf{P}(\eta^2 \cdot w)}{\mathbf{P}w}]^2,$$
$$a_2(\lambda_1) = \frac{\mathbf{P}(\eta^3 \cdot w)}{\mathbf{P}w} - \frac{\mathbf{P}(\eta \cdot w) \cdot \mathbf{P}(\eta^2 \cdot w)}{(\mathbf{P}w)^2},$$
$$a_3(\lambda_1) = \frac{\mathbf{P}(\eta^2 \cdot w)}{\mathbf{P}w} - [\frac{\mathbf{P}(\eta \cdot w)}{\mathbf{P}w}]^2,$$
$$b_1(\lambda_1) = \frac{\mathbf{P}(\eta \cdot Y^2 \cdot w)}{\mathbf{P}w} - \frac{\mathbf{P}(\eta \cdot w)\mathbf{P}(Y^2 \cdot w)}{(\mathbf{P}w)^2},$$
$$b_2(\lambda_1) = \frac{\mathbf{P}(\eta^2 \cdot Y^2 \cdot w)}{\mathbf{P}w} - \frac{\mathbf{P}(\eta^2 \cdot w)\mathbf{P}(Y^2 \cdot w)}{(\mathbf{P}w)^2}.$$

By repeatedly using Lemma 2.1, the consistency of $\hat{\lambda}_1$, and standard arguments we have

$$\sqrt{n}\{g_1^n(\hat{\lambda}_1) - g_n^1(\lambda_1)\} = \sqrt{n}\{\hat{\lambda}_1 - \lambda_1\}^T D_{\lambda_1} g_1(\lambda_1) + o_P(1),$$
$$\sqrt{n}\{g_2^n(\hat{\lambda}_1) - g_2^n(\lambda_1)\} = \sqrt{n}\{\hat{\lambda}_1 - \lambda_1\}^T D_{\lambda_1} g_2(\lambda_1) + o_P(1),$$
$$\sqrt{n}\{g_3^n(\hat{\lambda}_1) - g_3^n(\lambda_1)\} = \sqrt{n}\{\hat{\lambda}_1 - \lambda_1\}^T D_{\lambda_1} g_3(\lambda_1) + o_P(1).$$

Then $\sqrt{n}(\hat{\lambda}_2 - \lambda_2)$ may be rewritten as:

$$\sqrt{n}(\hat{\lambda}_2 - \lambda_2) = \sqrt{n}\Big(\frac{g_1^n(\hat{\lambda}_1)}{g_3^n(\hat{\lambda}_1)} - \frac{g_1^n(\lambda_1)}{g_3^n(\lambda_1)}\Big) + \sqrt{n}\Big(\frac{g_1^n(\lambda_1)}{g_3^n(\lambda_1)} - \lambda_2\Big) \tag{10}$$

For the first term of the right hand side of equation (10)

$$\begin{aligned}&\sqrt{n}\Big(\frac{g_1^n(\hat{\lambda}_1)}{g_3^n(\hat{\lambda}_1)} - \frac{g_1^n(\lambda_1)}{g_3^n(\lambda_1)}\Big)\\ &\quad= [g_3(\lambda_1)]^{-1}\sqrt{n}\{g_1^n(\hat{\lambda}_1) - g_1^n(\lambda_1)\} - [g_3(\lambda_1)]^{-2}[g_1(\lambda_1)]\sqrt{n}\{g_3^n(\hat{\lambda}_1) - g_3^n(\lambda_1))\} + o_P(1)\end{aligned} \tag{11}$$

by the consistency of $\hat{\lambda}_1$ and the law of large numbers. Combining the above results we obtain

$$\sqrt{n}\Big(\frac{g_1^n(\hat{\lambda}_1)}{g_3^n(\hat{\lambda}_1)} - \frac{g_1^n(\lambda_1)}{g_3^n(\lambda_1)}\Big) = \sqrt{n}\{\hat{\lambda}_1 - \lambda_1\}^T D_{\lambda_1}\{\frac{g_1(\lambda_1)}{g_3(\lambda_1)}\} + o_P(1) = \sqrt{n}\{\hat{\lambda}_1 - \lambda_1\}^T D_{\lambda_1}\lambda_2(\lambda_1) + o_P(1). \tag{12}$$

For the second term on the right hand side of equation (10) first note that by repeatedly using Lemma 2.2 and the law of large numbers

$$\begin{aligned}
&\sqrt{n}\{a_1^n(\lambda_1) - a_1(\lambda_1)\} =\\
&\quad(\mathbf{P}w)^{-1}\cdot\mathbb{G}_n(\{\eta^4 - 2\frac{\mathbf{P}(\eta^2 w)}{\mathbf{P}w}\eta^2 - a_1(\lambda_1) + [\frac{\mathbf{P}(\eta^2 w)}{\mathbf{P}w}]^2\}w) + o_P(1),\\
&\sqrt{n}\{a_2^n(\lambda_1) - a_2(\lambda_1)\} =\\
&\quad(\mathbf{P}w)^{-1}\cdot\mathbb{G}_n(\{\eta^3 - \frac{\mathbf{P}(\eta\cdot w)}{\mathbf{P}w}\eta^2 - \frac{\mathbf{P}(\eta^2 w)}{\mathbf{P}w}\eta - a_2(\lambda_1) + \frac{\mathbf{P}(\eta\cdot w)\cdot\mathbf{P}(\eta^2\cdot w)}{(\mathbf{P}w)^2}\}w) + o_P(1),\\
&\sqrt{n}\{a_3^n(\lambda_1) - a_3(\lambda_1)\} =\\
&\quad(\mathbf{P}w)^{-1}\cdot\mathbb{G}_n(\{\eta^2 - 2\frac{\mathbf{P}(\eta\cdot w)}{\mathbf{P}w}\eta - a_3(\lambda_1) + [\frac{\mathbf{P}(\eta\cdot w)}{\mathbf{P}w}]^2\}w) + o_P(1),\\
&\sqrt{n}\{b_1^n(\lambda_1) - b_1(\lambda_1)\} =\\
&\quad(\mathbf{P}w)^{-1}\cdot\mathbb{G}_n(\{\eta\cdot Y^2 - \frac{\mathbf{P}(\eta\cdot w)}{\mathbf{P}w}Y^2 - \frac{\mathbf{P}(Y^2 w)}{\mathbf{P}w}\eta - b_1(\lambda_1) + \frac{\mathbf{P}(\eta\cdot w)\mathbf{P}(Y^2\cdot w)}{(\mathbf{P}w)^2}\}w) + o_P(1),\\
&\sqrt{n}\{b_2^n(\lambda_1) - b_2(\lambda_1)\} =\\
&\quad(\mathbf{P}w)^{-1}\cdot\mathbb{G}_n(\{\eta^2 Y^2 - \frac{\mathbf{P}(\eta^2 w)}{\mathbf{P}w}Y^2 - \frac{\mathbf{P}(Y^2 w)}{\mathbf{P}w}\eta^2 - b_2(\lambda_1) + \frac{\mathbf{P}(\eta^2 w)\mathbf{P}(Y^2\cdot w)}{(\mathbf{P}w)^2}\}w) + o_P(1).
\end{aligned} \tag{13}$$

Next, by the law of large numbers

$$
\begin{aligned}
\sqrt{n}\{g_1^n(\lambda_1) - g_1(\lambda_1)\} =& a_1(\lambda_1)\sqrt{n}\{b_1^n(\lambda_1) - b_1(\lambda_1)\} - a_2(\lambda_1)\sqrt{n}\{b_2^n(\lambda_1) - b_2(\lambda_1)\} \\
& + b_1(\lambda_1)\sqrt{n}\{a_1^n(\lambda_1) - a_1(\lambda_1)\} - b_2(\lambda_1)\sqrt{n}\{a_2^n(\lambda_1) - a_2(\lambda_1)\} + o_P(1), \\
\sqrt{n}\{g_2^n(\lambda_1) - g_2(\lambda_1)\} =& - a_2(\lambda_1)\sqrt{n}\{b_1^n(\lambda_1) - b_1(\lambda_1)\} + a_3(\lambda_1)\sqrt{n}\{b_2^n(\lambda_1) - b_2(\lambda_1)\} \\
& - b_1(\lambda_1)\sqrt{n}\{a_2^n(\lambda_1) - a_2(\lambda_1)\} + b_2(\lambda_1)\sqrt{n}\{a_3^n(\lambda_1) - a_3(\lambda_1)\} + o_P(1), \\
\sqrt{n}\{g_3^n(\lambda_1) - g_3(\lambda_1)\} =& a_3(\lambda_1)\sqrt{n}\{a_1^n(\lambda_1) - a_1(\lambda_1)\} + a_1(\lambda_1)\sqrt{n}\{a_3^n(\lambda_1) - a_3(\lambda_1)\} \\
& - 2 \cdot a_2(\lambda_1)\sqrt{n}\{a_2^n(\lambda_1) - a_2(\lambda_1)\} + o_P(1).
\end{aligned}
$$

Now note that in the true value of λ we have

$$
\begin{aligned}
b_1(\lambda_1) &= \lambda_2 a_3(\lambda_1) + \lambda_3 a_2(\lambda_1), \\
b_2(\lambda_1) &= \lambda_2 a_2(\lambda_1) + \lambda_3 a_1(\lambda_1)
\end{aligned}
$$

and thus

$$
\begin{aligned}
g_1(\lambda_1) - \lambda_2 g_3(\lambda_1) &= 0, \\
g_2(\lambda_1) - \lambda_3 g_3(\lambda_1) &= 0.
\end{aligned}
$$

Accordingly

$$
\begin{aligned}
\sqrt{n}\{g_1^n(\lambda_1) - \lambda_2 g_3^n(\lambda_1)\} &= \sqrt{n}\{g_1^n(\lambda_1) - g_1(\lambda_1)\} - \lambda_2\sqrt{n}\{g_3^n(\lambda_1) - g_3(\lambda_1)\}, \\
\sqrt{n}\{g_2^n(\lambda_1) - \lambda_3 g_3^n(\lambda_1)\} &= \sqrt{n}\{g_2^n(\lambda_1) - g_2(\lambda_1)\} - \lambda_3\sqrt{n}\{g_3^n(\lambda_1) - g_3(\lambda_1)\}.
\end{aligned}
$$

Finally, combining all of the above and using the law of large numbers:

$$
\sqrt{n}\Big(\frac{g_1^n(\lambda_1)}{g_3^n(\lambda_1)} - \lambda_2\Big) = \{g_3(\lambda_1)\mathbf{P}w\}^{-1}\mathbf{G_n}(w\{a_1(\lambda_1)f^1 - a_2(\lambda_1)f^2\}), \tag{14}
$$

where

$$
\begin{aligned}
f_i^1 =&\eta_i Y_i^2 - \frac{\mathbf{P}(\eta \cdot w)}{\mathbf{P}w} Y_i^2 - \lambda_3 \eta_i^3 - \{\lambda_2 - \lambda_3 \frac{\mathbf{P}(\eta \cdot w)}{\mathbf{P}w}\}\eta_i^2 - \\
&\{\frac{\mathbf{P}(Y^2 w)}{\mathbf{P}w} - \lambda_2 \frac{2\mathbf{P}(\eta \cdot w)}{\mathbf{P}w} - \lambda_3 \frac{\mathbf{P}(\eta^2 w)}{\mathbf{P}w}\}\eta_i \\
&+ \frac{\mathbf{P}(\eta \cdot w)}{\mathbf{P}w}\{\frac{\mathbf{P}(Y^2 w)}{\mathbf{P}w} - \lambda_2 \frac{\mathbf{P}(\eta \cdot w)}{\mathbf{P}w} - \lambda_3 \frac{\mathbf{P}(\eta^2 w)}{\mathbf{P}w}\} \\
=&(\eta_i - \frac{\mathbf{P}(\eta \cdot w)}{\mathbf{P}w})\{Y_i^2 - E(Y_i^2|X_i)\}, \\
f_i^2 =&\eta_i^2 Y_i^2 - \frac{\mathbf{P}(\eta^2 \cdot w)}{\mathbf{P}w} Y_i^2 - \lambda_3 \eta_i^4 - \lambda_2 \eta_i^3 - \{\frac{\mathbf{P}(Y^2 w)}{\mathbf{P}w} - \lambda_3 \frac{2\mathbf{P}(\eta^2 \cdot w)}{\mathbf{P}w} - \lambda_2 \frac{\mathbf{P}(\eta \cdot w)}{\mathbf{P}w}\}\eta_i^2 \\
&+ \lambda_2 \frac{\mathbf{P}(\eta^2 w)}{\mathbf{P}w}\eta_i + \frac{\mathbf{P}(\eta^2 \cdot w)}{\mathbf{P}w}\{\frac{\mathbf{P}(Y^2 w)}{\mathbf{P}w} - \lambda_3 \frac{\mathbf{P}(\eta^2 \cdot w)}{\mathbf{P}w} - \lambda_2 \frac{\mathbf{P}(\eta \cdot w)}{\mathbf{P}w}\} \\
=&(\eta_i^2 - \frac{\mathbf{P}(\eta^2 \cdot w)}{\mathbf{P}w})\{Y_i^2 - E(Y_i^2|X_i)\}.
\end{aligned}
$$

Adding (12) and (14) we obtain

$$
\sqrt{n}\{\hat{\lambda}_2 - \lambda_2\} = \sqrt{n}\{\hat{\lambda}_1 - \lambda_1\}^T D_{\lambda_1}\lambda_2(\lambda_1) + \{g_3(\lambda_1)\mathbf{P}w\}^{-1}\mathbf{G_n}(w\{a_1(\lambda_1)f^1 - a_2(\lambda_1)f^2\}) + o_P(1)
$$

Similarly one may show

$$
\sqrt{n}\{\hat{\lambda}_3 - \lambda_3\} = \sqrt{n}\{\hat{\lambda}_1 - \lambda_1\}^T D_{\lambda_1}\lambda_3(\lambda_1) + \{g_3(\lambda_1)\mathbf{P}w\}^{-1}\mathbf{G_n}(w\{a_3(\lambda_1)f^2 - a_2(\lambda_1)f^1\}) + o_P(1).
$$

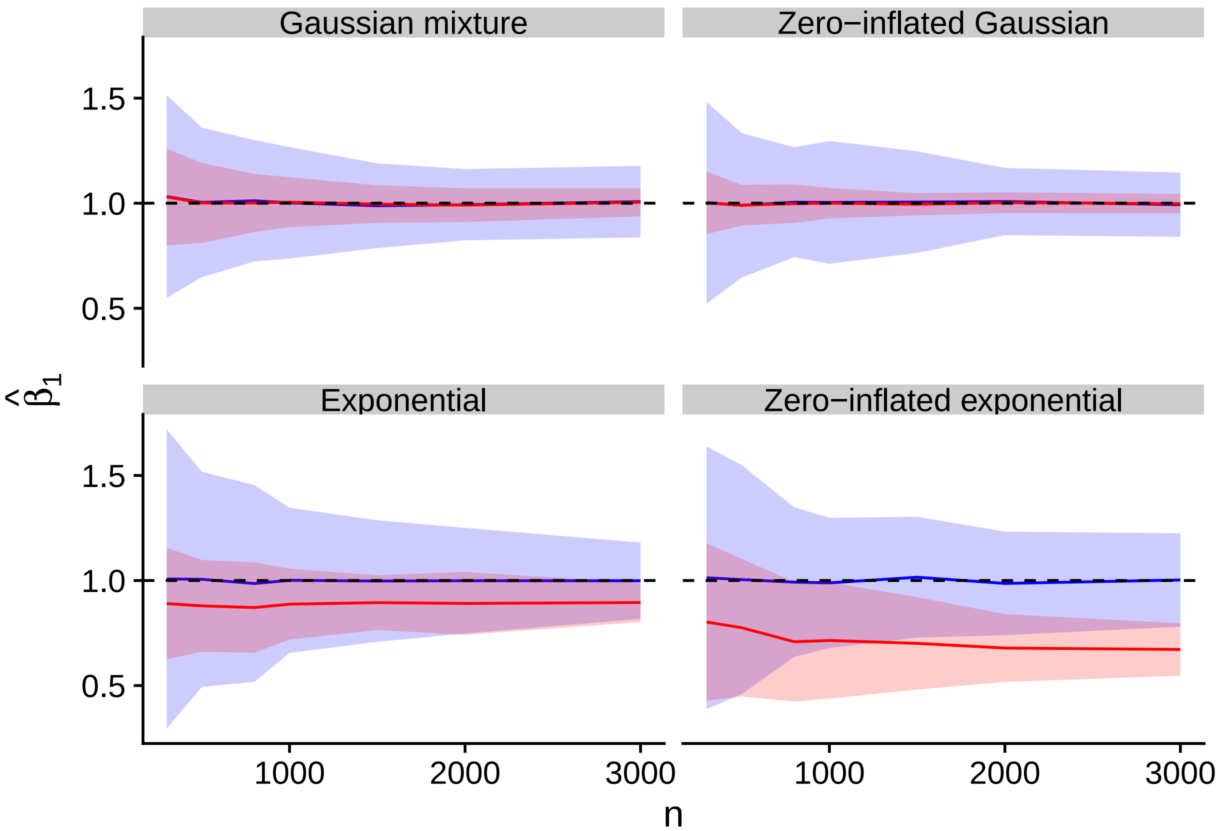

Figure 1: Average estimates of $\hat{\beta}$ using the moment mixture model (blue) and Gaussian mixture model (red) and corresponding 95% confidence pointwise confidence bands.

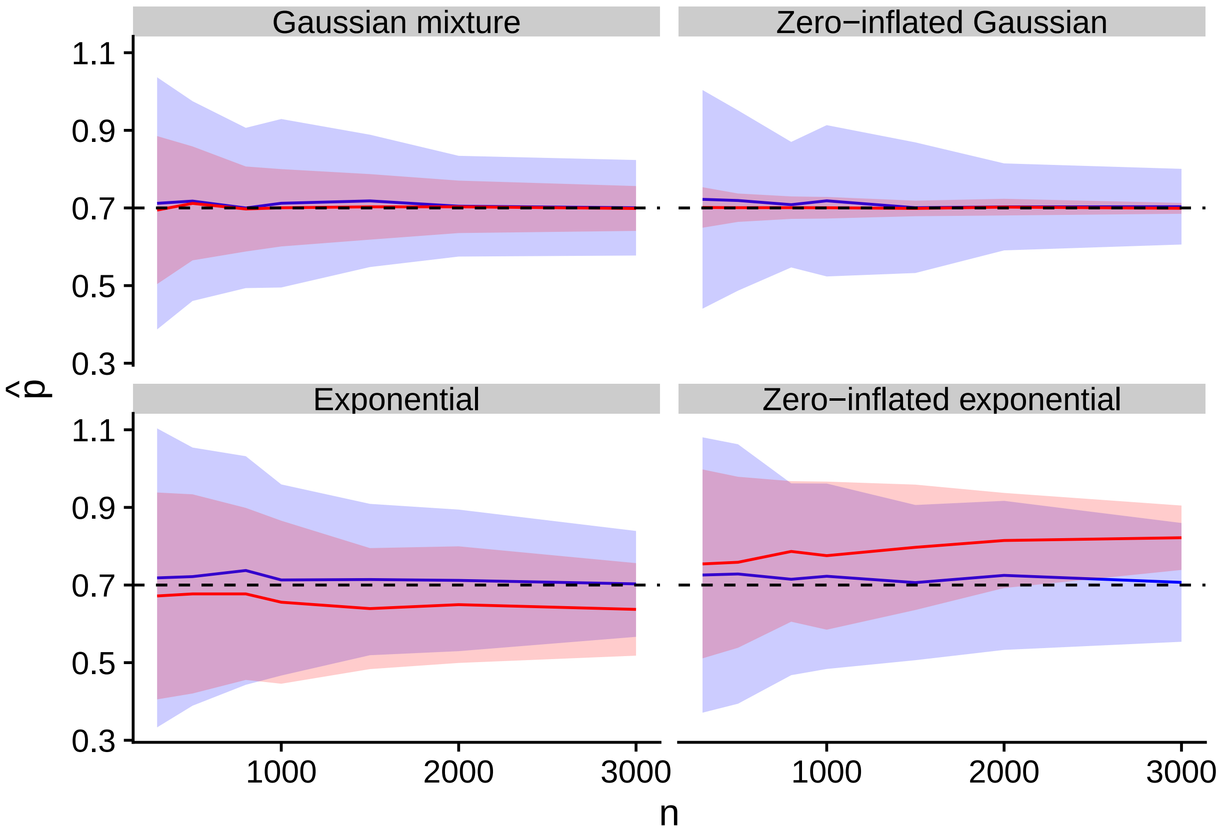

Figure 2: Average estimates of mixing proportion $\hat{p}$ using the moment mixture model (blue) and Gaussian mixture model (red) and corresponding 95% confidence pointwise confidence bands.

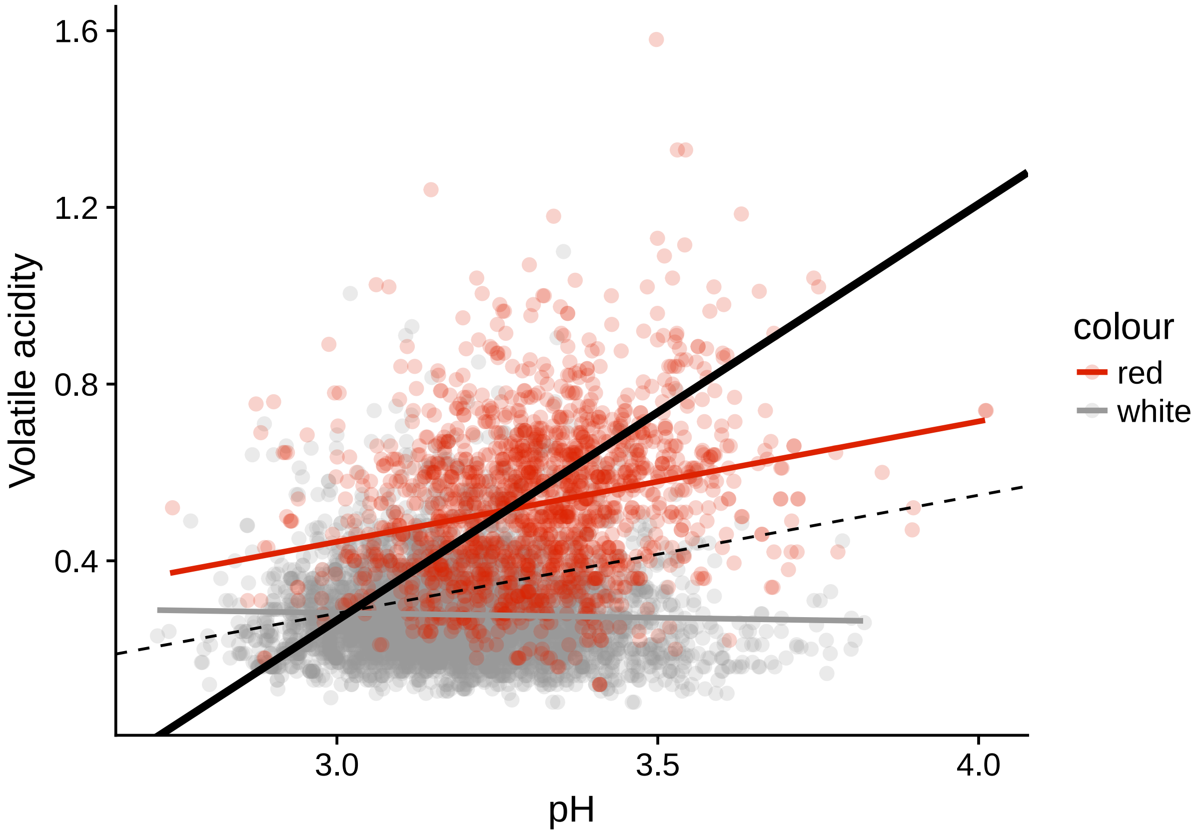

Figure 3: Volatile acidity vs. pH for white and red wines. The coloured lines show the estimated regression lines for the two corresponding types of wines, respectively. The dashed line shows the estimated regression line when a linear regression is used on all data points, while the solid line shows the estimated regression line for the moment mixture model.