Bayesian mixture models for count data

Charalampos Chanialidis

A Dissertation Submitted to the

University of Glasgow

for the degree of

Doctor of Philosophy

School of Mathematics & Statistics

November 2014

Abstract

Regression models for count data are usually based on the Poisson distribution. This thesis is concerned with Bayesian inference in more flexible models for count data. Two classes of models and algorithms are presented and studied in this thesis. The first employs a generalisation of the Poisson distribution called the COM-Poisson distribution, which can represent both overdispersed data and underdispersed data. We also propose a density regression technique for count data, which, albeit centered around the Poisson distribution, can represent arbitrary discrete distributions. The key contribution of this thesis are MCMC-based methods for posterior inference in these models.

One key challenge in COM-Poisson-based models is the fact that the normalisation constant of the COM-Poisson distribution is not known in closed form. We propose two exact MCMC algorithms which address this problem. One is based on the idea of retrospective sampling; we sample the uniform random variable u used to decide on the acceptance (or rejection) of the proposed new state of the unknown parameter first and then only evaluate bounds for the acceptance probability, in the hope that we will not need to

know the acceptance probability exactly in order to come to a decision on whether to accept or reject the newly proposed value. This strategy is based on an efficient scheme for computing lower and upper bounds for the normalisation constant. This procedure can be applied to a number of discrete distributions, including the COM-Poisson distribution. The other MCMC algorithm proposed is based on an algorithm known as the exchange algorithm. The latter requires sampling from the COM-Poisson distribution and we will describe how this can be done efficiently using rejection sampling.

We will also present simulation studies which show the advantages of using the COM-Poisson regression model compared to the alternative models commonly used in literature (Poisson and negative binomial). Three real world applications are presented: the number of emergency hospital admissions in Scotland in 2010, the number of papers published by Ph.D. students and fertility data from the second German Socio-Economic Panel.

COM-Poisson distributions are also the cornerstone of the proposed density regression technique based on Dirichlet process mixture models. Density regression can be thought of as a competitor to quantile regression. Quantile regression estimates the quantiles of the conditional distribution of the response variable given the covariates. This is especially useful when the dispersion changes across the covariates. Instead of estimating the conditional mean $\mathbb{E}(Y|X = x)$, quantile regression estimates the conditional quantile function $Q_Y(p|X = x)$ across different quantiles p where $p \in (0, 1)$. As a result, quantile regression models both location and shape shifts of the conditional distribution. This allows for a better understanding of how the covariates affect the conditional distribution of the response variable. Almost

all quantile regression techniques deal with a continuous response. Quantile regression models for count data have so far received little attention. A technique that has been suggested is adding uniform random noise ("jittering"), thus overcoming the problem that, for a discrete distribution, $Q_Y(p|X = x)$ is not a continuous function of the parameters of interest. Even though this enables us to estimate the conditional quantiles of the response variable, it has disadvantages. For small values of the response variable Y, the added noise can have a large influence on the estimated quantiles. In addition, the problem of "crossing quantiles" still exists for the jittering method. We eliminate all the aforementioned problems by estimating the density of the data, rather than the quantiles. Simulation studies show that the proposed approach performs better than the already established jittering method. To illustrate the new method we analyse fertility data from the second German Socio-Economic Panel.

Keywords: Quantile regression; Bayesian nonparametrics; Mixture models; COM-Poisson distribution; COM-Poisson regression, Markov chain Monte Carlo.

Acknowledgements

I know that I lack the writing skills to explain within a couple of sentences how I feel about my supervisors, Ludger and Tereza, but I will give it a try. Your patience, support, and guidance throughout these three years has been astounding. I can't think of a better pair of supervisors for a Ph.D. student to have. Thanks for putting up with me.

I am deeply grateful to all the people in the department. You are one of the (many) reasons why I will never forget my time in Glasgow.[1] I have to personally thank Beverley, Dawn, Jean, Kathleen, and Susan for their assistance over the years. I am pretty sure I have emailed you more times than what I have Ludger and Tereza together. A lot of thanks should also go to the academic staff of the department for providing a friendly environment for all the Ph.D. students. On a more personal note, I want to thank Agostino, Adrian, Dirk, and Marian for reasons that I would convey to them next time we meet.

I have made some really good friendships in Glasgow during the years. I feel really lucky that I had the opportunity to get to know Andrej, Chari,

[1] Living in Maryhill for a year is another reason.

Daniel, Gary, Helen, Kathakali, Lorraine, Maria, and Mustapha among others. Thanks for all the laughs, discussions, and fun we have had.

Thanks should also go to people that have never set foot on Glasgow but still manage to be there for me: Irina for all the transatlantic skype meetings we have had and are still having,[2] Grammateia and Thanos for all their help throughout the years, and finally my family for the same reasons that I have already explained in my M.Sc. thesis.[3]

Declaration

I have prepared this thesis myself; no section of it has been submitted previously as part of any application for a degree. I carried out the work reported in it, except where otherwise stated.

[2]Not letting me use the word *relish* on my post-doctoral application proved helpful.

[3]If you care that much, and know Greek, you can find the thesis online.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of methods

Quantile regression was first proposed by Koenker and Bassett (1978) as a more robust method to outliers compared to the classic linear regression (least squares regression). Since then it has been applied in areas such as economics (e.g. effects of union membership on wages (Chamberlain, 1994), hedge fund strategies (Meligkotsidou et al., 2009)), educational reform (e.g. effects of reducing class size on students (Levin, 2001)) and public health (e.g. pollution levels on upper quantiles (Lee and Neocleous, 2010)) among many others. One reason for the popularity of quantile regression is that it allows us to explore the entire conditional distribution, including the tails, of the response variable given the covariates. Thus, we can focus on the lower tail of the distribution if we are interested in poverty studies (which concern the low-income population) and on the upper tail if we are interested

in tax-policy studies that usually concern the high-income population.

In addition, if the assumptions made in least-squares regression such as Gaussianity or homoscedasticity do not hold, then, by just looking at the changes to the mean, we may under/overestimate or even fail to see what is happening to the conditional distribution of the response variable. Unlike least squares regression, where all inferences depend on the estimated parameter $\hat{\boldsymbol{\beta}}$, quantile regression allows us to be more precise since the estimated parameters $\hat{\boldsymbol{\beta}}_p$ depend on the quantile p. Quantile regression can suffer from the problem of "crossing quantile" curves, which is usually seen in sparse regions of the covariate space. This happens due to the fact that the estimated conditional quantile curve for a given $X = x$ is not necessarily a monotonically increasing function of p. This is a notable problem of quantile regression, for which there exists no general solution. Koenker (1984) considers parallel quantile planes in order to avoid the "crossing quantiles" problem. He (1997) and Wu and Liu (2009) propose methods to estimate the quantile curves while at the same time ensuring that they will be non-crossing. This problem also affects nonlinear quantile curves where different methods for solving it have been proposed. More information is given by Dette and Volgushev (2008); Chernozhukov et al. (2009, 2010) and Bondell et al. (2010).

Most quantile regression techniques deal with a continuous response. The problem with applying quantile regression to count data is that the distribution of the response variable is not continuous. As a result, the quantiles are not continuous either, and they cannot be expressed as a continuous function of the covariates. Machado and Santos Silva (2005) overcome this problem by adding uniform random noise ("jittering") to the counts. The general

idea is to construct a continuous variable Z whose conditional quantiles have a one-to-one relationship with the conditional quantiles of the counts Y and use this for inference. After estimating the conditional quantiles of Z we can now use the previous relationship to get the conditional quantiles of the counts Y. This approach eliminates the problem of having a non-continuous distribution for the response variable, but it has the drawback that for small values of Y the estimated conditional quantiles $Q_Y(p|X = x)$ will not be good estimates of the true conditional quantiles. This approach has been applied in the analysis of traffic accidents in Qin and Reyes (2011) and Wu et al. (2014), frequency of individual doctor visits in Winkelmann (2006) and Moreira and Barros (2010), and fertility data in Miranda (2008) and Booth and Kee (2009).

We overcome both aforementioned problems ("jittering" when Y takes small values and the "crossing quantiles" problem) by estimating the conditional density[1] of the response variable and by obtaining the quantiles through the density. The Bayesian density estimation methods that we will follow throughout the thesis are based on Dirichlet process models, which are also known as infinite mixture models. The idea behind mixture models is that the observed data cannot be characterised by a single distribution but instead by several; with the distribution used for a given observation chosen at random. In a sense we treat a population as if it consists of several subpopulations. We can apply these models to data where the observations come from different groups and the group memberships are not known, but also to represent

[1]The word "density" will be used both for the probability mass function in the discrete case and the probability density function in the continuous case.

multimodal distributions. An infinite mixture model can be thought of as a mixture model with a countably infinite number of components. It is different to a finite mixture model because it does not use a fixed number of components to model the data. The number of components can be inferred from the data using the Bayesian posterior inference scheme. Neal (2000) proposes different ways for sampling from the posterior of a Dirichlet process model.

In order to be able to estimate any form of conditional density, we assume that the conditional distribution of the counts Y can be expressed as a Dirichlet process mixture of regression models where the mixing weights vary with covariates. The weights are dependent on the distance between the values of the covariates as proposed by Dunson et al. (2007) when considering Bayesian methods for density regression. Density regression is similar to quantile regression in that it allows flexible modelling of the response variable Y given the covariates $X = x$. Features (mean, quantiles, spread) of the conditional distribution of the response variable vary with X, so, depending on the predictor values, features of the conditional distribution can change in a different way than the population mean. The difference between density regression and quantile regression is that density regression models the probability density function rather than directly modelling the quantiles. Specifically, we will assume that the conditional distribution of the counts can be expressed as a Dirichlet process mixture of COM-Poisson regression models.

The Conway-Maxwell-Poisson (or COM-Poisson) distribution was first proposed in Conway and Maxwell (1962) in the context of queuing systems with state-dependent service rates and brought back to surface by Shmueli et al.

(2005). Due to its extra parameter, compared to the Poisson distribution, it is flexible enough to handle any kind of dispersion. The main reason why the COM-Poisson is not used as much in practice is that its normalisation constant is not available in closed form and approximations to it are either computationally inefficient or not sufficiently exact. In the context of COM-Poisson mixtures, we overcome this problem by resorting to an MCMC strategy, known as the exchange algorithm (Murray et al., 2006). The key idea of the exchange algorithm is to introduce auxiliary data, which allows cancelling out the normalisation constants which are difficult to compute.

To recap, we will estimate the conditional density by bridging:

i) an MCMC algorithm for sampling from the posterior distribution of a Dirichlet process model, with a non-conjugate prior, found in Neal (2000).

ii) The MCMC algorithm in Dunson et al. (2007).

iii) A variation of the MCMC exchange algorithm of Murray et al. (2006).

Besides the above implementation of Bayesian density regression, we will also focus on the COM-Poisson regression model. Shmueli et al. (2005) describe methods for estimating the parameters of the COM-Poisson distribution and show its flexibility in fitting count data compared to other distributions. The advantage of this model is that it allows separation between a covariate's effect on the mean of the counts and on the variance of the counts. The disadvantage is, as we have already mentioned, that the normalisation constant has to be approximated. Minka et al. (2003) provide an asymp-

totic approximation that is only reasonably accurate in some parts of the parameter space.

We propose an exact MCMC algorithm that is based on the idea of retrospective sampling found in Papaspiliopoulos and Roberts (2008), and compute lower and upper bounds for the acceptance probability of the Metropolis-Hastings MCMC algorithm. The basic idea is that there is not always a need to know the acceptance probability of the MCMC exactly. Often we can make a decision (accept/reject) only based on lower and upper bounds for the acceptance probability. We will also show how one can sample from the COM-Poisson distribution and thus use the exchange algorithm for posterior inference in COM-Poisson regression models.

In Chapter 5 we will demonstrate this method using data on emergency hospital admissions in Scotland in 2010 where the main interest lies in the estimation of the variability of admissions, as it is considered a proxy for health inequalities. The COM-Poisson regression model is an ideal model for this data set, since it allows modelling the mean and the variance explicitly. As a result, we are able to identify areas with a high level of health inequalities. Furthermore, the results show that in order for the MCMC to make a decision between accepting or rejecting a move, the approximation of the bounds of the acceptance probability does not, usually, need to be precise.

1.2 Thesis organisation

Chapter 2 provides a literature review of distributions for count data, regression models for count data, quantile regression, mixture models, Dirichlet processes, Bayesian inference and Bayesian density regression models, which all form the background theory needed for the understanding of the thesis.

Chapter 3 introduces the proposed simulation techniques for intractable likelihoods, based on retrospective sampling and the exchange algorithm, and presents the MCMC algorithms for each one.

Chapters 4 and 5 are each split into two sections; the first section is related to the COM-Poisson regression model while the other is related to the Bayesian density regression model. These refer to the regression models (Chapter 4), and simulations and case studies (Chapter 5) for each model. The thesis is structured in this way for two reasons: to show the similarities and differences between the models, and to make it an easier read for people who are mainly interested in a specific topic.

Conclusions and future work are included in Chapter 6.

1.3 Contributions

The work presented in Chapter 3 (Section 3.2) and Chapter 5 (Section 5.2) has been published in Stat with the title *Retrospective MCMC sampling with an application to COM-Poisson regression* (Chanialidis et al., 2014) and was presented at the 1st International Conference of Statistical Distributions and Applications.

The work presented in Chapter 3 (Section 3.3) and Chapter 5 (Section 5.2) has been submitted for publication with the title *Efficient Bayesian inference for COM-Poisson regression models.*

The work presented in Chapter 4 has been published in the Proceedings of the 21st International Conference on Computational Statistics with the title *Bayesian density regression for count data* and was presented at the above conference and the 2nd Bayesian Young Statisticians conference.

All the above contributions can be found on my website.[2]

[2] http://www.chanialidis.com

Chapter 2

Review of background theory

2.1 Distributions for count data

2.1.1 Poisson distribution

Count data are typically used to model the number of occurrences of an event within a fixed period of time. Examples of count data may include

- the number of goals scored by a team.
- the number of telephone connections to a wrong number.
- the number of murders in a city.

The Poisson distribution is the most popular model used for modelling a discrete random variable Y. It is used to describe "rare" events and it is derived under three assumptions.

1. The probability of one event happening in a short interval is proportional to the length of the interval.

2. The number of events in non-overlapping intervals is independent, and

3. the probability of two events happening in a short interval is negligible in comparison to the probability of a single event happening.

The probability mass function of the Poisson(μ) distribution is

$$P(Y = y|\mu) = \exp\{-\mu\}\frac{\mu^y}{y!} \quad y = 0, 1, 2, \ldots \tag{2.1}$$

The mean and variance of a Poisson(μ) are respectively

$$\begin{aligned} \mathbb{E}[Y] &= \mu, \\ \mathbb{V}[Y] &= \mu. \end{aligned} \tag{2.2}$$

The equations in (2.2) show that the Poisson distribution assumes that the mean is equal to its variance; this is known as the equidispersion assumption. This assumption also implies that the Poisson distribution does not allow for the variance to be adjusted independently of the mean. In the presence of underdispersed data (variance is less than the mean) or overdispersed data (variance is greater than the mean) the Poisson distribution is not an appropriate model and one has to use another parametric model, with an additional parameter compared to the Poisson. For overdispersed data, one of the distributions that may provide a better fit is the negative binomial distribution.

2.1.2 Negative binomial distribution

The probability mass function of the negative binomial(r, p) distribution is

$$P(Y = y|r, p) = p^y(1-p)^r \binom{y+r-1}{y} \quad y = 0, 1, 2, \ldots \tag{2.3}$$

The mean and variance of a NB(r, p) are respectively

$$\begin{aligned} \mathbb{E}[Y] &= \frac{pr}{(1-p)}, \\ \mathbb{V}[Y] &= \frac{pr}{(1-p)^2}. \end{aligned} \tag{2.4}$$

An alternative formulation of the negative binomial distribution is

$$P(Y = y|\mu, k) = \frac{\Gamma(\frac{1}{k}+y)}{\Gamma(\frac{1}{k})y!}\left(\frac{k\mu}{1+k\mu}\right)^y\left(\frac{1}{1+k\mu}\right)^{\frac{1}{k}} \tag{2.5}$$

with mean and variance for the negative binomial(μ, k)

$$\begin{aligned} \mathbb{E}[Y] &= \mu, \\ \mathbb{V}[Y] &= \mu + k\mu^2. \end{aligned} \tag{2.6}$$

The first parameter of this formulation is the mean of the distribution whereas the second is referred to as the dispersion parameter. Large values of k are a sign of overdispersion, while when $k \to 0$ the variance of the distribution (cf. (2.6)) is equal to the mean and we have the Poisson model as a special case.

The negative binomial distribution can also be seen as a continuous mixture of Poisson distributions in which the mixing distribution of the Poisson parameter μ follows a gamma distribution. In this way, we treat the Poisson parameter μ as a random variable and we assign to it a gamma distribution.

Later, in 2.1.4, we will see that there are a plethora of choices for the mixing distribution.

Suppose that

$$\begin{aligned} y &\sim \text{Poisson}(\mu), \\ \mu &\sim \text{gamma}(a, b). \end{aligned} \tag{2.7}$$

Then

$$\begin{aligned} f(y) &= \int_0^\infty f(y, \mu)\ \mathrm{d}\mu \\ &= \int_0^\infty f(y|\mu) f(\mu)\ \mathrm{d}\mu \\ &= \int_0^\infty \exp\{-\mu\} \frac{\mu^y}{y!} \frac{b^a}{\Gamma(a)} \mu^{a-1} \exp\{-b\mu\}\ \mathrm{d}\mu \\ &= \frac{b^a}{\Gamma(a)} \frac{\mu^y}{y!} \int_0^\infty \mu^{y+a-1} \exp\{-(b+1)\mu\}\ \mathrm{d}\mu \\ &= \left(\frac{b}{b+1}\right)^a \left(\frac{1}{b+1}\right)^y \frac{\Gamma(y+a)}{\Gamma(y+1)\Gamma(a)}. \end{aligned} \tag{2.8}$$

which is the probability mass function of a negative binomial(r, p) with

$$p = \frac{b}{1+b}\,,\ r = a. \tag{2.9}$$

Equations (2.6) show that the negative binomial cannot model underdispersed data. Some well-known discrete distributions are summarised in Table 2.1.

Bernoulli distribution	$Y = 1$ if event $A = \{\text{success}\}$ occurs	θ	probability of success	$\mathbb{E}[Y] = \theta$
$R_y = \{0, 1\}$	$Y = 0$ otherwise			$\mathbb{V}[Y] = \theta(1-\theta)$
$P(0) = 1-\theta, P(1) = \theta$				
Binomial distribution	number of successes in independent Bernoulli trials	n	number of Bernoulli trials	$\mathbb{E}[Y] = n\theta$
$R_y = \{0, 1, \ldots, n.\}$	(sampling *with* replacement from population of type I	θ	probability of success	$\mathbb{V}[Y] = n\theta(1-\theta)$
$P(Y = y\|n, \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$	(proportion θ) and type II (proportion $1-\theta$) objects)			
Geometric distribution	number of independent trials required until first	θ	probability of success	$\mathbb{E}[Y] = \frac{1}{1-\theta}$
$R_Y = \mathbb{N}$	failure occurs			$\mathbb{E}[Y] = \frac{\theta}{(1-\theta)^2}$
$P(Y = y\|\theta) = \theta^{y-1}(1-\theta)$				
Poisson distribution	number of events in time interval	μ	average number of events	$\mathbb{E}[Y] = \mu$
$R_Y = \mathbb{N}_0$				$\mathbb{V}[Y] = \mu$
$P(Y = y\|\mu) = \exp\{-\mu\}\frac{\mu^y}{y!}$				
Negative binomial distribution	number of successes in independent Bernoulli trials	r	number of failures	$\mathbb{E}[Y] = \frac{pr}{(1-p)}$
$R_Y = \mathbb{N}_0$	before a specified number of failures occur	p	probability of success	$\mathbb{V}[Y] = \frac{pr}{(1-p)^2}$
$P(Y = y\|r, p) = p^y(1-p)^r\binom{y+r-1}{y}$				

Table 2.1: Overview of some well known discrete distributions.

2.1.3 COM-Poisson distribution

The COM-Poisson distribution (Conway and Maxwell, 1962) is a two-parameter generalisation of the Poisson distribution that allows for different levels of dispersion. The probability mass function of the COM-Poisson(λ, ν) distribution is

$$P(Y = y|\lambda, \nu) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)} \quad y = 0, 1, 2, \ldots$$
$$Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}, \tag{2.10}$$

for $\lambda > 0$ and $\nu \geq 0$.

The additional parameter ν, compared to the Poisson distribution, allows the COM-Poisson distribution to model underdispersed ($\nu > 1$) or overdispersed ($\nu < 1$) data. The Poisson distribution is a special case ($\nu = 1$). The ratio of two successive probabilities is

$$\frac{P(Y = y - 1)}{P(Y = y)} = \frac{y^\nu}{\lambda}. \tag{2.11}$$

The range of possible values of ν covers all different kinds of dispersion levels. Values of ν less than one correspond to flatter successive ratios compared to the Poisson distribution. This means that the distribution has longer tails (e.g. overdispersion). On the other hand, when ν is greater than one, we have underdispersion.

The COM-Poisson distribution is a generalisation of other well known discrete distributions:

- For $\nu = 0,\ \lambda < 1$ the distribution is a geometric(1-λ).

- For $\nu = 1$ the distribution is a Poisson(λ).
- For $\nu \to \infty$ it approaches a Bernoulli($\frac{\lambda}{\lambda+1}$) distribution.

For $\nu \neq 1$ the normalisation constant $Z(\lambda, \nu)$ does not have a closed form and has to be approximated.

Evaluating the normalisation constant $Z(\lambda, \nu)$

Minka et al. (2003) give an upper bound for the normalisation constant and an asymptotic approximation which is reasonably accurate for $\lambda > 10^{\nu}$. The upper bound on $Z(\lambda, \nu)$ is estimated using the fact that the series $\frac{\lambda^j}{(j!)^\nu}$ converges and $\lim_{j\to\infty} \frac{\lambda^j}{(j!)^\nu} = 0$.

As a result there exists a value k such that, for $j > k$,

$$\frac{\lambda}{j^\nu} < 1. \tag{2.12}$$

This ratio is monotonically decreasing, which means that for $j > k$, this series converges faster than a geometric series with multiplier given by (2.12). Minka et al. (2003) truncate the series at the k^{th} term such that

$$Z(\lambda, \nu) = \sum_{j=0}^{k} \frac{\lambda^j}{(j!)^\nu} + R_k, \tag{2.13}$$

where

$$R_k = \sum_{j=k+1}^{\infty} \frac{\lambda^j}{(j!)^\nu} \tag{2.14}$$

is the absolute truncation error.

The absolute truncation error R_k is bounded by

$$\frac{\lambda^{k+1}}{(k+1)!^\nu (1-\epsilon_k)}, \tag{2.15}$$

where ϵ_k is such that $\frac{\lambda}{(j+1)^\nu} < \epsilon_k$ for all $j > k$. A computational improvement that increases efficiency is to bound the relative truncation error given by

$$\frac{R_k}{\sum_{j=0}^{k} \frac{\lambda^j}{(j!)^\nu}}. \tag{2.16}$$

For $\nu \leq 1$, truncation of the infinite sum is costly since there is a large number of summations needed in order to achieve sensible accuracy. In that case, Minka et al. (2003) use an asymptotic approximation for $Z(\lambda, \nu)$,

$$Z(\lambda, \nu) = \frac{\exp\{\nu\lambda^{\frac{1}{\nu}}\}}{\lambda^{\frac{\nu-1}{2\nu}}(2\pi)^{\frac{\nu-1}{2}}\sqrt{\nu}}(1 + O(\lambda^{-\frac{1}{\nu}})). \tag{2.17}$$

The formula in (2.17) has been derived for integer ν. The main message from this formula is that the normalisation constant $Z(\lambda, \nu)$ grows rapidly as λ increases or ν decreases.

Estimating the COM-Poisson parameters

Shmueli et al. (2005) describe three methods for estimating the parameters of the COM-Poisson distribution and show its flexibility in fitting count data compared to other distributions.

1. The first method is based on equation (2.11). Taking a log of both sides of the equation

$$\log\left\{\frac{P(Y = y-1)}{P(Y = y)}\right\} = -\log\{\lambda\} + \nu\log\{y\}. \tag{2.18}$$

The ratio on the left hand side can be estimated by replacing the probabilities with the relative frequencies of $y - 1$ and y respectively. One

can plot these values versus $\log\{y\}$ for all the ratios that do not include the zero counts. A COM-Poisson would be an adequate model if the points fall on a straight line. If the data do appear to fit a COM-Poisson model, the parameters can be estimated by fitting a regression of $\log\left\{\frac{P(Y=y-1)}{P(Y=y)}\right\}$ on $\log\{y\}$.

2. The second method is based on the maximum likelihood approach. The likelihood for a set of n independent and identically distributed observations $y_1, y_2, \ldots, y_n$ is

$$\begin{aligned} L(y_1, y_2, \ldots, y_n|\lambda, \nu) &= \frac{\prod_{i=1}^{n} \lambda^{y_i}}{\prod_{i=1}^{n} (y_i!)^{\nu}} Z(\lambda, \nu)^{-n} \\ &= \lambda^{S_1} \exp\{-\nu S_2\} Z(\lambda, \nu)^{-n}. \end{aligned} \tag{2.19}$$

where $S_1 = \sum_{i=1}^{n} y_i$ and $S_2 = \sum_{i=1}^{n} \log\{y_i!\}$.

Equation (2.19) shows that (S_1, S_2) are sufficient statistics for $y_1, y_2, \ldots, y_n$, and that the COM-Poisson is a member of the exponential family since it can be expressed in the form

$$L(\boldsymbol{y}|\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta})\phi(\boldsymbol{y}) \exp\left\{\sum_{j=1}^{k} \pi_j(\boldsymbol{\theta}) t_j(\boldsymbol{y})\right\} \tag{2.20}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\intercal$ and $\boldsymbol{\theta} = (\lambda, \nu)$.

For the COM-Poisson case

$$\begin{aligned} \pi_1(\boldsymbol{\theta}) &= \log\{\lambda\}, & \pi_2(\boldsymbol{\theta}) &= -\nu, \\ t_1(\boldsymbol{y}) &= \sum_{i=1}^{n} y_i, & t_2(\boldsymbol{y}) &= \sum_{i=1}^{n} \log\{y_i!\}. \end{aligned} \tag{2.21}$$

3. The third method is based on Bayesian inference, see Section 2.5 for more information on Bayesian inference. This approach takes advantage of the exponential family structure of the COM-Poisson distribution to establish a conjugate family of priors. Kadane et al. (2006)

show that the conjugate prior density of the COM-Poisson distribution is of the form:

$$h(\lambda, \nu) = \lambda^{a-1} \exp\{-b\nu\} Z^{-c}(\lambda, \nu)\, k(a, b, c) \tag{2.22}$$

for $\lambda > 0$ and $\nu \geq 0$, where $k(a, b, c)$ is the normalisation constant. The posterior then is of the same form, with

$$a' = a + S_1,\ b' = b + S_2,\ c' = c + n. \tag{2.23}$$

The conjugate prior can be thought of as an extended bivariate gamma distribution. In order for equation (2.22) to constitute a density, it must be non-negative and integrate to one. The values of a, b, c that lead to a finite $k(a, b, c)^{-1}$, which is given by

$$k(a, b, c)^{-1} = \int_0^\infty \int_0^\infty \lambda^{a-1} \exp\{-b\nu\} Z^{-c}(\lambda, \nu)\ \mathrm{d}\lambda\ \mathrm{d}\nu, \tag{2.24}$$

will lead to a proper density. A necessary and sufficient condition for equation (2.22) to constitute a density is

$$\frac{b}{c} > \log\left\{\left\lfloor \frac{a}{c} \right\rfloor\right\} + \left(\frac{a}{c} - \left\lfloor \frac{a}{c} \right\rfloor\right) \log\left\{\frac{a}{c} + 1\right\} \tag{2.25}$$

where $\lfloor k \rfloor$ denotes the floor function which returns the highest integer smaller than, or equal to, k. Estimating the double integral in (2.24) is not straightforward since it includes an infinite sum. Kadane et al. (2006) calculate the double integral by using a non-equally spaced grid over the λ, ν space.

Approximations for the mean and variance of the COM-Poisson distribution

The COM-Poisson distribution belongs to the family of two-parameter power series distributions (Johnson et al., 2005). Moments of this distribution can then be obtained using the recursive formula:

$$\mathbb{E}[Y^{r+1}] = \begin{cases} \lambda \mathbb{E}[(Y+1)^{1-\nu}] & r = 0 \\ \lambda \frac{d}{d\lambda} \mathbb{E}[Y^r] + \mathbb{E}[Y]\mathbb{E}[Y^r] & r > 0 \end{cases} \tag{2.26}$$

Using i) the asymptotic approximation for the normalisation constant, equation (2.17), ii) equation (2.26), iii) and the fact that $\mathbb{V}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$, Shmueli et al. (2005) show that the mean and variance can be approximated by

$$\begin{aligned} \mathbb{E}[Y] &\approx \lambda^{\frac{1}{\nu}} + \frac{1}{2\nu} - \frac{1}{2}, \\ \mathbb{V}[Y] &\approx \frac{\lambda^{\frac{1}{\nu}}}{\nu}. \end{aligned} \tag{2.27}$$

Reparameterising the COM-Poisson distribution

The previous parameterisation of the COM-Poisson distribution, see (2.10), does not have a clear centering parameter, so we will use the reparameterisation $\mu = \lambda^{\frac{1}{\nu}}$ as proposed by Guikema and Coffelt (2008). The probability mass function of the COM-Poisson(μ, ν) becomes

$$\begin{aligned} P(Y = y|\mu,\nu) &= \left(\frac{\mu^y}{y!}\right)^\nu \frac{1}{Z(\mu,\nu)} \quad y = 0, 1, 2, \ldots \\ Z(\mu,\nu) &= \sum_{j=0}^{\infty} \left(\frac{\mu^j}{j!}\right)^\nu, \end{aligned} \tag{2.28}$$

for $\mu > 0$ and $\nu \geq 0$.

The mean and variance can be approximated by

$$\begin{aligned} \mathbb{E}[Y] &\approx \mu + \frac{1}{2\nu} - \frac{1}{2}, \\ \mathbb{V}[Y] &\approx \frac{\mu}{\nu}. \end{aligned} \tag{2.29}$$

Thus, in the new parameterisation μ closely approximates the mean, unless both μ and ν are small. The mode of the distribution is $\lfloor \mu \rfloor$, as this formulation is just a tempered Poisson distribution.

The fact that $Z(\mu, \nu)$, and thus the probability mass function $P(Y = y|\mu, \nu)$, is very expensive to compute, has been a key limiting factor for the use of the COM-Poisson distribution. In particular, in a Bayesian approach using the Metropolis-Hastings algorithm, each move requires an evaluation of $Z(\mu, \nu)$ in order to compute the acceptance probability. In Chapter 3 we will present two MCMC algorithms that do not need $Z(\mu, \nu)$ to be computed exactly. The first one (retrospective sampling algorithm) takes advantage of lower and upper bounds of the normalisation constant $Z(\mu, \nu)$ while the second one (exchange algorithm) requires no computation of $Z(\mu, \nu)$ at all.

2.1.4 Other distributions

Del Castillo and Pérez-Casany (1998) developed a family of distributions, known as weighted Poisson distributions, that can handle both underdispersed and overdispersed data. A discrete random variable Y is defined to have a weighted Poisson distribution if its probability mass function can be

written as

$$P(Y = y|\lambda, r, a) = \exp\{-\lambda\}\frac{\lambda^y w_y}{W y!} \quad y = 0, 1, 2, \ldots$$
$$W = \exp\{-\lambda\}\sum_{j=0}^{\infty}\frac{\lambda^j w_j}{j!} \tag{2.30}$$

where the weight function is defined as $w_y = (y+a)^r$ with $a \geq 0, r \in \mathbb{R}$. The Poisson distribution is a special case ($r = 0$). These distributions are used for modelling data with partial recording: when the event $Y = y$ occurs, a Poisson variable is recorded with probability proportional to w_y.

Ridout and Besbeas (2004) present a distribution for modelling underdispersed count data which is based on the weighted Poisson distribution. Its difference lies in the weights $w_y = \exp\{r|y - \lambda|\}$ which, in this case, are centered on the mean of the Poisson distribution. They refer to the distribution as the three-parameter exponentially weighted Poisson distribution (EWP_3). Cameron and Johansson (1997) used the Poisson polynomial distribution to model the number of takeover bids received by targeted firms. This distribution is another weighted Poisson distribution with weight function of the polynomial form

$$w_y = (1 + \sum_{j=1}^{k} a_j y^j)^2, \; a_j \in \mathbb{R}. \tag{2.31}$$

The COM-Poisson distribution can be seen as a weighted Poisson distribution with weight function $w_y = (y!)^{1-\nu}$ (Rodrigues et al., 2009).

Another distribution that can handle under- and overdispersion is the generalised Poisson distribution of Consul and Famoye (1992). A discrete random variable Y is defined to have a generalised Poisson distribution if its proba-

bility mass function can be written as

$$P(Y=y|\lambda,\theta)=\begin{cases}\exp\{-\lambda-\theta y\}\frac{\lambda(\lambda+\theta y)^{y-1}}{y!} & y=0,1,2,\ldots \\ 0 & \text{for } y>m \text{ when } \theta<0.\end{cases} \tag{2.32}$$

where $\lambda>0$, $\max\{-1,-\frac{\lambda}{4}\}\leq\theta\leq 1$ and m is the largest positive integer for which $\lambda+m\theta>0$ when θ is negative. For $\theta=0$, the generalised Poisson distribution reduces to the Poisson model, see (2.1). Positive (or negative) values of θ correspond to overdispersion (or underdipersion). A weakness of the generalised Poisson distribution is its inability to capture some levels of underdispersion since for large (in absolute value) negative values of θ the model in (2.32) is not a true probability distribution (unless truncated).

Rigby et al. (2008) present methods for modelling underdispersed and overdispersed data. The methods are classified into three main categories

1. *Ad hoc* methods.
2. Discretised continuous distributions.
3. Random effect at the observation level solutions.

The methods belonging in the first category do not assume an explicit distributional form for the discrete random variable. These methods require assumptions on the first two moments of the response variable such as the quasi-likelihood approach (Wedderburn, 1974). Alternative approaches include the pseudo-likelihood method (Carroll and Ruppert, 1982) and the double exponential family (Efron, 1986).

Discretised continuous distributions refer to methods which use continuous distributions to create a discrete one. For example, let $F_W(w)$ be the cumulative distribution function of a continuous random variable W defined in $\mathbb{R}^+$ then $f_Y(y) = F_W(y+1) - F_w(y)$ is a discrete distribution defined on $\mathbb{R}^+_y$.

Including an extra random effect variable is another way to handle overdispersion. Given a random effect variable γ, the response variable Y has a discrete probability function $f(y|\gamma)$ whereas γ has probability (density) function $f_\gamma(\gamma)$. The marginal probability function of Y is given by

$$f_Y(y) = \int f(y|\gamma) f_\gamma(\gamma)\ \mathrm{d}\gamma. \tag{2.33}$$

The negative binomial distribution, see (2.8), is an example of this method where the mixing distribution of the random effect follows a gamma distribution. Table 2.2 shows some overdispersed count data distributions $f_Y(y)$ along with their mixing distributions $f_\gamma(\gamma)$, where $f(y|\gamma)$ has a Poisson distribution. Among the overdispersed distributions seen in Table 2.2 are the Sichel, the Delaporte and the Poisson shifted generalised inverse Gaussian distribution.

The Sichel distribution is a three parameter distribution with probability mass function

$$\begin{aligned} P(Y = y|\mu,\sigma,\nu) &= \frac{(\frac{\mu}{c})^y K_{y+\nu}(a)}{y!(a\sigma)^{y+\nu} K_\nu\left(\frac{1}{\sigma}\right)} \quad y = 0,1,2,\ldots \\ c &= R_\nu\left(\frac{1}{\sigma}\right), \\ R_\lambda(t) &= \frac{K_{\lambda+1}(t)}{K_\lambda(t)}, \\ K_\lambda(t) &= \frac{1}{2}\int_0^\infty x^{\lambda-1}\exp\{-\frac{1}{2}t(x+\frac{1}{x})\}\ \mathrm{d}x, \end{aligned} \tag{2.34}$$

where $K_\lambda(t)$ is the modified Bessel function of the third kind.

The mean and variance for the Sichel(μ, σ, ν) are given by

$$\begin{aligned}\mathbb{E}[Y] &= \mu, \\ \mathbb{V}[Y] &= \mu + \mu^2 \left(\frac{2\sigma(\nu+1)}{c} + \frac{1}{c^2} - 1 \right).\end{aligned} \tag{2.35}$$

The Delaporte distribution is another three parameter distribution with probability mass function

$$\begin{aligned}P(Y = y|\mu, \sigma, \nu) &= \frac{\exp\{-\mu\nu\}}{\Gamma(\frac{1}{\sigma})} \left(1 + \mu\sigma(1-\nu)\right)^{-\frac{1}{\sigma}} S \quad y = 0, 1, 2, \ldots \\ S &= \sum_{j=0}^{y} \binom{y}{j} \frac{\mu^y \nu^{y-j}}{y!} \left(\mu + \frac{1}{\sigma(1-\nu)} \right)^{-j} \Gamma\left(\frac{1}{\sigma} + j \right)\end{aligned} \tag{2.36}$$

where the gamma function $\Gamma(x)$ is defined as

$$\Gamma(x) = \int_0^\infty x^{t-1} \exp\{-x\} \, \mathrm{d}x. \tag{2.37}$$

The mean and variance for the Delaporte(μ, σ, ν) are given by

$$\begin{aligned}\mathbb{E}[Y] &= \mu, \\ \mathbb{V}[Y] &= \mu + \mu^2\sigma(1-\nu)^2.\end{aligned} \tag{2.38}$$

More information on the Sichel, Delaporte, and other discrete univariate distributions can be found on Johnson et al. (2005).

Finally, Rigby et al. (2008) introduce a new four parameter distribution, the Poisson-shifted generalised inverse Gaussian distribution (PSGIG), which includes the Sichel and Delaporte distributions as a special and a limiting

case respectively. Its probability mass function is given by

$$P(Y = y|\mu, \sigma, \nu, \tau) = \frac{\exp\{-\mu\tau\}}{K_\nu(1/\sigma)} T \quad y = 0, 1, 2, \ldots$$
$$T = \sum_{j=0}^{y} \binom{y}{j} \frac{\mu^y \tau^{y-j} K_{\nu+j}(\delta)}{y! d^j (\delta\sigma)^{\nu+j}}. \tag{2.39}$$

The mean and variance for the PSGIG(μ, σ, ν, τ) are given by

$$\mathbb{E}[Y] = \mu,$$
$$\mathbb{V}[Y] = \mu + \mu^2(1-\tau)^2 \left(\frac{2\sigma(\nu+1)}{c} + \frac{1}{c^2} - 1 \right). \tag{2.40}$$

$f_Y(y)$: marginal	$f_\gamma(\gamma)$: mixing distribution
Negative binomial	Gamma
Poisson-inverse Gaussian	Inverse Gaussian
Sichel	Generalised inverse Gaussian
Delaporte	Shifted gamma
PSGIG	Shifted generalised inverse Gaussian
Poisson-Tweedie	Tweedie family
Zero-inflated Poisson	Binary
Zero-inflated negative binomial	Zero-inflated gamma

Table 2.2: Overdispersed count data distributions $f_Y(y)$ along with their mixing distributions $f_\gamma(\gamma)$, where $f(y|\gamma)$ has a Poisson distribution.

2.2 Regression models for count data

Linear model

In classic linear (least squares) regression the usual way of representing the data y_i, for $i = 1, 2, \ldots, n$, as a function of the k covariates $x_{i1}, \ldots, x_{ik}$ is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i \quad i = 1, \ldots, n. \tag{2.41}$$

In matrix form,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.42}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{2.43}$$

and β_i are the unknown parameters that need to be estimated and $\boldsymbol{\epsilon}$ is the random part of the model. One of the assumptions of classic regression is the independence of the errors with each other and with the covariates. In addition, the errors have zero mean and constant variance (homoscedasticity). Applying the zero mean assumption of the errors in equation (2.41),

$$\mathbb{E}(y_i|\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \tag{2.44}$$

where $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})^\intercal$. Least squares regression describes the behaviour of the location of the conditional distribution using the mean of the distribution to represent its central tendency. The residuals $\hat{\epsilon}_i$ are defined as

the differences between the observed and the estimated values. Minimising the sum of the squared residuals

$$\sum_{i=1}^{n} r(y_i - \boldsymbol{x}_i^\intercal \hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\intercal \hat{\boldsymbol{\beta}})^2 \tag{2.45}$$

where $r(u) = u^2$ is the quadratic loss function, gives the least squares estimator $\hat{\boldsymbol{\beta}}$ by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}. \tag{2.46}$$

The additional assumption that the errors $\boldsymbol{\epsilon}$ follow a Gaussian distribution,

$$\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n) \tag{2.47}$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix, provides a framework for testing the significance of the coefficients found in (2.46). Under this assumption the least-squares estimator is also the maximum-likelihood estimator. Taking expectations, with respect to $\boldsymbol{\epsilon}$, in equations (2.42) and (2.47) and by noting that a linear function of a normally distributed random variable is normally distributed itself we can rewrite the model in (2.42) as

$$\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n), \text{ where } \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}. \tag{2.48}$$

The model in (2.48) models the relationship between the mean of y_i, for $i = 1, 2, \ldots, n$, and the covariates linearly.

Generalised linear model

Equation (2.48) refers to data $\boldsymbol{y}$ that are normally distributed but can be generalised to any distribution belonging to the exponential family (Nelder and Wedderburn, 1972). These models are known as generalised linear models (GLM) and consist of three elements.

1. A probability distribution that belongs to the exponential family of distributions ("random component").

2. A linear predictor η_i ("systematic component") such that

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}. \tag{2.49}$$

3. A link function[1] g such that

$$\mathbb{E}[Y_i] = \mu_i = g^{-1}(\eta_i). \tag{2.50}$$

A GLM can be used for data that are not normally distributed and for situations where the relationship between the mean of the response variable and the covariates is not linear. The GLM includes many important distributions such as the Gaussian, Poisson, gamma and inverse Gaussian distributions (Dobson, 2001). The link and mean functions of some common distributions can be seen in Table 2.3.

[1]It is called the link function because it "links" the linear predictor η to the mean of the distribution μ.

Normal distribution $R_y = \mathbb{R}$ $f(y\|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$	$\boldsymbol{x}_i^\intercal \boldsymbol{\beta} = \mu_i$	$\mu_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}$
Exponential distribution $R_y = \mathbb{R}^+$ $f(y\|\lambda) = \lambda \exp\{-\lambda y\}$	$\boldsymbol{x}_i^\intercal \boldsymbol{\beta} = -\mu_i^{-1}$	$\mu_i = -(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^{-1}$
Gamma distribution $R_y = \mathbb{R}^+$ $P(Y = y\|a,b) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp\{-by\}$	$\boldsymbol{x}_i^\intercal \boldsymbol{\beta} = -\mu_i^{-1}$	$\mu_i = -(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^{-1}$
Poisson distribution $R_Y = \mathbb{N}_0$ $P(Y = y\|\mu) = \exp\{-\mu\}\frac{\mu^y}{y!}$	$\boldsymbol{x}_i^\intercal \boldsymbol{\beta} = \log\{\mu_i\}$	$\mu_i = \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}\}$
Inverse Gaussian distribution $R_y = \mathbb{R}^+$ $f(y\|\lambda,\mu) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left\{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right\}$	$\boldsymbol{x}_i^\intercal \boldsymbol{\beta} = -\mu_i^{-2}$	$\mu_i = -(\boldsymbol{x}_i^\intercal \boldsymbol{\beta})^{-\frac{1}{2}}$
Bernoulli distribution $R_y = \{0,1\}$ $P(0) = 1-\theta, P(1) = \theta$	$\boldsymbol{x}_i^\intercal \boldsymbol{\beta} = \log\left\{\frac{\mu_i}{1-\mu_i}\right\}$	$\mu_i = \frac{\exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}\}}$

Table 2.3: Link and mean functions of some common distributions.

Generalised additive model

Generalised additive models (Hastie and Tibshirani, 1986) are an extension to GLM's. The relationship between the linear predictor η_i and the covariates in a generalised additive model (GAM) is not restricted to be linear. For a GAM model we have

$$\eta_i = \beta_0 + \beta_1 f_1(x_{i1}) + \ldots + \beta_k f_k(x_{ik}), \tag{2.51}$$

instead of equation (2.49) which applies for a GLM. The unknown functions f_i are not restricted to have a specific parametric form (e.g. polynomial) and thus are flexible enough to explore any relationship between a covariate and (the mean of) the response variable. Instead of estimating single parameters (like the regression coefficients in a GLM), in generalised additive models, we find a nonparametric function that relates the mean of the response variable to the covariates. For a more detailed description of the generalised additive model see Hastie and Tibshirani (1990); Wood (2006).

2.2.1 Poisson regression

The Poisson regression model, which is a special case of the GLM, is the most common model used for count data. It has been used for modelling count data in many fields such as insurance (e.g. number of insurance claims (Heller et al., 2007)), public health (e.g. number of doctor visits (Winkelmann, 2004)), epidemiology (e.g. number of cancer incidences (Romundstad et al., 2001)), psychology (e.g. number of cases of substance abuse (Gagnon et al., 2008)), and many other research areas.

This model can be specified as

$$P(Y_i = y_i|\mu_i) = \exp\{-\mu_i\}\left(\frac{\mu_i^{y_i}}{y_i!}\right),$$
$$\log\{\mu_i\} = \boldsymbol{x}_i^\intercal\boldsymbol{\beta}, \tag{2.52}$$

with mean and variance

$$\mathbb{E}[Y_i] = \exp\{\boldsymbol{x}_i^\intercal\boldsymbol{\beta}\},$$
$$\mathbb{V}[Y_i] = \exp\{\boldsymbol{x}_i^\intercal\boldsymbol{\beta}\}. \tag{2.53}$$

Its assumption that the variance must be equal to the mean poses a problem when the data exhibit a different behaviour. Most of the proposed approaches to this problem focus on overdispersion (Del Castillo and Pérez-Casany, 1998; Ismail and Jemain, 2007).

2.2.2 Negative binomial regression

One way to handle this situation is to fit a parametric model that is more dispersed than the Poisson. A natural choice is the negative binomial. In this model

$$P(Y_i = y_i|\mu_i, k) = \frac{\Gamma\left(\frac{1}{k} + y_i\right)}{\Gamma\left(\frac{1}{k}\right) y_i!}\left(\frac{k\mu_i}{1+k\mu_i}\right)^{y_i}\left(\frac{1}{1+k\mu_i}\right)^{\frac{1}{k}},$$
$$\log\{\mu_i\} = \boldsymbol{x}_i^\intercal\boldsymbol{\beta}, \tag{2.54}$$

where the parameters μ_i and k represent the mean and the dispersion of the negative binomial. For this model, the mean and variance are

$$\mathbb{E}[Y_i] = \exp\{\boldsymbol{x}_i^\intercal\boldsymbol{\beta}\},$$
$$\mathbb{V}[Y_i] = \exp\{\boldsymbol{x}_i^\intercal\boldsymbol{\beta}\} + k\exp\{\boldsymbol{x}_i^\intercal\boldsymbol{\beta}\}^2. \tag{2.55}$$

The variance of a negative binomial model is a quadratic function of its mean. The negative binomial model approaches the Poisson(μ_i) model for $k \to 0$. For an extensive description on the negative binomial regression model see Hilbe (2007).

2.2.3 COM-Poisson regression

Sellers and Shmueli (2010) propose a COM-Poisson regression model based on the (λ, ν) formulation whereas Guikema and Coffelt (2008) propose a COM-Poisson generalised linear model based on the (μ, ν) reformulation; both formulations can be seen in Section 2.1.3. Modifying the latter model we have

$$\begin{aligned} P(Y_i = y_i|\mu_i, \nu_i) &= \left(\frac{\mu_i^{y_i}}{y_i!}\right)^{\nu_i} \frac{1}{Z(\mu_i, \nu_i)}, \\ Z(\mu_i, \nu_i) &= \sum_{j=0}^{\infty} \left(\frac{\mu_i^j}{j!}\right)^{\nu_i}, \\ \log\{\mu_i\} &= \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, \\ \log\{\nu_i\} &= -\boldsymbol{x}_i^\intercal \boldsymbol{c}. \end{aligned} \tag{2.56}$$

where Y is the dependent random variable being modelled, and $\boldsymbol{\beta}$ and $\boldsymbol{c}$ are the regression coefficients for the centering link function and the shape link function.

The mean and variance for the COM-Poisson model are approximated by

$$\mathbb{E}[Y_i] \approx \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}\},$$
$$\mathbb{V}[Y_i] \approx \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \boldsymbol{x}_i^\intercal \boldsymbol{c}\}. \tag{2.57}$$

The flexibility of this model can be seen by looking at the right hand side of (2.57). In this model larger values of $\boldsymbol{\beta}$ and $\boldsymbol{c}$ can be translated to higher mean and higher variance, respectively, for the response variable. The above approximations (especially the one for the mean) are really good when μ is large and ν is small. A better approximation for the mean of the COM-Poisson can be seen in page 20.

Sellers et al. (2012) present an overview of the different research areas in which the COM-Poisson model has been used. These include biology (Ridout and Besbeas, 2004), marketing (Kalyanam et al., 2007), and transportation (Lord et al., 2008) amongst others. More information on the COM-Poisson distribution and the COM-Poisson regression model is provided by Shmueli et al. (2005); Kadane et al. (2006); Sellers and Shmueli (2013).

2.2.4 Other regression models

Rigby and Stasinopoulos (2001, 2005) introduced the generalised additive models for location, scale, and shape (GAMLSS) as *semi-parametric* regression type models. They are *parametric*, in that they require a parametric distribution assumption for the response variable, and *semi* in the sense that the modelling of the parameters of the distribution, as functions of explanatory variables, may involve using nonparametric smoothing functions. They

overcome some of the limitations of the generalised linear model (GLM) and generalised additive model (GAM). In GAMLSS the exponential family distribution assumption for the response variable Y is relaxed and replaced by a general distribution family, including highly skew and/or kurtotic continuous and discrete distributions. The systematic part of the model, see (2.49), is expanded to allow modelling not only of the mean (or location) but other parameters of the distribution of Y as, linear and/or non-linear, parametric and/or smooth non-parametric functions of explanatory variables and/or random effects.

A GAMLSS model assumes that, for $i = 1, 2, \ldots, n$, independent and identically distributed observations y_i have probability (density) function $f_Y(y_i|\boldsymbol{\theta}_i)$ conditional on

$$\begin{aligned} \boldsymbol{\theta}_i &= (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) \\ &= (\mu_i, \sigma_i, \nu_i, \tau_i) \end{aligned} \tag{2.58}$$

which is a vector of four *distribution parameters*, each of which can be a function of the covariates. The first two population parameters μ_i and σ_i are characterised as location and scale parameters, while the remaining parameter(s) are characterised as shape parameters, e.g. skewness and kurtosis

parameters. Analogous to equation (2.49), a GAMLSS model is defined as

$$\begin{aligned} g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\gamma_{j1}, \\ g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 &= \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\gamma_{j2}, \\ g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 &= \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=3}^{J_3} \boldsymbol{Z}_{j3}\gamma_{j3}, \\ g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 &= \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=4}^{J_4} \boldsymbol{Z}_{j4}\gamma_{j4}, \end{aligned} \tag{2.59}$$

where $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}$ and $\boldsymbol{\eta}_k$ for $k = 1, 2, 3, 4$ are vectors of length n, $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{J_k k})^\intercal$ is a parameter vector of length J_k, $\boldsymbol{X}_k$ is a fixed known design matrix of order $n \times J_k$, Z_{jk} is a fixed known $n \times q_{jk}$ design matrix and γ_{jk} is a q_{jk} dimensional random variable. The model in (2.59) allows the user to model each distribution parameter as a linear function of the covariates and/or as linear functions of the random effects. The GAMLSS models presented in (2.59) is more general than the GLM, GAM, GLMM or GAMM in that all parameters (not just the mean) are modelled in terms of both fixed and random effects and that the distribution of the response variable is not limited to the exponential family. The form of the distribution for the response variable can be very general. More information on the available distributions can be found in Stasinopoulos and Rigby (2007).

If we let $\boldsymbol{Z}_{jk} = \boldsymbol{I}_n$ and $\gamma_{jk} = \boldsymbol{h}_{jk} = h_{jk}(\boldsymbol{x}_{jk})$ for $k = 1, 2, 3, 4$ then we have

the semi-parametric additive formulation of GAMLSS

$$
\begin{aligned}
g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} h_{j1}(\boldsymbol{x}_{j1}),\\
g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 &= \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} h_{j2}(\boldsymbol{x}_{j2}),\\
g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 &= \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=3}^{J_3} h_{j3}(\boldsymbol{x}_{j3}),\\
g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 &= \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=4}^{J_4} h_{j4}(\boldsymbol{x}_{j4}).
\end{aligned} \tag{2.60}
$$

The Sichel, Delaporte, and Poisson shifted generalised inverse Gaussian distribution, in Subsection 2.1.4, are some examples of distributions that can be fitted in the GAMLSSS model structure. For more information on the available distributions for the GAMLSS model see Rigby and Stasinopoulos (2005).

Finally, Sellers et al. (2012) present some distributions with the ability to handle underdispersed and/or overdispersed data. They refer to the regression models from the distributions mentioned in Subsection 2.1.4. A good source of reference for count data regression models is the book of Cameron and Trivedi (2013).

2.3 Quantile regression

The quantile function $Q(p)$ returns the value below which, random variables of the distribution would fall with probability p. A definition that eliminates

Distribution	$F_Y(y)$	$Q_Y(p)$
Exponential (λ)	$1-\exp\{-\lambda x\}$	$-\frac{\log\{1-p\}}{\lambda}$
Pareto (α,β)	$1-(\frac{\alpha}{y})^{\beta}$	$\alpha(1-p)^{-\frac{1}{\beta}}$
Uniform (α,β)	$\frac{y-\alpha}{\beta-\alpha}$	$p(\beta-\alpha)+\alpha$
Logistic (μ,s)	$\frac{1}{1+\exp\{-(y-\mu)/s\}}$	$\mu - s\log\left\{\frac{1-p}{p}\right\}$

Table 2.4: Quantile functions of the exponential, Pareto, uniform and logistic distributions.

the problem of having a non-continuous cumulative distribution function is

$$Q(p) = \inf\{x \in \mathbb{R} : p \leq F(x)\}. \tag{2.61}$$

It returns the value x such that

$$F(x) = P(X \leq x) = p \tag{2.62}$$

where $F(x)$ is the cumulative distribution function (c.d.f.). Quantile regression was first proposed in Koenker and Bassett (1978) as a more robust method to outliers compared to the classic linear regression (least squares regression). It extends the concept of a quantile function to estimating the conditional quantile distributions $Q_Y(p|X = \boldsymbol{x})$ and gives a more complete picture of what the relation between the covariates and the response variable is. In addition, it models location shifts and shape shifts of the conditional distribution. In quantile regression, similar to (2.41) and (2.42),

$$y_i = \beta_{0,p} + \beta_{1,p}x_{i1} + \ldots + \beta_{k,p}x_{ik} + \epsilon_{i,p} \quad i = 1, \ldots, n. \tag{2.63}$$

and in matrix form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_p + \boldsymbol{\epsilon}_p \tag{2.64}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \boldsymbol{\beta}_p = \begin{pmatrix} \beta_{0,p} \\ \beta_{1,p} \\ \vdots \\ \beta_{k,p} \end{pmatrix}, \boldsymbol{\epsilon}_p = \begin{pmatrix} \epsilon_{1,p} \\ \epsilon_{2,p} \\ \vdots \\ \epsilon_{n,p} \end{pmatrix} \tag{2.65}$$

Unlike least squares regression where all inferences depend on the estimated parameter $\hat{\boldsymbol{\beta}}$, quantile regression allows us to be more precise since the estimated parameters $\hat{\boldsymbol{\beta}}_p$ are not constant across the conditional distribution but depend on the quantile p.

Analogous to the equations of least squares regression,

$$\begin{aligned} \mathbb{E}[\epsilon_i | X_i = \boldsymbol{x}_i] &= 0, \\ \mathbb{E}[y_i | X_i = \boldsymbol{x}_i] &= \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, \end{aligned} \tag{2.66}$$

in quantile regression

$$\begin{aligned} Q_Y(\epsilon_{i,p} | X_i = \boldsymbol{x}_i) &= 0, \\ Q_Y(p | X_i = \boldsymbol{x}_i) &= \boldsymbol{x}_i^\intercal \boldsymbol{\beta}_p. \end{aligned} \tag{2.67}$$

Estimating the parameter $\hat{\boldsymbol{\beta}}_p$ requires minimising the sum of the absolute residuals

$$\hat{\boldsymbol{\beta}}_p = \underset{\boldsymbol{\beta}_p \in \mathbb{R}^k}{\arg\min} \sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x}_i^\intercal \boldsymbol{\beta}_p), \tag{2.68}$$

where

$$\begin{aligned} \rho_p(u) &= \left(p I_{[0,\infty)}(u) + (1-p) I_{(-\infty,0)}\right) |u|, \\ &= \left(p - I_{(-\infty,0)}\right) u. \end{aligned} \tag{2.69}$$

is the absolute loss functionThis function is also known as check function.

In general, a closed form solution does not exist since the check function is not differentiable at the origin. Rewriting $\boldsymbol{y}$ as a function of only positive elements

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta}_p + \boldsymbol{\epsilon}_p \\ &= \boldsymbol{X}(\boldsymbol{\beta}_{1,p} - \boldsymbol{\beta}_{2,p}) + (\boldsymbol{\epsilon}_{1,p} - \boldsymbol{\epsilon}_{2,p}), \end{aligned} \tag{2.70}$$

where $\boldsymbol{\beta}_{1,p}, \boldsymbol{\beta}_{2,p}, \boldsymbol{\epsilon}_{1,p}, \boldsymbol{\epsilon}_{2,p} \geq 0$ and setting

$$\begin{aligned} A &= (X, -X, I_n, -I_n), \\ \boldsymbol{z} &= (\boldsymbol{\beta}_{1,p}^\intercal, \boldsymbol{\beta}_{2,p}^\intercal, \boldsymbol{\epsilon}_{1,p}^\intercal, \boldsymbol{\epsilon}_{2,p}^\intercal), \\ \boldsymbol{c} &= (\mathbf{0}^\intercal, \mathbf{0}^\intercal, p \times \boldsymbol{l}^\intercal, (1-p) \times \boldsymbol{l}^\intercal)^\intercal. \end{aligned} \tag{2.71}$$

where I_n is a n dimensional identity matrix, $\mathbf{0}^\intercal$ is a $k \times 1$ vector of zeros and $\boldsymbol{l}$ is a $n \times 1$ vector of ones reduces the previous problem to

$$\min_{\boldsymbol{z}} \boldsymbol{c}^\intercal \boldsymbol{z} \text{ subject to: } A\boldsymbol{z} = Y \tag{2.72}$$

which can be solved using linear programming techniques. If the design matrix $\boldsymbol{X}$ is of full column rank there exists a solution for the above problem. For more information see Buchinsky (1998); Schulze (2004).

An application of quantile regression can be seen in Koenker and Hallock (2001) in which they present data for 235 European working-class households and model the relationship between food expenditure and income. This relationship is known as an Engel curve. Engel curves are also used to describe how the demanded quantity of a good or service changes as the consumer's income level changes, and they are used for tax policy and measuring inflation among other things.

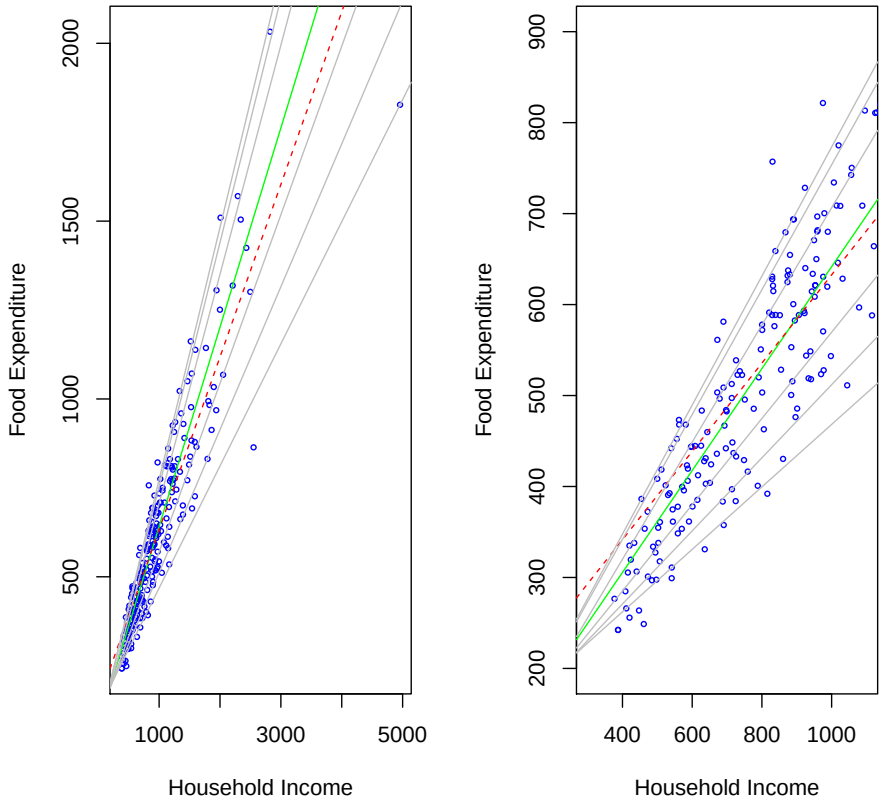


Figure 2.1: Quantile regression lines for $p \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$ are all in gray except the median that is in green. Mean regression line is the dashed line in red. The right-hand panel shows an enlarged version of the bottom-left part of the plot in the left panel.

Figure 2.1 shows seven quantile regression lines for different values of $p \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$, where the median is indicated by the darker solid line while the least squares estimate of the conditional mean function is indicated by the dashed line. The last two are quite different due to the non-robustness of least squares. There are two households with high income and low food expenditure that drive the conditional mean downwards. As a result, the mean regression line is a very poor estimate on the poorest households since the conditional mean is above almost all of them. The spacing of the quantile regression lines reveals that the conditional distribution of food expenditure is skewed to the left due to the narrower spacing of the upper quantiles that indicates high density and a short upper tail.

Another application of quantile regression comes from a cross-sectional study that measures growth and development of the Dutch population between the ages of 0 and 21 years (Buuren and Fredriks, 2001). Focusing on the Body Mass Index (BMI) of 7294 boys, Figure 2.2 shows that the tails and the mean of the conditional distributions vary differently with age. The BMI is defined as the individual's body mass divided by the square of his height. Its measure is kg/m^2. A person is considered to be healthy weight when his BMI is between 18.5 and 25. Conditional quantiles are important because they provide the basis for developing growth charts and establishing health standards. The rate of change with age, particularly for ages less than 10, is different for each conditional quantile. For higher quantiles there is a drop in BMI until the age of 5 and then it rises up to 25 (close to being overweight). The BMI of underweight people seem to not increase as much as those with healthy weight. A general review of quantile regression and its application

areas can be found in Yu et al. (2003).

Quantile regression for count data

When being applied to discrete distributions, such as those arising from count data, quantile regression methods need to be "tweaked" to deal with the fact the quantiles are not continuous any more. Machado and Santos Silva (2005) introduce a continuous auxiliary variable $Z = Y + U$ where Y is the count variable and U is a uniform random variable in the interval $[0, 1)$. The density of the new variable is

$$f(z) = \begin{cases} p_0 & \text{if } 0 \leq z < 1, \\ p_1 & \text{if } 1 \leq z < 2, \\ \vdots & \vdots \end{cases} \tag{2.73}$$

and the cumulative distribution function is

$$F(z) = \begin{cases} p_0 z & \text{if } 0 \leq z < 1, \\ p_0 + p_1(z-1) & \text{if } 1 \leq z < 2, \\ \vdots & \vdots \end{cases} \tag{2.74}$$

where $p_i = P(Y = i)$.

As a result, the quantiles of Z are given by

$$Q_Z(p) = \begin{cases} \frac{p}{p_0} & \text{for } p < p_0, \\ 1 + \frac{p-p_0}{p_1} & \text{for } p_0 \leq p < p_0 + p_1, \\ \vdots & \vdots \end{cases} \tag{2.75}$$

From (2.75) we can see that the lower bound of the quantiles of Z is p. The conditional quantiles $Q_Z(p|X_i = \boldsymbol{x}_i)$ are then specified as

$$Q_Z(p|X_i = \boldsymbol{x}_i) = p + \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}_p\} \tag{2.76}$$

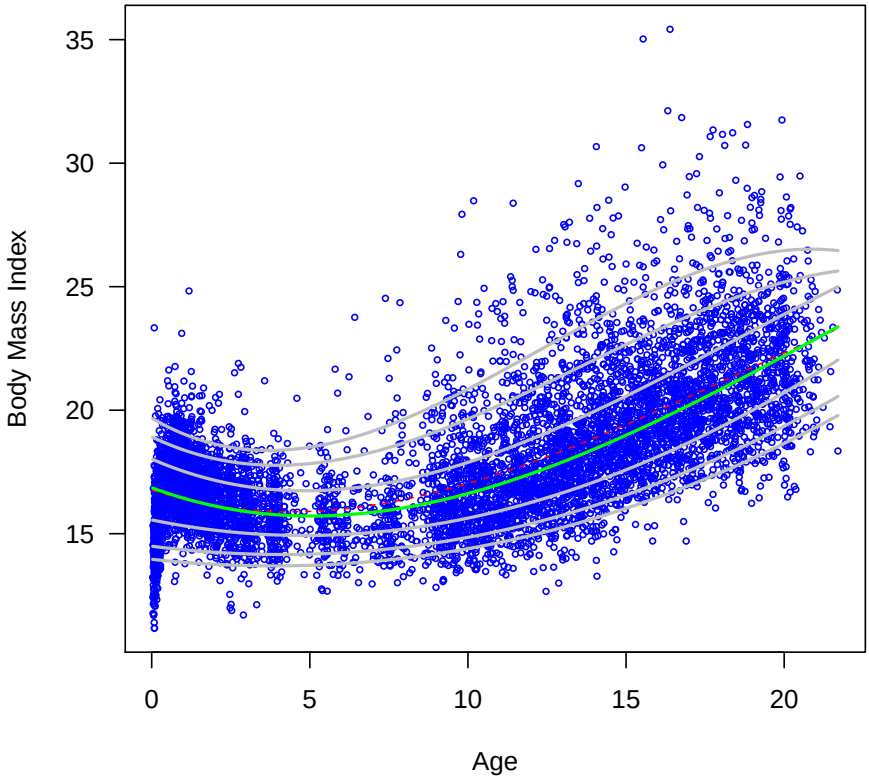


Figure 2.2: Quantile regression curves for $p \in \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$ are all in gray except the median that is in green. Mean regression curve is the dashed line in red.

where p is added in order to reflect the lower bound (2.75). Afterwards, the variable Z is transformed in such a way that the new quantile function is linear in the parameters

$$Q_{T(Z;p)}(p|X_i = \boldsymbol{x}_i) = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}_p \tag{2.77}$$

where

$$T(Z;p) = \begin{cases} \log\{Z - p\} & \text{for } Z > p, \\ \log\{\varsigma\} & \text{for } Z \leq p, \end{cases} \tag{2.78}$$

with ς being a small positive number. The parameters $\boldsymbol{\beta}_p$ are estimated by running a quantile regression of $T(Z;p)$ on $\boldsymbol{x}_i$. This is allowed, since quantiles are invariant to monotonic transformations and to censoring from below up to the quantile of interest (see Powell, 1986, 1991; Neocleous and Portnoy, 2008). The conditional quantiles of interest, $Q_Y(p|X_i = \boldsymbol{x}_i)$ can be found from the previous quantiles

$$Q_Y(p|X_i = \boldsymbol{x}_i) = \lceil Q_Z(p|X_i = \boldsymbol{x}_i) - 1 \rceil \tag{2.79}$$

where $\lceil p \rceil$ denotes the ceiling function which returns the smallest integer greater than, or equal to, p. Finally the quantiles of Z are found through

$$Q_Z(p|X_i = \boldsymbol{x}_i) = T^{-1}(Q_{T(Z;p)}(p|X_i = \boldsymbol{x}_i)). \tag{2.80}$$

While the jittering approach eliminates the problem of a discrete response distribution, for small values of the response variable Y, the mean and the variance in the transformed variable Z will be mainly due to the added noise, resulting in poor estimates of the conditional quantiles $Q_Y(p|X_i = \boldsymbol{x}_i)$. As an example, when $Y = 0$ the term $\log\{Z - p\} = \log\{U - p\}$ could go from $-\infty$ to 0, simply due to the added noise.

2.4 Mixture models

2.4.1 Finite mixture model

Mixture models are widely used for density estimation and clustering. The idea behind mixture models is that the observed data cannot be characterised by a simple distribution but instead by several; for each observation one of these distributions is selected at random. In a sense we treat a population as if it is made from several subpopulations. We can apply these models to data where the observations come from various groups and the group members are not known, but also to provide approximations for multimodal distributions. The general form of a finite mixture model with K groups is

$$f(y) = \sum_{j=1}^{K} p_j f(y|\theta_j) \tag{2.81}$$

where p_i are the weights and $f(y|\theta_i)$ are the probability distribution functions for each group. This can also be viewed as

$$\begin{aligned} c_i &\sim \text{Mult}(1, \boldsymbol{\pi}), \\ Y_i &\sim F(|\theta_{c_i}). \end{aligned}$$

where $\boldsymbol{\pi} = (p_1, \ldots, p_K)$. Introducing prior distributions on $\boldsymbol{\pi}$ and $\theta_1, \ldots, \theta_K$,

$$\begin{aligned} \boldsymbol{\pi} &\sim \text{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right), \\ c_i &\sim \text{Mult}(1, \boldsymbol{\pi}), \\ \theta_1, \ldots, \theta_K &\sim G_0, \\ Y_i &\sim F(|\theta_{c_i}). \end{aligned} \tag{2.82}$$

After drawing the weights for each of the K mixture components, each observation gets allocated to a mixture component. The latent variable c_i indicates the cluster to which the i^{th} observation belongs. We then draw parameters for each mixture component from a distribution G_0 and think of the observations Y_i as coming from a distribution with parameters according to the cluster observation i belongs to.

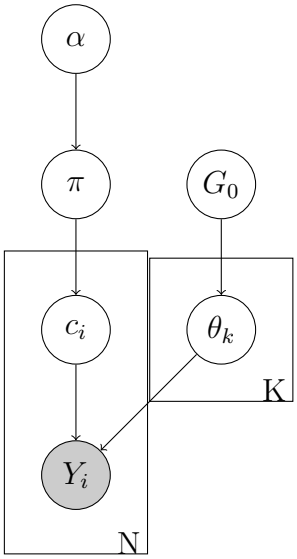


Figure 2.3: Graphical representation of the finite mixture model.

Alternative method

The standard frequentist method used to fit finite mixture models is the expectation-maximisation (EM) algorithm (Dempster et al., 1977) which converges to a zero-gradient point of the likelihood of the mixture parameters. The EM algorithm interprets the data Y as *incomplete* and assumes that there is a missing auxiliary variable, the group assignments. The *complete* likelihood function has a less complicated form and can be easily maximised.

The algorithm is an iterative procedure which alternates between two steps: the expectation step in which it computes the conditional expectation of the complete log-likelihood given the data Y and the current estimate of the parameters $\hat{\boldsymbol{\theta}}$, and the *maximisation* step in which it computes the estimates that maximise the expected log-likelihood found in the previous *expectation* step. The key property of the algorithm is that the *incomplete* log-likelihood increases after each iteration. There are drawbacks connected with the EM algorithm. These refer to slow convergence, choice of initial values in order to reach the global maximum in fewer iterations, and the choice of a suitable stopping rule which detects that the algorithm has reached its global maximum. For more information on how to deal with these problems of the EM see Pilla and Lindsay (2001); Biernacki et al. (2003); Karlis and Xekalaki (2003)

Titterington et al. (1985); Lindsay (1995); McLachlan and Peel (2004); Frühwirth-Schnatter (2006) provide comprehensive information on the history, applications, and theory of mixture models.

Parametric models that use a fixed number of mixture components can suffer, depending on the number of mixture components used, from under- or overfitting of data. As a result, model selection becomes both important and difficult. The nonparametric approach is to use a model with unbounded complexity. This can be seen as letting $K \to \infty$ in equation (2.82). If the parameters θ_k and mixture proportions $\boldsymbol{\pi}$ are integrated out, the only variables left are the latent variables c_i that do not grow with K. It can be shown that the number of components used to model n datapoints is approximately $(O(a \log(n))$. At most, n components will be associated with the data ("ac-

tive" clusters)[2] so the model stays well defined even when $K \to \infty$. Using this model eliminates the need to estimate the number of mixture components and is a very popular method especially when the number of components grows without bound as the amount of data grows.

2.4.2 Dirichlet distribution

The Dirichlet distribution is a multivariate generalisation of the beta distribution and is used as a prior over probability distributions over finite space. The probability density function of the Dirichlet distribution is

$$f(\pi_1, \ldots, \pi_{k-1}, a_1, \ldots, a_k) = \frac{1}{B(a)} \prod_{i=1}^{k} \pi_i^{a_i - 1} \qquad \pi_1, \ldots, \pi_k > 0 \tag{2.83}$$

where

$$\begin{aligned} \sum_{i=1}^{k} \pi_i &= 1, \\ B(a) &= \prod_{i=1}^{k} \frac{\Gamma(a_i)}{\Gamma(\sum_{i=1}^{k} a_i)}, \\ a &= (a_1, \ldots, a_k), \ a_i > 0. \end{aligned} \tag{2.84}$$

Analogous to the beta distribution being the conjugate prior for the binomial distribution, the Dirichlet distribution is the conjugate prior for the multinomial distribution.

[2] Imagine having $n = 1000$ number of observations. In this case, the number of clusters c_i could go up to 1000.

2.4.3 Dirichlet process

A Dirichlet process (Ferguson, 1973; Antoniak, 1974) can be seen as a infinite dimensional Dirichlet distribution. In the same way as the Gaussian process is a distribution over functions the Dirichlet process (DP) is a distribution over distributions meaning that every draw from a Dirichlet process is a distribution. One of the reasons why Gaussian processes are so popular in regression and classification is because they do not restrict the data to a specific model. In a similar way, using Dirichlet processes bypasses the need to estimate the "correct" number of components in a mixture model.

A Dirichlet process is a common choice for a prior of an unknown distribution over infinite space. Instead of choosing a prior with a specific parametric form, we incorporate the "uncertainty" about the prior distribution by sampling from a class of distributions. Thus, by relaxing the parametric assumptions of the prior, we overcome the restrictions that those assumptions have on the observed data. These draws are discrete probability distributions over infinite sample space that can be written as an infinite sum of atoms.

The original definition of Ferguson (1973) defines the Dirichlet process using partitions of the sample space Ω. If the distribution of G is a draw from the Dirichlet process then for any partition of Ω of the form $(B_1, B_2, \ldots, B_k)$ the vector of associated probabilities has a Dirichlet distribution, i.e.

$$(G(B_1), \ldots, G(B_k)) \sim \text{Dirichlet}(\alpha G_0(B_1), \ldots, \alpha G_0(B_k)) \tag{2.85}$$

where α can be thought of as the precision or concentration parameter and

G_0 the mean of the Dirichlet process (known as the base measure), since

$$\mathbb{E}[G(B)] = G_0(B), \quad \mathbb{V}[G(B)] = \frac{G_0(B)(1 - G_0(B))}{1 + \alpha}. \tag{2.86}$$

An interesting feature of the Dirichlet process is that even if the base measure G_0 is continuous, G is almost surely discrete. Thus if we draw a sample $\theta_1, \ldots, \theta_n$ from G the sample might exhibit ties, i.e. there is non-zero probability that the same value is drawn more than once. If G is continuous the marginal distribution of the θ_i is continuous, but their joint distribution is not. If we let $\alpha \to 0$ then $\theta_1, \ldots, \theta_n$ are identical with probability one (one cluster). If we let $\alpha \to +\infty$ then the θ_i are i.i.d. draws from G_0 (n clusters).

Because of the discrete "spiky" nature of the joint distribution one typically does not use a Dirichlet process to model the response variable of interest itself, but to model the parameters of the distribution of the response variable. This can been viewed as employing a kernel to "smooth" the spiky density. This will be discussed in more detail in Section 2.4.4.

We can view the Dirichlet process as a limiting case of the finite mixture model (2.82). If $\alpha_1 = \ldots = \alpha_K$ (equation (2.84)) and $K \to +\infty$ and $\alpha_1 \to 0$ such that $\alpha_1 K \to \alpha > 0$ then the joint distribution of the $\theta_{c_1}, \ldots, \theta_{c_n}$ tends to the same distribution as obtained from the Dirichlet process.

"Stick-breaking" process

A way of visualising the previous statement, is with the "stick-breaking" process (Sethuraman, 1991). Thinking of a draw G from a Dirichlet Process

as an infinite sum of point masses

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}, \tag{2.87}$$

where δ_{θ_k} is the Dirac delta function taking the value ∞ at θ_k and 0 everywhere else. Each θ_k is a draw from the base distribution G_0 and β_k follows a beta distribution

$$\theta_k \sim G_0, \quad \beta_k \sim \text{beta}(1, \alpha), \tag{2.88}$$

and the mixing proportions of each component are found using

$$\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i). \tag{2.89}$$

The reason why it is called "stick-breaking" is because it is like having a stick of length 1, breaking it at β_1, and then assigning π_1 to be the length of stick we just broke off. Now recursively break the other portion to obtain π_2, π_3 and so forth. Figure 2.4 shows a random draw for the probability weights (of each possible cluster) for different values of α for 100 observations. It can be seen that as α increases, the number of mixture components with non-zero weights tends to be higher.

2.4.4 Dirichlet process mixture model

We can think of a Dirichlet process mixture model as

$$\begin{aligned} G &\sim \text{DP}(\alpha, G_0), \\ \theta_i | G &\sim G, \\ Y_i | \theta &\sim F(|\theta_i), \end{aligned} \tag{2.90}$$

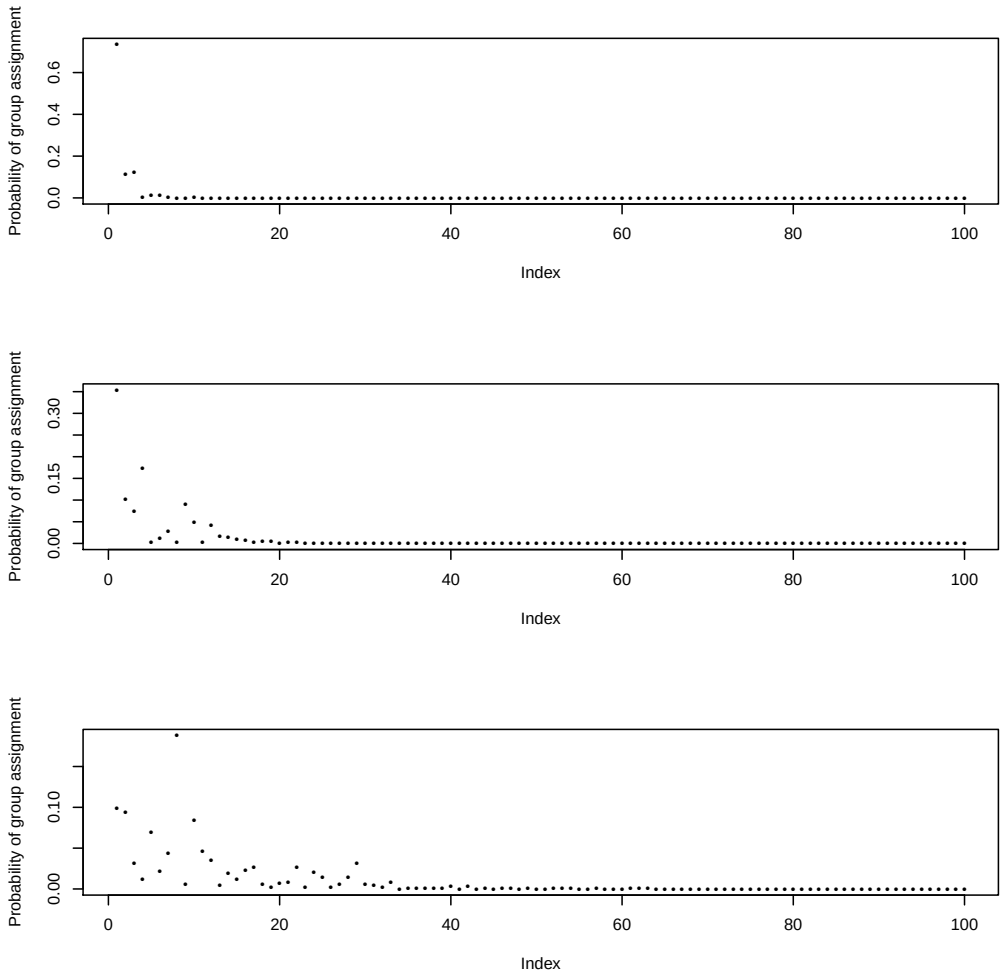


Figure 2.4: One random draw for the probability weights for 100 observations for $\alpha = 1, 5, 10$.

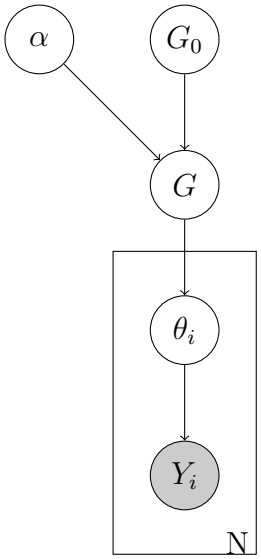


Figure 2.5: Graphical representation of the Dirichlet process mixture model.

where G_0 could be a Gaussian distribution and G an infinite pair of θ_k's that each one consists of a mean and a standard deviation. Ferguson (1983) points out that these models can be seen as countably infinite mixtures since realisations of the Dirichlet process are discrete with probability one. For a large enough sequence of draws from G the value of any draw will be repeated by another one. This is equivalent to saying that when we keep drawing the same parameters, the observations that have those parameters belong to the same cluster. In other words, the discreteness of the Dirichlet process gives the Dirichlet process mixture model its clustering property. If $\theta_1^*, \ldots, \theta_m^*$ are the unique values among all the θ_i and n_k are the number of repeats of θ_k^* parameters the predictive distribution can be written as

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n}[\alpha G_0 + \sum_{k=1}^{m} n_k \delta_{\theta_k^*}]. \tag{2.91}$$

The probability of observing a repeated value is proportional to the number

of times it is observed. Clusters with many observations will tend to get bigger. Equation (2.91) relies on the exchangeability of the observations.[3] Equation (2.91) is an important property, it is the basis of Gibbs sampling algorithms for Dirichlet process mixture models and gives these a intuitive interpretation, discussed in the next section.

Chinese restaurant process

The predictive distribution given in equation (2.91) leads to the popular interpretation of the Dirichlet process as a basic form of the Chinese restaurant process (CRP). The basic CRP is a process in which n customers sit down in a Chinese restaurant with an infinite number of tables (with infinite capacity). Then

- the first customer sits at the first table.
- The m^{th} customer has two options. He either sits at a table already occupied with probability proportional to the people sitting there or he sits at a new table with probability proportional to α.

The above procedure defines an exchangeable distribution over the partition of the customers but also over the permutations of the customers. As more customers come in, more tables become occupied but as expected from the definition of the CRP, large enough tables tend to grow larger faster[4]. Thinking of the customers as the data and the tables as the clusters we can associate

[3] Observations are called exchangeable when the joint probability distribution is invariant under any permutation.

[4] Also known as the "rich get richer" phenomenon.

the Chinese restaurant process to the Dirichlet process mixture models. The CRP view also is the basis for generalisations of Dirichlet process mixture models.

Escobar (1994) used a Dirichlet process prior to provide a nonparametric Bayesian estimate of a vector of Gaussian means. This method is based on Gibbs sampling and can be easily implemented for models based on conjugate prior distributions. For non-conjugate priors, straightforward Gibbs sampling requires the estimation of a (usually) difficult numerical integration. West (1994) used a Monte Carlo approximation to this integral. MacEachern and Müller (1998) proposed an exact approach for handling non-conjugate priors which uses a mapping from a set of auxiliary parameters to the set of parameters in use. Walker and Damien (1998) applied a different auxiliary variable method which still requires the computation of an integral. Neal (2000) reviews all the above methods and proposes a new auxiliary variable method that improves on the previous ones by Walker and Damien (1998) and MacEachern and Müller (1998).

Two short introductions to the Dirichlet process and the Dirichlet process mixture model can be found in Orbanz and Teh (2010); Teh (2010). For more information refer to MacEachern (1994); Escobar and West (1994); Frigyik et al. (2010); Jara et al. (2011).

2.4.5 Flexibility of the COM-Poisson mixture model

As we have already mentioned we will use a mixture of COM-Poisson distributions to approximate the unknown probability density. This allows us to

estimate any probability density no matter its complexity.

Approximating underdispersed distributions

An advantage of the COM-Poisson regression model is in the way it captures underdispersion. We illustrate this advantage by trying to approximate a binomial distribution with large mean and small variance. Assume that the number of trials of the binomial distribution is large ($n = 60$) and the probability of success is high ($p = 0.99$). In this case the mean and variance of the distribution we are trying to approximate are

$$\begin{aligned} \mathbb{E}[Y] &= 59.4, \\ \mathbb{V}[Y] &= 0.594. \end{aligned} \tag{2.92}$$

We will approximate this distribution with a Poisson, a COM-Poisson and a mixture of COM-Poisson distributions. In both cases, the COM-Poisson and Poisson approximations that minimise the Kullback-Leibler divergence are plotted. Figure 2.6 shows the "best" approximations for the binomial distribution (plotted in black). By "best" approximations we mean the ones that minimise the Kullback-Leibler (KL) divergence which is defined as

$$\mathrm{KL}(P, Q) = \sum_i \log\left\{\frac{P(i)}{Q(i)}\right\} P(i). \tag{2.93}$$

where P is the true distribution and Q is an approximation of P.

Approximating overdispersed dstributions

In the case of approximating overdispersed distributions, such as the geometric, Figure 2.7 shows that the COM-Poisson distribution (plotted in red) does well enough by itself and there is no need for using more than one as an approximation of a geometric (plotted in black). On the other hand, for small values of the probability parameter p of the geometric, a single Poisson distribution (plotted in green) cannot approximate accurately the geometric. In that case one needs a mixture of Poisson or negative binomial distributions.

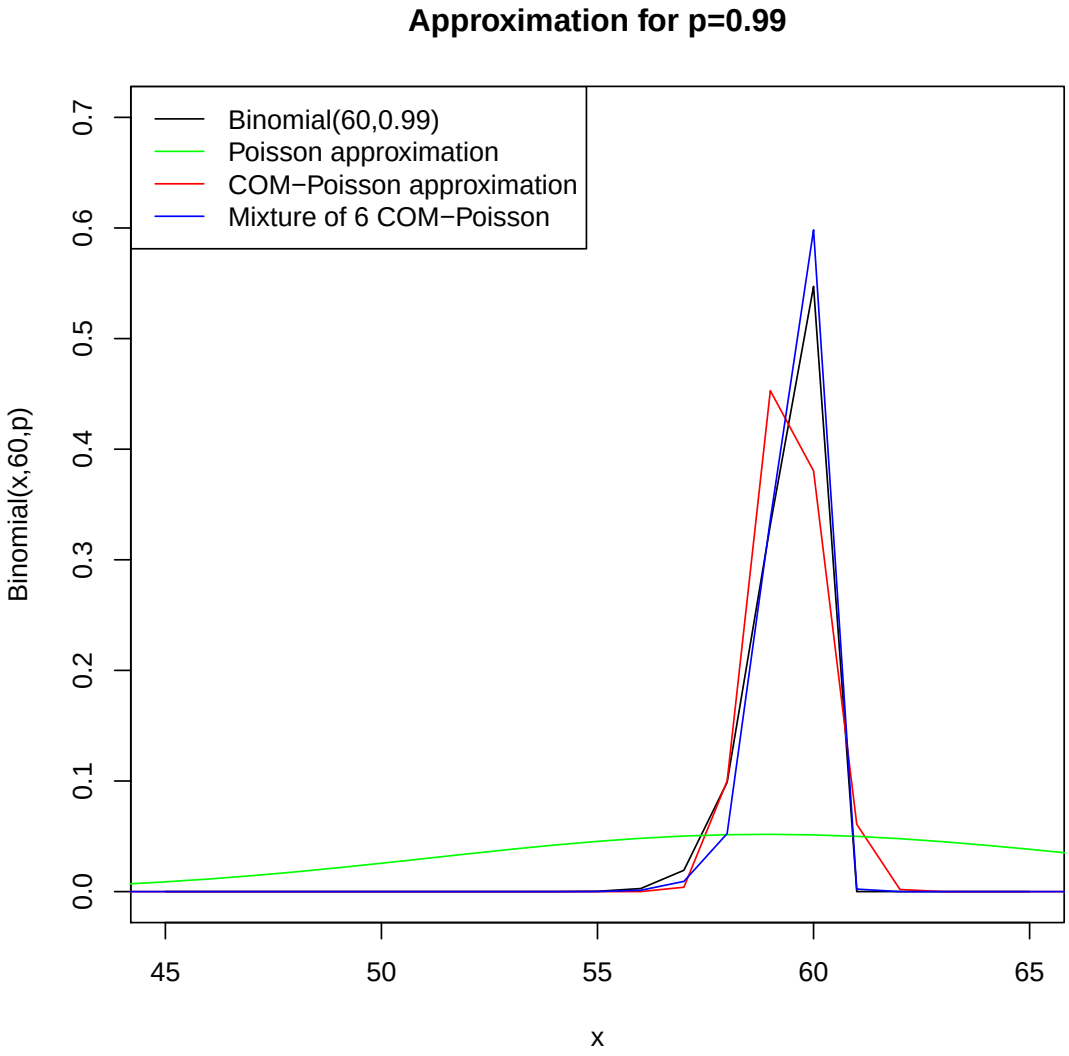


Figure 2.6: Approximating a binomial with large mean and small variance with a Poisson (green line), a COM-Poisson (red line) and a mixture of COM-Poisson distributions (blue line).

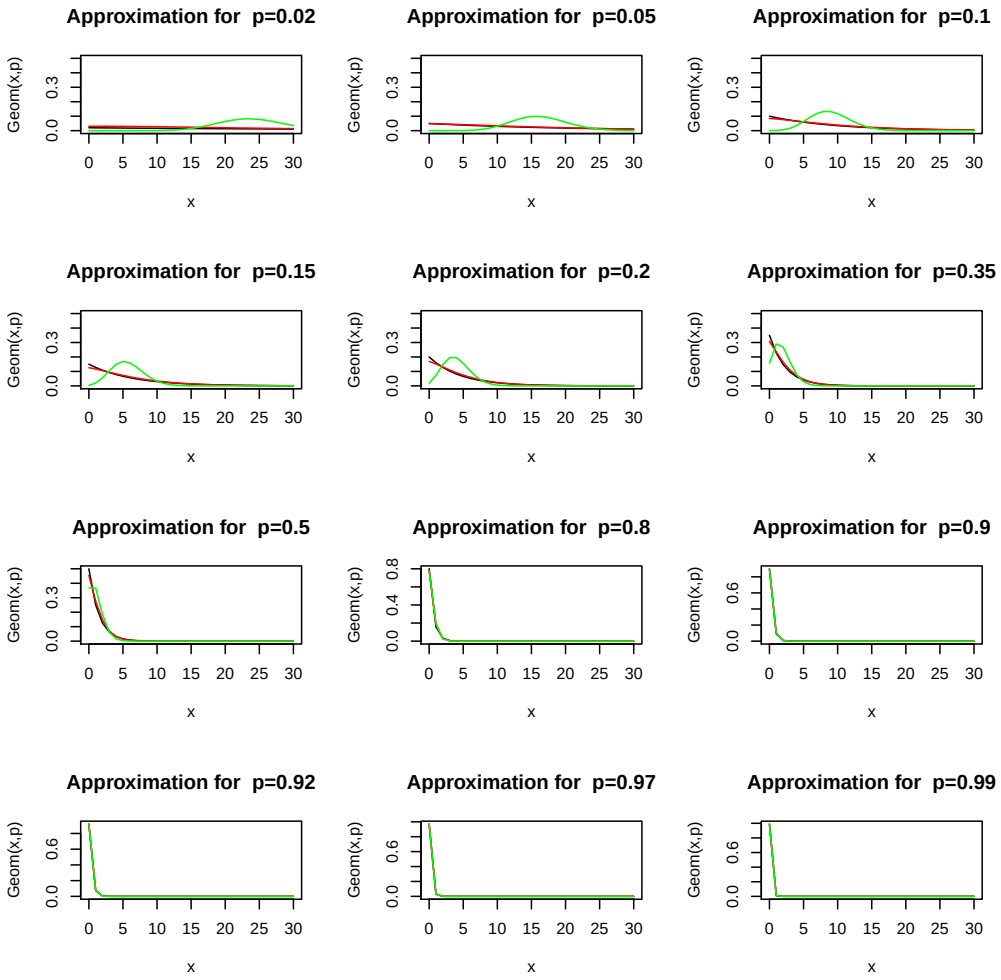


Figure 2.7: Approximating a geometric distribution (in black) with a Poisson (in green), and a COM-Poisson distribution (in red).

2.5 Bayesian inference

Bayesian and frequentist approaches to inference work within the same overall framework: there is a population parameter θ which we want to make inferences about, and a likelihood distribution $p(y|\theta)$ which determines the likelihood of observing the data y, under different parameter values θ. The crucial difference is that in the Bayesian approach θ is treated as a random variable. Thus, we can specify a probability distribution $p(\theta)$ for the parameter θ. This distribution, known as the prior distribution, represents our beliefs about the distribution of θ prior to observing any data.

Bayesian inference can be thought of as the mechanism for drawing inference from the combined knowledge of our prior beliefs (represented in the prior distribution) and the data (represented in the likelihood). This knowledge is expressed in the posterior distribution $\pi(\theta|y)$; which can be thought of as representing our beliefs about the distribution of θ after we have observed the data. Expressing Bayes' theorem in terms of random variables we get

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)\mathrm{d}\theta}, \\ &= \frac{p(y|\theta)p(\theta)}{p(y)}, \\ &\propto p(y|\theta)p(\theta). \end{aligned} \tag{2.94}$$

The denominator in the second line in (2.94) is a normalisation constant which is independent of the parameter θ. This constant, also known as the evidence, can be difficult to compute especially when θ is multivariate. There are situations though in which $p(y)$ can be computed easily. One of them is when the posterior distribution belongs to the same family of distributions

as the prior distribution. This can be checked using the final line in (2.94) and is known as conjugacy whereas the prior is called conjugate prior.

We will introduce the basics behind Bayesian inference but more information can be found in Gelman et al. (2004); Robert and Casella (2009). Another, more introductory, good source of reference is Rogers and Girolami (2011).

Prior distribution

We have already mentioned that any inference on θ depends on the prior distribution and the data. Different prior distributions will lead to different inferences for the unknown parameter θ. This can also be seen in the final line of (2.94). The prior distribution is chosen to represent our beliefs (or ignorance) about θ and therefore it can be chosen to be informative or vague. Informative prior distributions express both our knowledge and our uncertainty about θ whereas vague priors are selected such that they have little effect on the posterior distribution. This means that inferences on θ will not be affected by our beliefs but only from the data (through the likelihood).

Posterior inference

Bayesian inference, like frequentist inference, uses point estimates and/or interval estimates, known as credible intervals. Usually, the posterior mean or median are used to describe the location of the posterior distribution while the posterior variance (or standard deviation) can be used to describe the scale of the posterior distribution. A credible interval is the Bayesian

analogue of the confidence interval, although the two have different interpretations. A 95% credible interval of $[0.03, 0.06]$ has the straightforward interpretation that there is a 95% probability that θ lies between 0.03 and 0.06.

2.5.1 Stochastic simulation

In simple models, the posterior $p(\theta|y)$ from (2.94) can be obtained in closed form. If this is not the case one has to resort to other strategies, such as stochastic simulation. Monte Carlo methods are techniques that use simulation to obtain summary statistics (mean, variance, quantiles) of the posterior distribution. There are ways that one can generate independent identically distributed samples from standard distributions but for higher dimensions this is not trivial. Markov chain Monte Carlo methods are methods for drawing samples from the posterior distribution no matter its complexity. These draws will however not be independent, but form a Markov chain. A Markov chain is a sequence of random variables $(\theta_n)_{n=0}^{\infty}$ in which the future is independent of the past, given the present. This can be seen as:

$$\pi(\theta_{n+1}|\theta_0, \theta_1, \ldots, \theta_n) = \pi(\theta_{n+1}|\theta_n). \tag{2.95}$$

For the MCMC methods to work, we have to use a Markov chain that has as a stationary distribution the posterior distribution $\pi(\theta) = p(\theta|y)$. The main idea behind these methods is that after a period of time (more on that in Subsection 2.5.2), they will consist of a sample from the posterior distribution.

Metropolis-Hastings algorithm

In the Metropolis-Hastings algorithm of Hastings (1970), a Markov chain is constructed in which, when the current state is θ, a candidate state θ^* is drawn from a proposal distribution $h(\theta^*|\theta)$ and then accepted with probability p where

$$p = \min\left\{1, \frac{\pi(\theta^*)h(\theta|\theta^*)}{\pi(\theta)h(\theta^*|\theta)}\right\}, \tag{2.96}$$

where $\pi(\theta^*)$ is the target distribution at θ^*. If θ^* is rejected, the chain remains at the current state θ. In order to accept the candidate θ^* with probability p, the acceptance probability p is compared to a random variable $u \sim \text{Unif}(0, 1)$ and θ^* is accepted if $u < p$.

The Metropolis-Hastings algorithm is a generalisation of other well known MCMC algorithms:

- The Metropolis algorithm of Metropolis et al. (1953), in which the proposal distribution is symmetric e.g. $h(\theta^*|\theta) = h(\theta|\theta^*)$.
- The independent Metropolis, in which the proposal distribution is independent of the current state e.g. $h(\theta^*|\theta) = h(\theta^*)$.
- Gibbs sampling, in which the proposal distribution is a conditional distribution of the target distribution e.g. $h(\theta^*|\theta) = \pi(\theta^*)$. In the case of the Gibbs sampler $\pi(\theta^*)$ is the full conditional distribution given the other parameters. The Gibbs sampler can be seen as a special case of the Metropolis-Hastings algorithm where each move is always accepted with probability one.

A more thorough explanation of the Metropolis-Hastings algorithm and the developments after its emergence, can be found in Chib and Greenberg (1995); Casella and Robert (2011).

2.5.2 MCMC diagnostics

A crucial part of every MCMC algorithm is checking whether (and when) it shows evidence of convergence. We have mentioned that the algorithm must be run long enough to draw samples from the posterior distribution, but how long is that specifically?

A first step would be to examine the trace plots and posterior densities of every parameter. Usually this is done by running multiple MCMC chains, with different starting values. For the MCMC algorithms to show signs of convergence they must gravitate towards the same range of values for each parameter. The point at which this happens gives an idea of the period that we can discard, known as burn-in period.

Gelman et al. (2004) propose running multiple chains and provide a statistic, $\hat{R}$, that checks for signs of convergence. This diagnostic uses an analysis of variance to assess convergence. Specifically, it calculates the between-chain variance (B) and within-chain variance (W), and assesses whether they are different enough. Assuming one is running m chains, each of length n,

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\overline{\theta}_{.j} - \overline{\theta}_{..})^2, \quad W = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \overline{\theta}_{.j})^2\right) \tag{2.97}$$

where

$$\overline{\theta}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij}, \quad \overline{\theta}_{..} = \frac{1}{m}\sum_{j=1}^{m}\overline{\theta}_{.j} \tag{2.98}$$

An estimate of the marginal posterior of θ can be calculated as

$$\hat{\mathbb{V}}[\theta|y] = \frac{n-1}{n}W + \frac{1}{n}B. \tag{2.99}$$

$\hat{R}$ is defined as

$$\hat{R} = \sqrt{\frac{\hat{\mathbb{V}}(\theta|y)}{W}} \tag{2.100}$$

and provides an estimate of the potential reduction in the scale of θ as the number of simulations tends to infinity. A value of $\hat{R}$ close to 1 suggests that the chains have converged to the target distribution; Gelman et al. (2004) recommend to run the chains until the value of $\hat{R}$ for each parameter is below 1.1.

Due to the nature of a Markov chain, the posterior sample given from an MCMC chain is correlated. Less autocorrelation in the parameter of the posterior sample indicates better mixing of the chain and faster convergence. This can be checked using autocorrelation plots. High autocorrelation can be dealt with by adjusting the variance of the proposal distribution, jointly updating the parameters that are highly correlated, keeping only every j^{th} iterate from the sample (known as thinning), and/or choosing a different density for the candidate states.

2.6 Bayesian density regression for continuous data

Density regression allows flexible modelling of the response variable Y given the covariates $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})^\intercal$. Features (mean, quantiles, spread) of the conditional distribution of the response variable given the covariates, vary with $\boldsymbol{x}_i$. So, depending on the predictor values, features of the conditional distribution can change in a different way than the population mean.

2.6.1 Dunson *et al.* model

Dunson et al. (2007) propose Bayesian methods for density regression allowing the probability distribution of the response variable to change flexibly with predictors. The conditional distribution of the response variable is expressed as a mixture of regression models where the mixing weights vary with predictors. They propose as a prior for the uncountable collection of the mixture distributions a weighted mixture of Dirichlet process priors. The weights are dependent on the distance between predictors' values. Predictors that are "close" to each other tend to have higher weight. This procedure results in a generalisation of the Pólya urn scheme that does not rely on the exchangeability of the subjects.

A large number of mixtures of normal densities can be used to approximate any smooth density accurately and based on that, Dunson et al. (2007) focus

on the following mixture of regression models:

$$f(y_i|\boldsymbol{x}_i) = \int f(y_i|\boldsymbol{x}_i, \boldsymbol{\phi}_i) G_{\boldsymbol{x}_i}(\mathrm{d}\boldsymbol{\phi}_i), \tag{2.101}$$

where

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{\phi}_i) = N(y_i; \boldsymbol{x}_i^\intercal \boldsymbol{b}_i, \sigma_i^2), \tag{2.102}$$

with $\boldsymbol{\phi}_i = (\boldsymbol{b}_i^\intercal, \sigma_i^2)^\intercal$ and $G_{\boldsymbol{x}_i}$an unknown mixture distribution that changes according to the location of $\boldsymbol{x}_i$.

Assuming $G_{\boldsymbol{x}} \equiv G$ (mixing distribution does not depend on $\boldsymbol{x}$) where $G \sim DP(a, G_0)$; we end up with a DP mixture of normal linear regression models. This is equivalent to

$$f(y_i|\boldsymbol{x}_i) = \sum_{h=1}^{\infty} \pi_h N(y_i; \boldsymbol{x}_i^\intercal \boldsymbol{b}_h, \sigma_i^2), \tag{2.103}$$

with $\boldsymbol{\pi} = (\pi_h)_{h=1}^{\infty}$ an infinite sequence of weights that sum to one and $\boldsymbol{\phi}_h = (\boldsymbol{b}_h^\intercal, \sigma_h^2)^\intercal$, atoms sampled independently from the base distribution G_0. Conditioning on the allocation variable $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)^\intercal$, where $Z_i = h$ means that the i^{th} observation belongs to the h^{th} mixture component, the mean of this model is:

$$\begin{aligned}
\mathbb{E}[Y_i|\boldsymbol{x}_i] &= \sum_{h=1}^{\infty} P(Z_i = h)\mathbb{E}[Y_i|\boldsymbol{x}_i, Z_i = h], \\
&= \sum_{h=1}^{\infty} \pi_h \boldsymbol{x}_i^\intercal \boldsymbol{b}_h, \\
&= \boldsymbol{x}_i^\intercal \sum_{h=1}^{\infty} \pi_h \boldsymbol{b}_h, \\
&= \boldsymbol{x}_i^\intercal \overline{\boldsymbol{b}_h}.
\end{aligned} \tag{2.104}$$

where $\overline{\boldsymbol{b}_h} = \sum_{h=1}^{\infty} \pi_h \boldsymbol{b}_h$. Since π_h does not depend on $\boldsymbol{x}_i$ the dispersion stays the same across the conditional mean. In other words the model effectively

allows the distribution of the residuals to be non-Gaussian, but this distribution is the same across the entire covariate space. The reason for this is that despite incorporating an infinite number of normal linear regression models, the model assumes that the weights are constant. It does not take into account the values of the predictor $\boldsymbol{x}_i$. This goes against our expectation of "similar" predictor values leading to "similar" distribution for the response variable. For a graphical explanation, see Figure (2.8), where we assume that the regression model has two covariates ($\boldsymbol{x}_i = (x_{i1}, x_{i2})^\intercal$) and given a new predictor value $\boldsymbol{x}_7$ our goal is to estimate $f(y_7|\boldsymbol{x}_7)$. The predictors are clustered into three different groups (different colour in each group). Each group has its own parameters θ_k^* for $k = 1, 2, 3$. Taking two possible values for the new predictor $\boldsymbol{x}_7$ (with the + sign), it seems logical for the one on the bottom left to belong to the "red" cluster and the one on the top left to a new cluster. Assuming $G_{\boldsymbol{x}} \equiv G$ with $G \sim DP(a, G_0)$ clusters the predictors by just considering the number of predictors belonging to the same cluster. The parameter $\boldsymbol{\phi}_7 = (\boldsymbol{b}_7^\intercal, \sigma_7^2)^\intercal$ for the new predictor is going to be one of the already existing parameters (of the three clusters) or a new value from G_0. Thus, the probability of $\boldsymbol{x}_7$ belonging to the "blue" cluster would be higher than to each of the other two clusters.

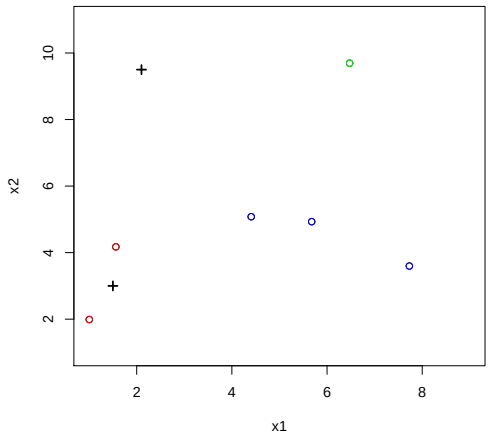


Figure 2.8: The predictors are clustered into three different groups with a different colour for each group. Two possible values of a new predictor are shown with a + sign. Having a Dirichlet process as a mixing distribution will give high probability to the new predictor belonging to the "blue" cluster (due to the "rich get "richer" phenomenon) even though this does not seem realistic.

2.6.2 Weighted mixtures of Dirichlet process priors

Overcoming the previous restriction is done by proposing a prior $G_{\boldsymbol{x}}$ as a weighted mixture of Dirichlet process priors. After placing a Dirichlet process-distributed probability measure at each of the sample predictors' values

$$G^*_{\boldsymbol{x}_i} \sim DP(a, G_0) \text{ for } i = 1, \ldots, n \tag{2.105}$$

then mixing across these measures

$$G_{\boldsymbol{x}} = \sum_{i=1}^{n} b_i(\boldsymbol{x}) G^*_{\boldsymbol{x}_i} \tag{2.106}$$

constructs an uncountable collection of probability measures for all possible values $\boldsymbol{x} = (1, x_1, \ldots, x_k)^\intercal$, where $\boldsymbol{b}(\boldsymbol{x}) = (b_1(\boldsymbol{x}), \ldots, b_n(\boldsymbol{x}))'$ is a weight function such that $b_i(\boldsymbol{x}) \geq 0, i = 1, \ldots, n$, and $\boldsymbol{b}(\boldsymbol{x})' 1_n = 1$, with 1_n the $n \times 1$ vector of 1s. The proposed weight function has the form:

$$\boldsymbol{b}(\boldsymbol{x}) = \gamma_i K(\boldsymbol{x}, \boldsymbol{x}_i) / \sum_{l=1}^{n} \gamma_l K(\boldsymbol{x}, \boldsymbol{x}_l) \quad i = 1, \ldots, n, \tag{2.107}$$

where $\gamma = (\gamma_1, \ldots, \gamma_n)^\intercal$ represent location weights and K is a kernel function, such as

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\{-\psi \|\boldsymbol{x} - \boldsymbol{x}'\|^2\}. \tag{2.108}$$

The above formulation ensures that distributions close to $\boldsymbol{x}$ are assigned high weight in the prior for $G_{\boldsymbol{x}}$, especially if these locations have high γ values.

2.6.3 Importance of location weights

Allowing the weights for some or most of the locations to be close to zero is equivalent to $\gamma_i / \sum_l \gamma_l \to 1$, which results in the i^{th} component of the weight function being one and all the others are zero. This means that no matter how "close" or "far" the predictor values of a new observation are from $\boldsymbol{x}_i$ we will still have the same mixing measure $G^*_{\boldsymbol{x}_i} \equiv G$. This gives us the aforementioned DP mixture model. On the other hand, assigning uniform weights to each location is not a good choice since the number of subjects at any location will tend to be small. Thus, the weights γ_i are an important part of this formulation since the allocation of subjects to different basis locations allows the mixture distribution to change with predictors. A logical choice for γ_i would favour a few dominant locations, the number of which would

increase with sample size. This is accomplished by proposing the prior

$$\gamma_i \sim \text{gamma}(k, n \times k) \qquad k \sim \text{log-N}(\mu_k, \sigma_k^2), \tag{2.109}$$

with μ_k and σ_k^2 placing high probability on small, but not minute, values.

2.6.4 Generalised Pólya urn scheme

Marginalising across $G_{\boldsymbol{x}}$, the weighted mixture of DP priors, results in a generalisation of the Pólya urn scheme, which incorporates weights that depend on the distance between subjects' predictor values. Dunson et al. (2007) finally show that for any $n \times n$ matrix $\boldsymbol{B}$, with elements $(b_{ij})_{i,j=1}^n$ satisfying $0 \leq b_{ij} \leq 1$ and $\boldsymbol{b}_i^\intercal \mathbf{1}_n = 1$, there is a unique $n \times (n-1)$ matrix $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^\intercal$ having row vectors $\boldsymbol{w}_i = (w_{i,1}, \ldots, w_{i,i-1}, w_{i,i+1}, \ldots, w_{i,n})^\intercal$, with $0 \leq w_{ij} \leq 1$ such that the conditional distribution of ϕ_i is equivalent to:

$$(\phi_i | \boldsymbol{\phi}^{(i)}, \boldsymbol{X}, a, \boldsymbol{B}) = \frac{a}{a + w_i} G_0 + \sum_{j \neq i} \frac{w_{ij}}{a + w_i} \delta_{\phi_j}. \tag{2.110}$$

Letting $\boldsymbol{p}_{i0} = \boldsymbol{p}_0(\boldsymbol{x}_i)$ denote the $n \times 1$ vector of probabilities corresponding to $P(M_{i+} = m | \boldsymbol{x}_i)$, for $m = 0, \ldots, n-1$ with $M_{i+} = \sum_{j \neq i} M_{ij}$, $\boldsymbol{p}_{ij} = \boldsymbol{p}_j(\boldsymbol{x}_i)$ denote the $(n-1) \times 1$ vector of probabilities corresponding to $P(M_{ij} = 1, M_{i+} = m | \boldsymbol{x}_i)$, for $m = 1, \ldots, n-1$ and

$$\begin{aligned} \boldsymbol{\Gamma}_0 &= \left(1, \frac{a}{a+1}, \frac{a}{a+2}, \ldots, \frac{a}{a+n-1}\right)^\intercal, \\ \boldsymbol{\Gamma}_1 &= \left(\frac{1}{a+1}, \frac{1}{a+2}, \ldots, \frac{1}{a+n-1}\right)^\intercal \end{aligned} \tag{2.111}$$

we have that

$$w_{ij} = \frac{a \boldsymbol{p}_{ij}^\intercal \boldsymbol{\Gamma}_1}{\boldsymbol{p}_{i0}^\intercal \boldsymbol{\Gamma}_0} \tag{2.112}$$

is a set of weights between 0 and 1 that depend on $\boldsymbol{B}$ and a where

$$\boldsymbol{B} = (\boldsymbol{b}(\boldsymbol{x}_1), \ldots, \boldsymbol{b}(\boldsymbol{x}_n))^\intercal. \tag{2.113}$$

and $w_i = \sum_{j \neq i} w_{ij}$.An assumption for the clustering property of the DP is that the data are exchangeable which does not hold now since the subjects' predictor values are informative about the clustering. Instead of assuming exchangeability, weights w_{ij} that depend on the subjects' relative predictor values are included in the conditional distribution. Letting $w_{ij} = 1$ for all i, j we obtain the Pólya urn conditional distribution.

Derivation of generalised Pólya urn scheme

Dunson et al. (2007) condition on the allocation of subjects to mixture components $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\intercal$ and using the Pólya urn result they obtain the conditional prior

$$(\phi_i | \boldsymbol{Z}, \boldsymbol{\phi}^{(i)}, \boldsymbol{X}, a) \sim \frac{a}{a + \sum_{j \neq i}(Z_j = Z_i)} G_0 + \frac{1}{a + \sum_{j \neq i}(Z_j = Z_i)} \sum_{j \neq i} \mathbf{1}(Z_j = Z_i)\delta_{\phi_j} \tag{2.114}$$

where $M_{ij} = \mathbf{1}(Z_j = Z_i)$ is a $0 - 1$ indicator that subjects i and j belong to the same cluster.

The probability of $\boldsymbol{M}_i = \{M_{ij}, j \neq i\} = \boldsymbol{m}_i = \{m_{ij}, j \neq i\}$, for $\boldsymbol{m}_i \in$

$\{0,1\}^{n-1}$ is,

$$\begin{aligned} P(\boldsymbol{M}_i = \boldsymbol{m}_i) &= \sum_{j=1}^{n} Pr(Z_i = j) \prod_{h \neq i} P(Z_h = j)^{m_{ih}} (1 - P(Z_h = j))^{1-m_{ih}} \\ &= \sum_{j=1}^{n} b_j(\boldsymbol{x}_i) \prod_{h \neq i} b_j(\boldsymbol{x}_h)^{m_{ih}} (1 - b_j(\boldsymbol{x}_h))^{1-m_{ih}} \\ &= \sum_{j=1}^{n} b_{ij} \prod_{h \neq i} b_{hj}^{m_{ih}} (1 - b_{hj})^{1-m_{ih}}. \end{aligned} \tag{2.115}$$

Finally, after marginalising across the distribution for $\boldsymbol{M}_i$:

$$\begin{aligned} (\phi_i | \boldsymbol{\phi}^{(i)}, \boldsymbol{X}, a, \boldsymbol{B}) \sim & \sum_{h \neq i} \sum_{m_{ih}=0}^{1} \{ \sum_{j=1}^{n} b_{ij} \prod_{l \neq i} b_{lj}^{m_{il}} (1 - b_{lj})^{1-m_{il}} \} \\ & \times \Big(\frac{a}{a + \sum_{l \neq i} m_{il}} G_0 + \frac{1}{a + \sum_{l \neq i} m_{il}} \sum_{l \neq i} m_{il} \delta_{\phi_l} \Big). \end{aligned} \tag{2.116}$$

The above formula is a generalisation of the Blackwell and MacQueen Pólya urn scheme.

We can express Equation (2.116) as

$$\begin{aligned} (\phi_i | \boldsymbol{\phi}^{(i)}, \boldsymbol{X}, a, \boldsymbol{B}) &= \boldsymbol{p}_{i0}^{\intercal} \boldsymbol{\Gamma}_0 G_0 + \sum_{j \neq i} \boldsymbol{p}_{ij}^{\intercal} \boldsymbol{\Gamma}_1 \delta_{\phi_j} \\ &= \boldsymbol{p}_0(\boldsymbol{x}_i)^{\intercal} \boldsymbol{\Gamma}_0 G_0 + \sum_{j \neq i} \boldsymbol{p}_j(\boldsymbol{x}_i)^{\intercal} \boldsymbol{\Gamma}_1 \delta_{\phi_j}. \end{aligned} \tag{2.117}$$

where $\boldsymbol{p}_{i0}^{\intercal} \mathbf{1}_n = 1$ and $\boldsymbol{p}_{ij}^{\intercal} \mathbf{1}_{n-1} \leq 1$. This is in the form of a weighted average of Blackwell and MacQueen Pólya urn distributions. Finally, for any $n \times n$ matrix $\boldsymbol{B}$, with elements $(b_{ij})_{i,j=1}^{n}$ satisfying $0 \leq b_{ij} \leq 1$ and $\boldsymbol{b}_i^{\intercal} \mathbf{1}_n = 1$, there is a unique $n \times (n-1)$ matrix $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^{\intercal}$ having row vectors $\boldsymbol{w}_i = (w_{i,1}, \ldots, w_{i,i-1}, w_{i,i+1}, \ldots, w_{i,n})^{\intercal}$, with $0 \leq w_{ij} \leq 1$ such that the above expression is equivalent to:

$$(\phi_i | \boldsymbol{\phi}^{(i)}, \boldsymbol{X}, a, \boldsymbol{B}) = \frac{a}{a + w_i} G_0 + \sum_{j \neq i} \frac{w_{ij}}{a + w_i} \delta_{\phi_j}, \tag{2.118}$$

2.6.5 MCMC algorithm

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\intercal$ denote the $k \leq n$ distinct values of $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)^\intercal$ and let $\boldsymbol{S} = (S_1, \ldots, S_n)^\intercal$ be a vector of indicators denoting the global configuration of subjects to distinct values $\boldsymbol{\theta}$, with $S_i = h$ indexing the location of the i^{th} subject within the $\boldsymbol{\theta}$ vector. In addition, let $\boldsymbol{C} = (C_1, \ldots, C_k)^\intercal$ with $C_h = j$ denoting that θ_h is an atom from the basis distribution, $G^*_{\boldsymbol{x}_j}$. Hence $C_{S_i} = Z_i = j$ denotes that subject i is drawn from the j^{th} basis distribution.

Excluding the i^{th} subject, $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta} \backslash \{\phi_i\}$ denotes the $k^{(i)}$ distinct values of $\boldsymbol{\phi}^{(i)} = \boldsymbol{\phi} \backslash \{\phi_i\}$, $\boldsymbol{S}^{(i)}$ denotes the configuration of subjects $\{1, \ldots, n\} \backslash \{i\}$ to these values and $\boldsymbol{C}^{(i)} = (C_1^{(i)}, \ldots, C_{k^{(i)}}^{(i)})^\intercal$ indexes the DP component numbers for the elements of $\boldsymbol{\theta}^{(i)}$.

Conditioning on $\boldsymbol{Z}^{(i)}$ but marginalising over Z_i, Dunson et al. (2007) obtain the following conditional prior for ϕ_i:

$$(\phi_i | \boldsymbol{Z}^{(i)}, \boldsymbol{\phi}^{(i)}, \boldsymbol{X}, a) \sim \sum_{i=1}^{n} \frac{a b_{ij}}{a + \sum_{l \neq i} \mathbf{1}(Z_l = j)} G_0 + \sum_{m \neq i} \Big\{ \sum_{j=1}^{n} \frac{b_{ij} \mathbf{1}(Z_m = j)}{a + \sum_{l \neq i} \mathbf{1}(Z_l = j)} \Big\} \delta_{\phi_m}. \tag{2.119}$$

Grouping the subjects in the same cluster, we obtain the expression

$$(\phi_i | \boldsymbol{S}^{(i)}, \boldsymbol{C}^{(i)}, \boldsymbol{\theta}^{(i)}, \boldsymbol{X}, a) \sim w_{i0} G_0 + \sum_{h=1}^{k^{(i)}} w_{ih} \delta_{\theta_h^{(i)}}, \tag{2.120}$$

where the probability weights on the various components are defined as

$$\begin{aligned} w_{i0} &= \sum_{j=1}^{n} \frac{a b_{ij}}{a + \sum_{l \neq i} \mathbf{1}(C^{(i)}_{S_l^{(i)}} = j)}, \\ w_{ih} &= \frac{b_{i, C_h^{(i)}} \sum_{m \neq i} \mathbf{1}(S_m^{(i)} = h)}{a + \sum_{l \neq i} \mathbf{1}(C^{(i)}_{S_l^{(i)}} = C_h)}, \; h = 1, \ldots, k^{(i)}. \end{aligned} \tag{2.121}$$

Updating the prior with the likelihood for the data $\boldsymbol{y}$, we obtain the conditional posterior:

$$(\phi_i|\boldsymbol{S}^{(i)}, \boldsymbol{C}^{(i)}, \boldsymbol{\theta}^{(i)}, \boldsymbol{X}, a) \sim q_{i0}G_{i,0} + \sum_{h=1}^{k^{(i)}} q_{ih}\delta_{\theta_h^{(i)}}, \tag{2.122}$$

where $G_{i,0}(\phi)$ is the posterior obtained by updating the prior $G_0(\phi)$ and the likelihood $f(y_i|\boldsymbol{x}_i, \phi)$:

$$\begin{aligned} G_{i,0}(\phi) &= \frac{G_0(\phi)f(y_i|\boldsymbol{x}_i, \phi)}{\int f(y_i|\boldsymbol{x}_i, \phi)\mathrm{d}G_0(\phi)} \\ &= \frac{G_0(\phi)f(y_i|\boldsymbol{x}_i, \phi)}{h_i(y_i|\boldsymbol{x}_i)}, \end{aligned} \tag{2.123}$$

where

$$\begin{aligned} q_{i0} &= cw_{i0}h_i(y_i|\boldsymbol{x}_i), \\ q_{ih} &= cw_{ih}f(y_i|\boldsymbol{x}_i, \theta_h), \ h = 1, \ldots, k^{(i)} \end{aligned} \tag{2.124}$$

and c is a normalisation constant.

The MCMC algorithm alternates between the following steps:

Step 1: Update S_i for $i = 1, \ldots, n$, by sampling from the multinomial conditional posterior $P(S_i = h) = q_{ih}, h = 1, \ldots, k^{(i)}$. When $S_i = 0$, sample $\phi_i \sim G_{i,0}$ and $C_{S_i} \sim$ multinomial $(\{1, \ldots, n\}, \boldsymbol{b}_i)$.

Step 2: Update the parameters θ_h, for $h = 1, \ldots, k$ by sampling from the conditional posterior distribution

$$(\theta_h|\boldsymbol{S}, \boldsymbol{C}, \boldsymbol{\theta}^{(h)}, k, \boldsymbol{y}, \boldsymbol{X}) \sim \prod_{i:S_i=h} f(y_i|\boldsymbol{x}_i, \theta_h)\}G_0(\theta_h), \tag{2.125}$$

that has a simple form when G_0 is chosen to be a conjugate prior.

Step 3: Update C_h, for $h = 1, \ldots, k$, by sampling from the multinomial conditional with

$$(C_h|\boldsymbol{S}, \boldsymbol{C}^{(h)}, \boldsymbol{\theta}, k, \boldsymbol{y}, \boldsymbol{X}) \sim \frac{\prod_{i:S_i=h} b_{ij}}{\sum_{l=1}^{n} \prod_{i:S_i=h} b_{il}}, \; j = 1, \ldots, n. \tag{2.126}$$

Step 4: Update the weights γ_j for $j = 1, 2, \ldots, n$, by using a data augmentation approach motivated by Dunson and Stanford (2005); Holmes et al. (2006). Letting $K_{ij} = \exp\{-\psi\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\}$ and $K^*_{ij} = \frac{K_{ij}}{\sum_{l\neq j}\gamma_l K_{il}}$, the conditional likelihood for γ_j is

$$L(\gamma_j) = \prod_{i=1}^{n} \left(\frac{\gamma_j K^*_{ij}}{1+\gamma_j K^*_{ij}}\right)^{\mathbf{1}(C_{S_i}=j)} \left(\frac{1}{1+\gamma_j K^*_{ij}}\right)^{\mathbf{1}(C_{S_i}\neq j)}. \tag{2.127}$$

This likelihood can be obtained by using $\mathbf{1}(C_{S_i} = j) = \mathbf{1}(Z^*_{ij} > 0)$, with $Z^*_{ij} \sim \text{Poisson}(\gamma_j \xi_{ij} K^*_{ij})$ and $\xi_{ij} \sim \text{exponential}(1)$. Updating $\{Z^*_{ij}, \xi_{ij}\}$ and $\{\gamma_j\}$ in Gibbs steps:

1. let $Z^*_{ij} = 0$ if $\mathbf{1}(C_{S_i} \neq j)$ and otherwise $Z^*_{ij} \sim \text{Poisson}(\gamma_j \xi_{ij} K^*_{ij})\mathbf{1}(Z^*_{ij} > 0)$, for all i and j;
2. $\xi_{ij} \sim \text{gamma}(1 + Z^*_{ij}, 1 + \gamma_j K^*_{ij})$, for all i and j;
3. letting $\text{gamma}(a_\gamma, b_\gamma)$ denote the prior for γ_j,

$$\gamma_j \sim \text{gamma}(a_\gamma + \sum_{i=1}^{n} Z^*_{ij}, b_\gamma + \sum_{i=1}^{n} \xi_{ij} K^*_{ij}). \tag{2.128}$$

To compare with the MCMC algorithm for count data see Section 4.2.

2.6.6 Clustering properties

From equations (2.120) and (2.121) we have that

- The probability of $\phi_i \notin \boldsymbol{\phi}^{(i)}$ is proportional to the precision parameter α. The larger the value of α is, the higher the probability that the i^{th} subject will go to a new cluster.
- The probability of $\phi_i \in \boldsymbol{\phi}^{(i)}$ is proportional to the number of subjects that have predictor values close to $\boldsymbol{x}_i$.
- The probability that the subjects $\boldsymbol{x}_i, \boldsymbol{x}_j$ will be in the same cluster is inversely proportional to their distance.

Thus, the allocation of subjects to clusters is controlled by α and the distance between each subject's predictor values. Observations in sparse regions are more likely to be allocated to a new cluster.

2.6.7 Predictive density and simulation examples

The primary goal of Dunson et al. (2007) is to estimate the predictive density of a future observation y_{new} from a new subject with predictors $\boldsymbol{x}_{\text{new}}$. They show that

$$(y_{\text{new}}|\boldsymbol{x}_{\text{new}}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{\theta}, k, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta, \tau, \boldsymbol{X}) = \omega_{n,0}(\boldsymbol{x}_{\text{new}}) N(y_{\text{new}}; \boldsymbol{x}_{\text{new}}^\intercal \boldsymbol{\beta}, \tau^{-1} + \boldsymbol{x}_{\text{new}}^\intercal \Sigma_\beta \boldsymbol{x}_{\text{new}}) + \sum_{h=1}^{k} \omega_{n,h}(\boldsymbol{x}_{\text{new}}) N(y_{\text{new}}; \boldsymbol{x}_{\text{new}}^\intercal \boldsymbol{\beta}_h, \tau^{-1}) \tag{2.129}$$

which is a finite mixture of normal linear regression models. The probability weights depend on the location of $\boldsymbol{x}_{\text{new}}$, which allows deviations on the density across the space of the covariates. Finally, to examine any changes in the

conditional density across the covariate space we can calculate the expected predictive density by using a large number of iterations (after convergence) and averaging over them. Another way to estimate the mean regression curve can be by sampling with the above probabilities from the same iterations and averaging over them. Dunson et al. (2007) simulate data under the normal linear regression model

$$f(y_i|\boldsymbol{x}_i) = N(y_i; -1 + x_{i1}, 0.01) \tag{2.130}$$

where $x_{i1} \sim \text{Unif}(0, 1)$ and $\boldsymbol{x}_i = (1, x_{i1})^\intercal$ and show that the predictive densities, across the covariate space, are very close to the true densities.

To show the "strength" of the method (Dunson et al., 2007), we simulate data from Gaussian distributions with different assumptions for the mean and variance.

1. Adding more covariates,

$$f(y_i|\boldsymbol{x}_i) = N(y_i; -2 + 5x_{i1} - 3x_{i2}^2, 0.01) \tag{2.131}$$

   where $x_{i1}, x_{i2} \sim \text{Unif}(0, 1)$ and $\boldsymbol{x}_i = (1, x_{i1}, x_{i2})^\intercal$. Figure 2.9 shows the predictive density of y_{new} at the $0.1, 0.25, 0.5, 0.75$ and 0.9 quantiles of the empirical distribution of x_{i1} with the covariate x_{i2} fixed at its sample mean. The predictive densities are very close to the true densities.

2. Assuming there is a non-constant variance,

$$f(y_i|\boldsymbol{x}_i) = N(y_i; -2 + 5x_{i1}, (1 + x_{i1})^2) \tag{2.132}$$

   where $x_{i1} \sim \text{Unif}(0, 1)$ and $\boldsymbol{x}_i = (1, x_{i1})^\intercal$. Figure 2.10 shows that the predictive densities capture the non-constant variance across the quantiles.

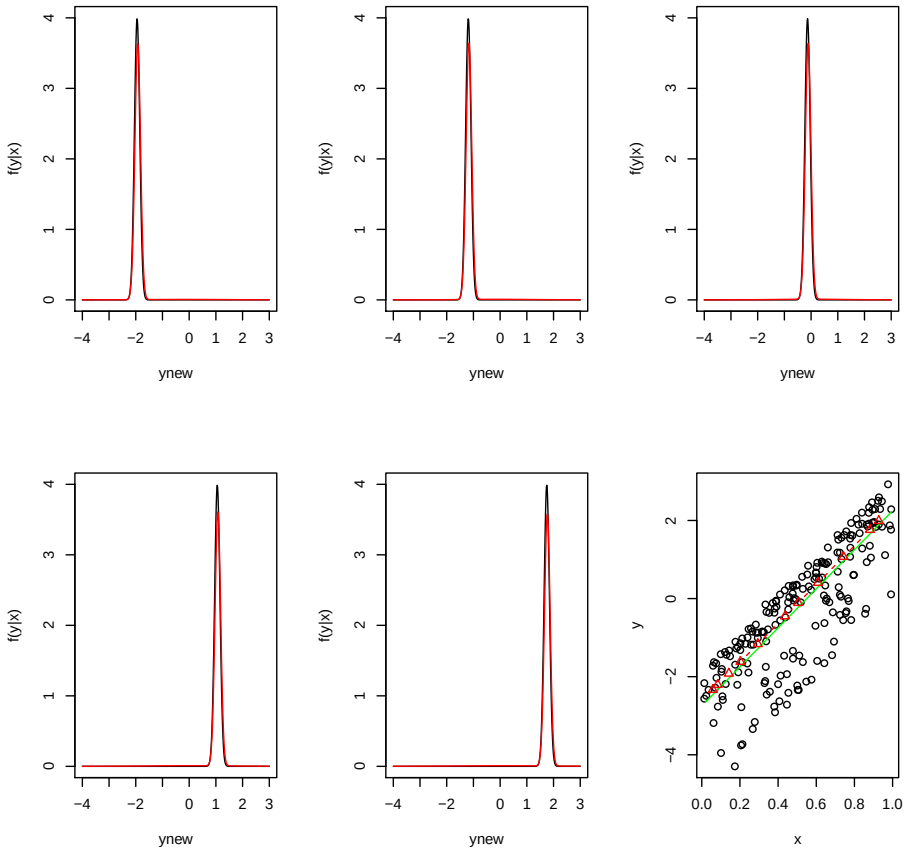


Figure 2.9: True conditional densities of $y|x$ are represented with a black line and posterior mean estimates are with a red line where the covariate x_{i2} fixed at its sample mean. The first five plots refer to the quantiles $q = 0.1, 0.25, 0.5, 0.75, 0.9$ of the empirical distribution of x_{i1} with the covariate x_{i2} fixed at its sample mean. The final plot has the data along with the true mean regression line in green and the estimated regression line in red.

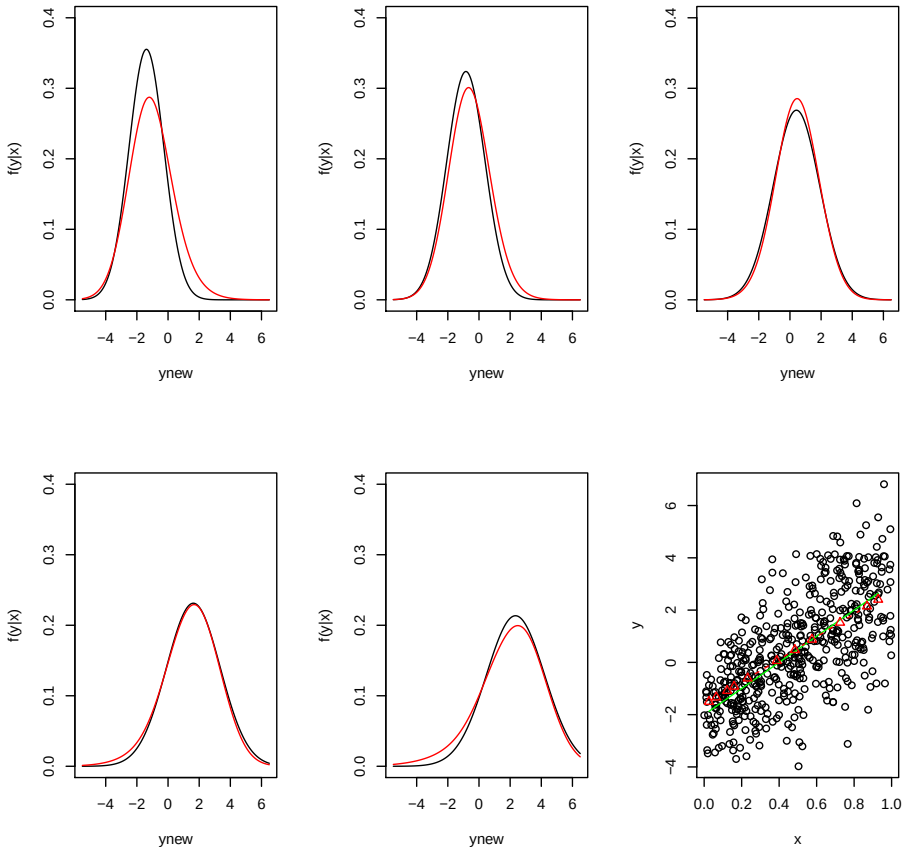


Figure 2.10: True conditional densities of $y|x$ are represented with a black line and posterior mean estimates are with a red line. The first five plots refer to the quantiles $q = 0.1, 0.25, 0.5, 0.75, 0.9$ of the empirical distribution of x_{i1}. The final plot has the data along with the true mean regression line in green and the estimated regression line in red.

3. Finally, as a more interesting case we simulate data from a mixture of two normal linear regression models, with the mixture weights depending on the predictor, with the error variance differing and with a non-linear mean function for the second component:

$$f(y_i|\boldsymbol{x}_i) = \exp\{-2x_{i1}\}N(y_i; x_{i1}, 0.01) + (1 - \exp\{-2x_{i1}\})N(y_i; x_{i1}^4, 0.04) \tag{2.133}$$

where $x_{i1} \sim \text{Unif}(0, 1)$ and $\boldsymbol{x}_i = (1, x_{i1})^\intercal$. It is clear from Figure 2.11 that the estimated conditional densities are close to the true conditional densities across the whole space.

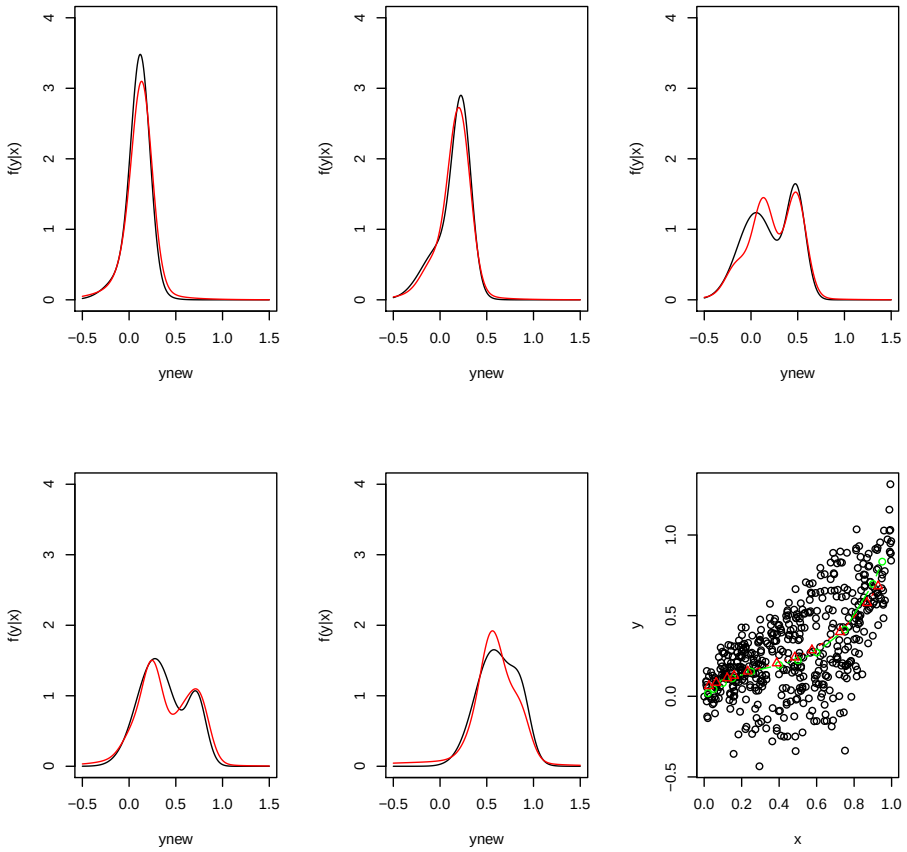


Figure 2.11: True conditional densities of $y|x$ (for the third simulated example) are in black and posterior mean estimates are in red. First five plots refer to the quantiles $q = 0.1, 0.25, 0.5, 0.75, 0.9$ of the empirical distribution of x_{i1}. The final plot has the data along with the true mean regression line in green and the estimated regression line in red.

2.6.8 Other approaches to density regression

Another approach to density regression is to model the joint density (of the response variable and the covariates) and obtain the conditional density as a byproduct. Müller et al. (1996) focus on estimating the regression function $g(\boldsymbol{x}_i) = \mathbb{E}[Y_i|\boldsymbol{x}_i]$ based on the complete data $\boldsymbol{d}_i = (y_i, \boldsymbol{x}_i)$. They assume that the distribution of $\boldsymbol{d}$ has a mixture form and fit a Dirichlet process of multivariate Gaussian distributions to the complete data. They evaluate the predictive expectation $\mathbb{E}[Y_i|\boldsymbol{x}_i, \boldsymbol{D}]$ where $\boldsymbol{D} = \{\boldsymbol{d}_i; i = 1, \ldots, n\}$, from the induced conditional distributions.

MacEachern (1999, 2001) proposed the dependent Dirichlet process (DDP), which generalises the "stick-breaking" construction, see page 51, as an alternative approach to define a dependent prior model for a set of random measures $G_{\boldsymbol{x}}$. This is done by assuming fixed weights $\boldsymbol{\pi} = (\pi_h)_{h=1}^{\infty}$ while allowing the atoms $\boldsymbol{\theta} = (\theta_h)_{h=1}^{\infty}$ to vary with $\boldsymbol{x}_i$ according to a stochastic process. Another way to allow dependence in random measures is to allow the random measures to depend on a shared set of latent factors, which are assigned Dirichlet process priors (Gelfand and Kottas, 2001). The dependent Dirichlet process has been applied to spatial modelling (Gelfand et al., 2005), functional data (Dunson and Herring, 2006), and time series (Caron et al., 2008) applications.

Griffin and Steel (2006) incorporate dependency by allowing the ordering of the random variables $\boldsymbol{\beta} = (\beta_h)_{h=1}^{\infty}$ in the "stick-breaking" construction to depend on predictors. They propose an order-based DDP which does not have the fixed weights assumption of the "simple" DDP.

Dunson (2007) proposed an empirical Bayes approach to density regression, relying on a local mixture of parametric regression models which borrows information by using a kernel-weighted urn scheme. This urn scheme incorporates two smoothing parameters which control the generation of new clusters and borrowing of information. Dunson and Park (2008) propose a class of kernel "stick-breaking" processes for uncountable collections of dependent random probability measures to be used as a prior for $G_{\boldsymbol{x}}$. This prior does not depend on the sample data and induces a covariate-dependent prediction rule upon marginalisation. Hannah et al. (2009) generalise existing Bayesian nonparametric regression models to a variety of response distributions and propose Dirichlet process mixtures of generalised linear models (GLM) which allow both continuous and categorical inputs, and can model the same type of responses that a GLM can. Ghosh et al. (2010) develop a Bayesian density regression model which is based on a logistic Gaussian processes instead of the popular "stick-breaking" process. Wang and Dunson (2011) develop a density regression model that incorporates stochastic-ordering constraints which are natural when a response tends to increase or decrease monotonically with a predictor. For an overview of Bayesian nonparametric inference in density estimation, density regression, and model validation see Müller and Quintana (2004). In the programming language R, *DPpackage* (Jara et al., 2011) includes functions to perform inference via simulation from the posterior distributions for Bayesian nonparametric and semiparametric models. For non-Bayesian approaches to density regression we refer the reader to Fan et al. (1996); Hall et al. (1999, 2004).

Chapter 3

Simulation techniques for intractable likelihoods

In this chapter we propose two simulation techniques for intractable likelihoods. The first one is based on a technique known as retrospective sampling (Papaspiliopoulos and Roberts, 2008) and computes lower and upper bounds for the likelihood. As a result, the target density and the acceptance probability of the Metropolis-Hastings algorithm can be bounded. These bounds can be arbitrarily tight if needed, thus the MCMC algorithm will eventually accept or reject the candidate state $\boldsymbol{\theta}^*$.

The second technique uses rejection sampling to draw from the COM-Poisson distribution and takes advantage of the exchange algorithm (Murray et al., 2006), which is able to draw posterior samples from distributions with unknown normalisation constants. The resulting MCMC algorithms for both techniques sample from the target of interest.

The MCMC algorithms for both techniques are presented and at the end of this chapter we use both algorithms to estimate the parameters μ and ν of a COM-Poisson distribution. The algorithms give similar results in terms of acceptance rates, summary statistics and autocorrelation of the posterior sample.

3.1 Intractable likelihoods

The normalisation constant in the distribution of a random variable may not be available in closed form; in such cases the calculation of the likelihood can be computationally expensive. Specifically, the probability density function $p(y|\boldsymbol{\theta})$ can be written as

$$p(y|\boldsymbol{\theta}) = \frac{q_{\boldsymbol{\theta}}(y)}{Z(\boldsymbol{\theta})} \tag{3.1}$$

where $q_{\boldsymbol{\theta}}(y)$ is the unnormalised density and the normalisation constant $Z(\boldsymbol{\theta}) = \sum_y p(y, \theta)$ or $Z(\boldsymbol{\theta}) = \int p(y, \boldsymbol{\theta})\, \mathrm{d}y$ is unknown. If we use an MCMC algorithm for posterior inference we need to evaluate the acceptance ratio $\min\{1, a\}$ of the Metropolis-Hastings algorithm, which involves the calculation of

$$\begin{aligned} a &= \frac{p(\boldsymbol{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \frac{h(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{h(\boldsymbol{\theta}^*|\boldsymbol{\theta})}, \\ &= \left\{ \prod_{y_i} \frac{q_{\boldsymbol{\theta}^*}(y_i) Z(\boldsymbol{\theta})}{q_{\boldsymbol{\theta}}(y_i) Z(\boldsymbol{\theta}^*)} \right\} \frac{p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta})} \frac{h(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{h(\boldsymbol{\theta}^*|\boldsymbol{\theta})}, \end{aligned} \tag{3.2}$$

where $p(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$, $h(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ is the proposal distribution and[1]

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{y_i} p(y_i|\boldsymbol{\theta}). \tag{3.3}$$

If the proposal distribution $h(|)$ is symmetric in its arguments, then a simplifies to

$$\begin{aligned} a &= \frac{p(\boldsymbol{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \\ &= \left\{ \prod_{y_i} \frac{q_{\boldsymbol{\theta}^*}(y_i)Z(\boldsymbol{\theta})}{q_{\boldsymbol{\theta}}(y_i)Z(\boldsymbol{\theta}^*)} \right\} \frac{p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta})} \end{aligned} \tag{3.4}$$

Both expressions (3.2) and (3.4) cannot be computed due to the ratio of the unknown normalisation constants.

Approximations of the likelihood or approximate Bayesian computation (ABC) methods can be used, but the resulting algorithms may not sample from the target of interest. We will follow two different approaches that both lead to an exact MCMC algorithm.

3.2 Retrospective sampling in MCMC

Ishwaran and James (2001) developed a method for sampling from a Dirichlet process that does not rely on analytically integrating out components of the hierarchical model. This method requires truncating the Dirichlet process prior and as a result it introduces an error. Papaspiliopoulos and Roberts

[1] To simplify the notation we distinguish between the marginal and joint (unnormalised) density by using a scalar (regular) and vector (bold) argument, respectively.

(2008) proposed an exact way of sampling from a Dirichlet process, using the method of Ishwaran and James (2001), known as retrospective sampling. This technique exchanges the order of simulation to implement simulation of infinite dimensional random variables in finite time. Amongst other purposes, it has been used for simulation of diffusion sample paths by Beskos et al. (2006) and Sermaidis et al. (2013). We propose a novel retrospective sampling technique for MCMC.

As mentioned in Chapter 2, in simulation-based Bayesian inference we are interested in drawing samples from the posterior distribution of the parameters $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, where $p(\boldsymbol{\theta})$ denotes the prior distribution of $\boldsymbol{\theta}$. In the Metropolis-Hastings algorithm a Markov chain is constructed in which, when the current state is $\boldsymbol{\theta}$, a candidate state $\boldsymbol{\theta}^*$ is drawn from a proposal distribution $h(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ and then accepted with probability $\min\{1, a\}$ with a as set out in equation (3.2). If $\boldsymbol{\theta}^*$ is rejected, the chain remains at the current state $\boldsymbol{\theta}$. In order to accept the candidate $\boldsymbol{\theta}^*$ with probability $\min\{1, a\}$, the acceptance ratio a is compared to a random $u \sim \text{Unif}(0, 1)$ and $\boldsymbol{\theta}^*$ is accepted if $u < a$.

The key idea of the proposed algorithm is that the acceptance ratio a from (3.2) needs to be known exactly only if the random variable u and a are very close. To be able to exploit this idea we need to exchange the order of simulation. We first draw the uniform random variable u which is used to decide on the outcome of the Metropolis-Hastings acceptance/rejection move and then perform any calculations needed on the acceptance ratio. Depending on the value of u, the acceptance ratio (which involves the normalisation constant Z both in the numerator and denominator) may not be needed to be known

exactly.

Suppose we have a sequence of increasingly and arbitrarily precise lower and upper bounds for $p(\boldsymbol{y}|\boldsymbol{\theta})$, denoted by $\check{p}(\boldsymbol{y}|\boldsymbol{\theta})$ and $\hat{p}(\boldsymbol{y}|\boldsymbol{\theta})$, respectively. Plugging these bounds into (3.2) yields lower and upper bounds for the acceptance ratio

$$\begin{aligned}\check{a}_n &= \frac{\check{p}_n(\boldsymbol{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{\hat{p}_n(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}\frac{h(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{h(\boldsymbol{\theta}^*|\boldsymbol{\theta})},\\ \hat{a}_n &= \frac{\hat{p}_n(\boldsymbol{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{\check{p}_n(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}\frac{h(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{h(\boldsymbol{\theta}^*|\boldsymbol{\theta})}.\end{aligned} \tag{3.5}$$

By construction $\check{a}_n \leq a \leq \hat{a}_n$ and we will assume that we can make bounds arbitrarily precise, i.e. $\check{a}_n \to a$ and $\hat{a}_n \to a$ as $n \to \infty$. The proposed algorithm for deciding on the acceptance of $\boldsymbol{\theta}^*$ then proceeds as follows.

1. Draw $u \sim \text{Unif}(0,1)$ and set the number of refinements $n = 0$.
2. Compute $\check{a}_n$ and $\hat{a}_n$ and compare them to u.
   - If $u \leq \check{a}_n$, accept the candidate value.
   - If $u > \hat{a}_n$, reject the candidate value.
   - If $\check{a}_n < u < \hat{a}_n$, refine the bounds, i.e increase n and return to step 2.

Figure 3.1 illustrates this idea for three possible realisations u_1, u_2, u_3, of the Unif$(0,1)$ distribution, each one leading to a different outcome (accept, refine, reject). Panel 3.1a shows the Metropolis-Hastings strategy where a is the exact value of the acceptance ratio and u_1, u_2, u_3 are the three realisations of the Unif$(0,1)$ distribution. In the case of u_2 we reject the candidate value $\boldsymbol{\theta}^*$

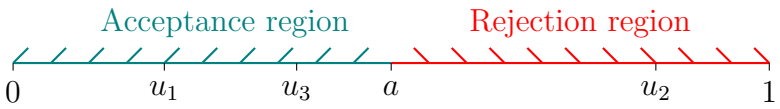
since $u_2 > a$ whereas in the other two cases (u_1 or u_3) we accept the candidate value $\boldsymbol{\theta}^*$ since $u_1, u_3 < a$. The acceptance and rejection regions can also be seen in the figure. Panel 3.1b shows the retrospective sampling strategy along with the refinement region where we compute lower and upper bounds for the acceptance ratio. For the realisations u_1 and u_2 of the Unif$(0, 1)$ distribution we can immediately make a decision since they do not fall into the refinement region. In the case of u_3 we have to refine the bounds. Panel 3.1c shows the new bounds, where the refined lower bound is above u_3 and as a result we accept the candidate value $\boldsymbol{\theta}^*$.

Because the bounds $\check{a}_n$ and $\hat{a}_n$ are arbitrarily tight, the algorithm will eventually accept or reject a candidate value $\boldsymbol{\theta}^*$. The algorithm will reach exactly the same decision on acceptance or rejection as a vanilla Metropolis-Hastings step with the likelihoods computed to full precision, however it has the key advantage that it can reach a decision much more quickly.

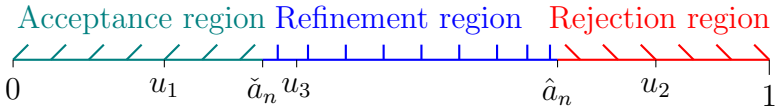
We will assume in the following that the bounds $\check{p}(\boldsymbol{y}|\boldsymbol{\theta})$ and $\hat{p}(\boldsymbol{y}|\boldsymbol{\theta})$ are obtained by bounding each contribution, i.e.

$$\check{p}(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_i \check{p}(y_i|\boldsymbol{\theta}) \qquad \hat{p}(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_i \hat{p}(y_i|\boldsymbol{\theta}) \tag{3.6}$$

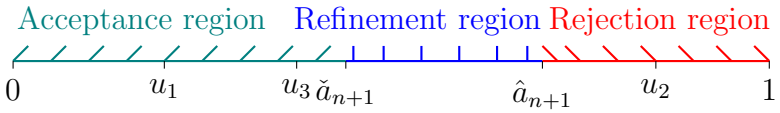
with $\check{p}(y_i|\boldsymbol{\theta}) \leq p(y_i|\boldsymbol{\theta}) \leq \hat{p}(y_i|\boldsymbol{\theta})$. We will now explain how this algorithm can be used in the context of the COM-Poisson distribution. For notational convenience we will omit the subscript i.



(a) Metropolis-Hastings scheme



(b) Retrospective sampling scheme (before refinement)



(c) Retrospective sampling scheme (after refinement)

Figure 3.1: Illustration of the retrospective sampling algorithm (panels b and c) in contrast to the standard Metropolis-Hastings algorithm (panel a).

3.2.1 Piecewise geometric bounds

This section explains how the bounds required in the previous algorithm can be constructed for discrete distributions with probability mass function given by

$$p(y|\boldsymbol{\theta}) = \frac{q_{\boldsymbol{\theta}}(y)}{Z(\boldsymbol{\theta})}, \tag{3.7}$$

where the normalisation constant $Z(\boldsymbol{\theta}) = \sum_y q_{\boldsymbol{\theta}}(y)$ is not available in closed form. The COM-Poisson distribution, in Subsection 2.1.3, is an example of such a distribution.

For ease of presentation we will assume that only computing the right tail is computationally challenging. At the end of the section we will explain how the method can be applied to bounding the left tail as well.

A simple way of reducing the computational burden is to compute the normalisation constant $Z(\boldsymbol{\theta})$ up to a k^{th} term and use this as a lower bound for $Z(\boldsymbol{\theta})$. An upper bound can be obtained by also considering an upper bound for the remaining terms. For the approach to be computationally efficient, k should be chosen to be not too large, which in turn implies that the upper bound for the remaining terms will be rather loose.

On the other hand, if the ratio of consecutive probabilities is bounded by constants over a certain range of y

$$\check{b}_{y_0,y_1} \leq \frac{q_{\boldsymbol{\theta}}(y+1)}{q_{\boldsymbol{\theta}}(y)} \leq \hat{b}_{y_0,y_1}, \quad y \in \{y_0, y_0+1, \ldots, y_1-1\}, \tag{3.8}$$

then tighter bounds can be obtained at little excess computational cost.

We will now construct bounds based on the constants $\check{b}_{y_0,y_1}, \hat{b}_{y_0,y_1}$. These tighter bounds are based on including piecewise bounds on a sequence of increasingly large blocks of probabilities in the tails. This corresponds to using the following lower bound and upper bound for $Z(\boldsymbol{\theta})$:

$$\begin{aligned} \check{Z}(\boldsymbol{\theta}) &= E(\boldsymbol{\theta}) + \check{C}(\boldsymbol{\theta}), \\ \hat{Z}(\boldsymbol{\theta}) &= E(\boldsymbol{\theta}) + \hat{C}(\boldsymbol{\theta}) + \hat{R}(\boldsymbol{\theta}), \end{aligned} \tag{3.9}$$

where $E(\boldsymbol{\theta}) = \sum_{j=0}^{k_1} q_{\boldsymbol{\theta}}(j)$ is obtained by computing the sum of the first k_1 terms exactly. $\check{C}(\boldsymbol{\theta})$ and $\hat{C}(\boldsymbol{\theta})$ are piecewise bounds on blocks of probabilities, computed as set out below. $\hat{R}(\boldsymbol{\theta})$ is an upper bound on the remaining terms.

If (3.8) holds, then for all $j \in \{0, \ldots, r\}$ with $r = y_1 - 1 - y_0$

$$(\check{b}_{y_0,y_1})^j \leq \frac{q_{\boldsymbol{\theta}}(y_0+j)}{q_{\boldsymbol{\theta}}(y_0)} = \frac{q_{\boldsymbol{\theta}}(y_0+1)}{q_{\boldsymbol{\theta}}(y_0)} \cdots \frac{q_{\boldsymbol{\theta}}(y_0+j)}{q_{\boldsymbol{\theta}}(y_0+j-1)} \leq (\hat{b}_{y_0,y_1})^j. \tag{3.10}$$

We can rewrite the sum of the block of $r+1$ probabilities as

$$\sum_{j=0}^{r} q_{\boldsymbol{\theta}}(y_0+j) = q_{\boldsymbol{\theta}}(y_0)\sum_{j=0}^{r}\frac{q_{\boldsymbol{\theta}}(y_0+j)}{q_{\boldsymbol{\theta}}(y_0)}. \tag{3.11}$$

Taking advantage of (3.10) and (3.11) we obtain the bounds:

$$\begin{aligned}
\check{c}_{y_0,y_1}(\boldsymbol{\theta}) &= q_{\boldsymbol{\theta}}(y_0)\sum_{j=0}^{r}(\check{b}_{y_0,y_1})^j = q_{\boldsymbol{\theta}}(y_0)\frac{1-(\check{b}_{y_0,y_1})^{r+1}}{1-\check{b}_{y_0,y_1}} \leq \sum_{j=0}^{r} q_{\boldsymbol{\theta}}(y_0+j),\\
\hat{c}_{y_0,y_1}(\boldsymbol{\theta}) &= q_{\boldsymbol{\theta}}(y_0)\sum_{j=0}^{r}(\hat{b}_{y_0,y_1})^j = q_{\boldsymbol{\theta}}(y_0)\frac{1-(\hat{b}_{y_0,y_1})^{r+1}}{1-\hat{b}_{y_0,y_1}} \geq \sum_{j=0}^{r} q_{\boldsymbol{\theta}}(y_0+j).
\end{aligned} \tag{3.12}$$

These bounds are computed in blocks of probabilities.

Denote by $s = (k_1, \ldots, k_{l_n})$ the sequence of end-points of the piecewise bounds, we then define

$$\begin{aligned}
\check{C}(\boldsymbol{\theta}) &= \sum_{i=1}^{l_n-1}\check{c}_{k_i,k_{i+1}-1}(\boldsymbol{\theta}),\\
\hat{C}(\boldsymbol{\theta}) &= \sum_{i=1}^{l_n-1}\hat{c}_{k_i,k_{i+1}-1}(\boldsymbol{\theta}),
\end{aligned} \tag{3.13}$$

which satisfies

$$\check{C}(\boldsymbol{\theta}) \leq \sum_{j=k_1}^{k_{l_n}-1} q_{\boldsymbol{\theta}}(j) \leq \hat{C}(\boldsymbol{\theta}). \tag{3.14}$$

Finally, using (3.10) and (3.11) for the tail of the distribution, we get

$$\sum_{j=0}^{\infty} q_{\boldsymbol{\theta}}(y_{k_{l_n}}+j) = q_{\boldsymbol{\theta}}(y_{k_{l_n}})\sum_{j=0}^{\infty}\frac{q_{\boldsymbol{\theta}}(y_{k_{l_n}}+j)}{q_{\boldsymbol{\theta}}(y_{k_{l_n}})}. \tag{3.15}$$

and

$$\hat{c}_{k_{l_n},\infty}(\boldsymbol{\theta}) = q_{\boldsymbol{\theta}}(y_{k_{l_n}})\sum_{j=0}^{\infty}(\hat{b}_{k_{l_n},\infty})^j = q_{\boldsymbol{\theta}}(y_{k_{l_n}})\frac{1}{1-\hat{b}_{k_{l_n},\infty}} \geq \sum_{j=k_{l_n}}^{\infty} q_{\boldsymbol{\theta}}(j). \tag{3.16}$$

Thus we obtain the desired result that

$$\check{Z}(\boldsymbol{\theta}) = E(\boldsymbol{\theta}) + \check{C}(\boldsymbol{\theta}) \leq \sum_{j=0}^{\infty} q_{\boldsymbol{\theta}}(j) \leq E(\boldsymbol{\theta}) + \hat{C}(\boldsymbol{\theta}) + \hat{R}(\boldsymbol{\theta}) = \hat{Z}(\boldsymbol{\theta}) \quad (3.17)$$

The bounds on the normalisation constant $\check{Z}(\boldsymbol{\theta}), \hat{Z}(\boldsymbol{\theta})$ and the number of exact terms k_1 should be indexed by n, the number of refinements, since every time there is a need for refinement we have to choose a larger k_1 and the bounds will be different. For the rest of the section we will assume that n is fixed. The bounds are increasingly tight as long as $k_1 \to \infty$. In practice the values of $k_1, \ldots, k_{l_n}$ are chosen depending on $\boldsymbol{\theta}$ and on the magnitude of the previous contribution to the sum made. In our experience, choosing k_s such that $k_{s+1} - k_s = d^s$ for $d \approx 2$ works well in practice.

So far we have set out how to compute bounds for the right tail of the distribution. If the mode of the distribution is large, then it is advisable to use the same strategy for the left tail too. The approach is essentially the same, with the main difference being that the bounds are computed right to left and that the summation will stop at zero. Instead of (3.9), we will have

$$\begin{aligned} \check{Z}(\boldsymbol{\theta}) &= \check{E}(\boldsymbol{\theta}) + \check{C}(\boldsymbol{\theta}), \\ \hat{Z}(\boldsymbol{\theta}) &= \hat{E}(\boldsymbol{\theta}) + \hat{C}(\boldsymbol{\theta}) + \hat{R}(\boldsymbol{\theta}), \end{aligned} \quad (3.18)$$

where $\check{E}(\boldsymbol{\theta}), \hat{E}(\boldsymbol{\theta})$ are lower and upper bounds on $E(\boldsymbol{\theta}) = \sum_{j=0}^{k_1} q_{\boldsymbol{\theta}}(j)$. In a similar fashion as (3.13), one can denote by $t = (t_1, \ldots, t_{l_n} = k_1)$ the sequence of end-points of the piecewise bounds and then define

$$\begin{aligned} \check{E}(\boldsymbol{\theta}) &= \sum_{i=1}^{l_n - 1} \check{e}_{t_i, t_{i+1}-1}(\boldsymbol{\theta}), \\ \hat{E}(\boldsymbol{\theta}) &= \sum_{i=1}^{l_n - 1} \hat{e}_{t_i, t_{i+1}-1}(\boldsymbol{\theta}). \end{aligned} \quad (3.19)$$

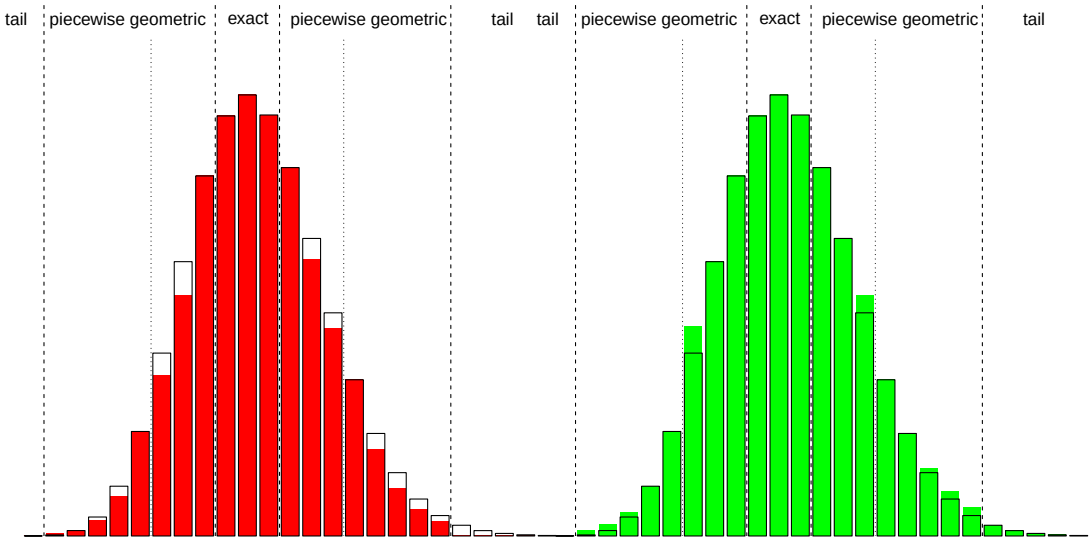


Figure 3.2: Computing the lower and upper bounds of the normalisation constant in blocks of probabilities. We first compute exactly the probabilities close to the mode of the distribution, and then compute the lower and upper bounds in blocks of probabilities where we only have to compute exactly the first probability of each block.

A graphical explanation of the procedure can be seen in Figure 3.2 where the first two blocks are comprised of three and five probabilities respectively.

Bounds on the COM-Poisson normalisation constant

In the COM-Poisson case, the bounds in (3.8) and (3.12) are

$$\begin{aligned}\check{b}_{y_0,y_1} &= \left(\frac{\mu}{y_1}\right)^{\nu},\\ \hat{b}_{y_0,y_1} &= \left(\frac{\mu}{y_0+1}\right)^{\nu},\end{aligned} \tag{3.20}$$

for $y_0, y_1 > \lfloor \mu \rfloor$ and

$$\begin{aligned}\check{c}_{y_0,y_1}(\boldsymbol{\theta}) &= q(y_0)\frac{1-\left(\frac{\mu^{r+1}}{y_1^{r+1}}\right)^{\nu}}{1-\left(\frac{\mu}{y_1}\right)^{\nu}},\\ \hat{c}_{y_0,y_1}(\boldsymbol{\theta}) &= q(y_0)\frac{1-\left(\frac{\mu^{r+1}}{(y_0+1)^{r+1}}\right)^{\nu}}{1-\left(\frac{\mu}{y_0+1}\right)^{\nu}},\end{aligned} \tag{3.21}$$

where $r = y_1 - 1 - y_0$ and $q(y_0) = \left(\frac{\mu^{y_0}}{y_0!}\right)^{\nu}$ is the unnormalised density of the COM-Poisson. Bounds for the left tail of the distribution are computed in a similar way.

Figure 3.3 shows the lower and upper bounds of the normalisation constant as a function of the computed number of terms. The piecewise geometric bounds are plotted in dotted lines, the bounds and the asymptotic approximation of Minka et al. (2003) are plotted with solid lines and a blue line respectively. The true value of the normalisation constant is shown in red. The top panel of Figure 3.3 shows the bounds and true value of the normalisation constant $Z(\mu, \nu)$ for three different pairs of (μ, ν). In all three pairs the second parameter $\nu = 0.2 < 1$. We can see from the top panel that the piecewise geometric bounds do not need to compute a lot of terms to approximate the normalisation constant precisely. As was expected, the asymptotic approximation gets better as μ decreases.

A similar result can be seen on the bottom panel of Figure 3.3. In the three parameter pairs of the bottom panel the second parameter is increased every time. The advantage of using the piecewise geometric bounds, thus first computing the mode of the distribution and afterwards the terms close to the mode, can be seen clearly by looking at the final plot. This plot shows the bounds and true value of the normalisation constant $Z(\mu, \nu)$ where $\mu = 40$ and $\nu = 2$. The bounds in Minka et al. (2003) need to compute three times more terms to be as precise as the piecewise geometric bounds (Chanialidis et al., 2014).

Bounds on the Weighted Poisson distribution

The bounds, found in (3.8), for the weighted Poisson distribution of Del Castillo and Pérez-Casany (1998) for $r > 0$, see (2.30), are

$$\begin{aligned} \check{b}_{y_0,y_1} &= \frac{\lambda}{y_1}\frac{w_{y_1}}{w_{y_1-1}}, \\ \hat{b}_{y_0,y_1} &= \frac{\lambda}{y_0+1}\frac{w_{y_0+1}}{w_{y_0}}. \end{aligned} \tag{3.22}$$

Similar bounds can be constructed for $r < 0$.

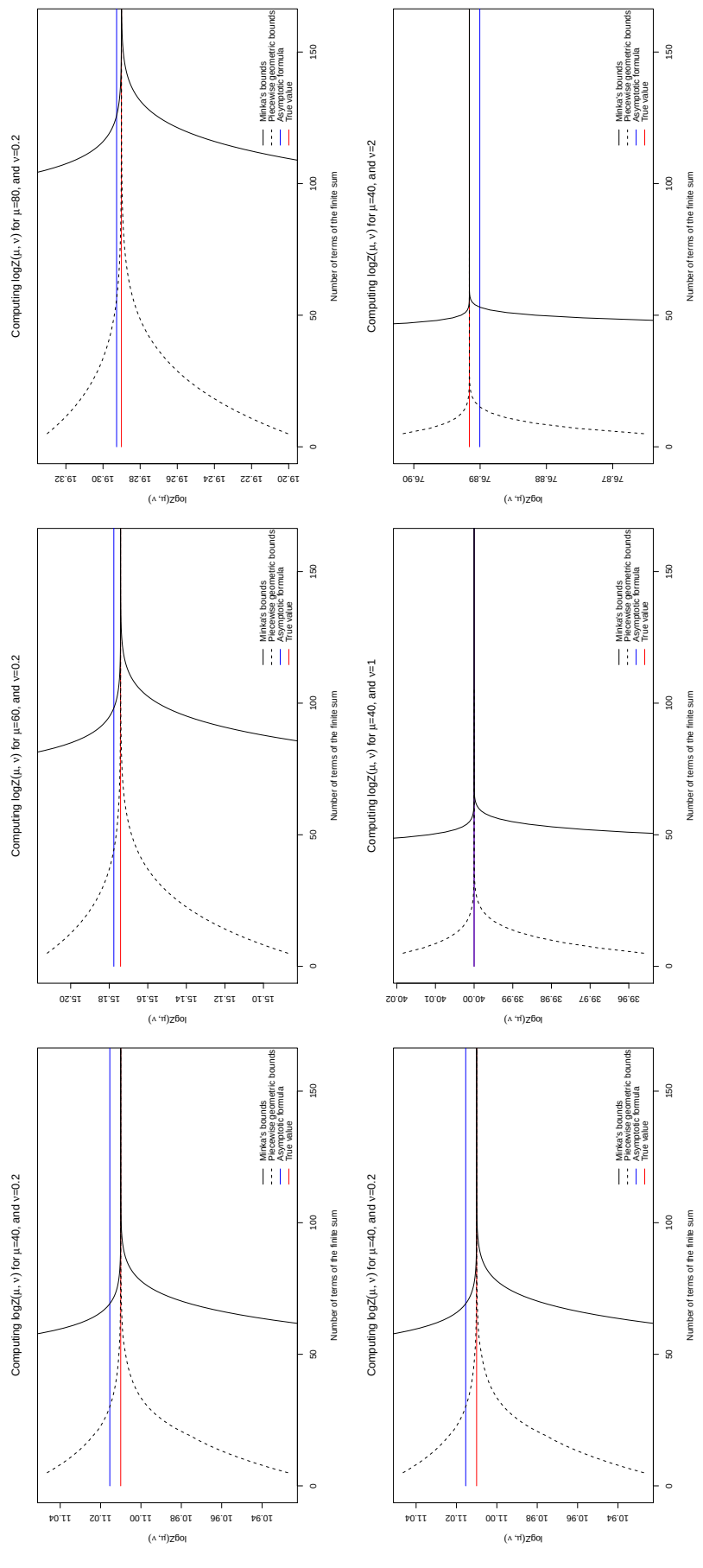


Figure 3.3: Bounds for the normalisation constant $Z(\mu, \nu)$ for different values of μ and ν.

3.2.2 MCMC for retrospective algorithm

We present an overview of the MCMC algorithm that takes advantage of the bounds for the acceptance probability, see (3.5), and the bounds for the normalisation constant $Z(\mu, \nu)$, see (3.9).

Let $y_1, y_2, \ldots, y_n$ be independent and identically distributed observations from the COM-Poisson distribution with unknown parameters μ and ν. Keeping up with the notation on page 87 this can be seen as having an unknown parameter $\boldsymbol{\theta} = (\mu, \nu)$.

Before we give the full MCMC algorithm we explain how the decision on acceptance or rejection of a single move is decided. Suppose we have a new candidate $\boldsymbol{\theta}^*$, drawn from a proposal $h(|\boldsymbol{\theta})$.

i. Draw one realisation $u \sim \text{Unif}(0, 1)$.

ii. For $n = 1, 2, \ldots$ repeat the following steps until a decision can be made

   α. Evaluate the bounds $\check{a}_n$ and $\hat{a}_n$ from (3.5) by computing the current bounds $\check{Z}_n(\boldsymbol{\theta})$, $\hat{Z}_n(\boldsymbol{\theta})$, $\check{Z}_n(\boldsymbol{\theta}^*)$ and $\hat{Z}_n(\boldsymbol{\theta}^*)$ for the current precision setting (number of exact terms and block lengths of piecewise geometric bounds).

   β. If $u \leq \check{a}_n$, accept $\boldsymbol{\theta}^*$.

   If $u > \hat{a}_n$ reject $\boldsymbol{\theta}^*$.

   Otherwise ($\check{a}_n < u < \hat{a}_n$), refine precision settings (by increasing the number of exact terms and/or changing the block lengths of

piecewise geometric bounds[2]) and go back to step α.

The steps of the MCMC algorithm can be summarised as

1. Start with an initial state $\boldsymbol{\theta}_0 = (\mu_0, \nu_0)$ for the unknown parameter $\boldsymbol{\theta}$.

2. For $k = 1, 2, \ldots, N$.

   (a) Propose a candidate state μ_k^* for the first parameter of the COM-Poisson distribution.

   As a result the candidate state for $\boldsymbol{\theta}$ becomes $\boldsymbol{\theta}_k^* = (\mu_k^*, \nu_{k-1})$.

   (b) Perform the above steps for obtaining an accept/reject decision.

   In the case of acceptance set $\mu_k = \mu_k^*$. In the case of rejection $\mu_k = \mu_{k-1}$.

   (c) Propose a candidate value ν_k^* for the second parameter of the COM-Poisson distribution.

   As a result the candidate state for $\boldsymbol{\theta}$ becomes $\boldsymbol{\theta}_k^* = (\mu_k, \nu_k^*)$.

   (d) Perform the above steps for obtaining an accept/reject decision.

   In the case of acceptance set $\nu_k = \nu_k^*$. In the case of rejection $\nu_k = \nu_{k-1}$.

3. After discarding an initial number of draws (burn-in period), the remaining draws can be regarded as a sample from the target distribution.

[2]see page 95 for details.

3.3 Exchange algorithm

3.3.1 Algorithm

Møller et al. (2006) presented an MCMC algorithm, known as exchange algorithm, for cases where the likelihood function involves an intractable normalisation constant that is a function of the parameters. The only assumption for the algorithm to work is to be able to draw independent samples from the unnormalised density. Møller et al. (2006) introduces an auxiliary variable $\boldsymbol{x}$ on the same space as the data $\boldsymbol{y}$ and extend the target distribution

$$\pi(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta}_0), \tag{3.23}$$

for some fixed $\boldsymbol{\theta}_0$. In this case, the proposal distribution for the joint update $(\boldsymbol{\theta}^*, \boldsymbol{x}^*)$ is

$$h(\boldsymbol{\theta}^*, \boldsymbol{x}^*|\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}) = h_1(\boldsymbol{x}^*|\boldsymbol{\theta}^*)h_2(\boldsymbol{\theta}^*|\boldsymbol{\theta}, \boldsymbol{y}) \tag{3.24}$$

which corresponds to the usual change in parameters $\boldsymbol{\theta} \to \boldsymbol{\theta}^*$, followed by a choice for the auxiliary variable. If one chooses

$$h_1(\boldsymbol{x}^*|\boldsymbol{\theta}^*) = \frac{q_{\boldsymbol{\theta}^*}(\boldsymbol{x}^*)}{Z(\boldsymbol{\theta}^*)} \tag{3.25}$$

where q and Z are the unnormalised likelihood and the normalisation constant respectively, the Metropolis-Hastings acceptance ratio becomes

$$\begin{aligned} a &= \frac{p(\boldsymbol{x}^*|\boldsymbol{\theta}_0, \boldsymbol{y})}{p(\boldsymbol{x}|\boldsymbol{\theta}_0, \boldsymbol{y})} \frac{p(\boldsymbol{\theta}^*|\boldsymbol{y})}{p(\boldsymbol{\theta}|\boldsymbol{y})} \frac{h_1(\boldsymbol{x}|\boldsymbol{\theta})}{h_1(\boldsymbol{x}^*|\boldsymbol{\theta}^*)} \frac{h_2(\boldsymbol{\theta}|\boldsymbol{\theta}^*, \boldsymbol{y})}{h_2(\boldsymbol{\theta}^*|\boldsymbol{\theta}, \boldsymbol{y})} \\ &= \frac{p(\boldsymbol{x}^*|\boldsymbol{\theta}_0, \boldsymbol{y})}{p(\boldsymbol{x}|\boldsymbol{\theta}_0, \boldsymbol{y})} \frac{Z(\boldsymbol{\theta})q_{\boldsymbol{\theta}^*}(\boldsymbol{y})p(\boldsymbol{\theta}^*)}{Z(\boldsymbol{\theta}^*)q_{\boldsymbol{\theta}}(\boldsymbol{y})p(\boldsymbol{\theta})} \frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})Z(\boldsymbol{\theta}^*)}{q_{\boldsymbol{\theta}^*}(\boldsymbol{x}^*)Z(\boldsymbol{\theta})} \frac{h_2(\boldsymbol{\theta}|\boldsymbol{\theta}^*, \boldsymbol{y})}{h_2(\boldsymbol{\theta}^*|\boldsymbol{\theta}, \boldsymbol{y})} \\ &= \frac{q_{\boldsymbol{\theta}^*}(\boldsymbol{y})p(\boldsymbol{\theta}^*)}{q_{\boldsymbol{\theta}}(\boldsymbol{y})p(\boldsymbol{\theta})} \frac{h_2(\boldsymbol{\theta}|\boldsymbol{\theta}^*, \boldsymbol{y})}{h_2(\boldsymbol{\theta}^*|\boldsymbol{\theta}, \boldsymbol{y})} \frac{p(\boldsymbol{x}^*|\boldsymbol{\theta}_0, \boldsymbol{y})}{p(\boldsymbol{x}|\boldsymbol{\theta}_0, \boldsymbol{y})} \frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{q_{\boldsymbol{\theta}^*}(\boldsymbol{x}^*)}, \end{aligned}$$

where, unlike Equation (3.4), every term can be computed. The crucial assumption is that we can draw independent samples from (3.25). The choice of $\boldsymbol{\theta}_0$, even though is not necessary to construct a valid Metropolis-Hastings algorithm, is important on the efficiency of the Markov chain. A point estimate on $\boldsymbol{\theta}_0$ could be the maximum pseudo-likelihood estimate based on the observations $\boldsymbol{y}$.

The aforementioned algorithm can have poor mixing among the parameters. Murray et al. (2006) overcame this problem by introducing auxiliary variables $(\boldsymbol{\theta}^*, \boldsymbol{y}^*)$ and sampling from an augmented distribution

$$\pi(\boldsymbol{\theta}^*, \boldsymbol{y}^*, \boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{y}^*|\boldsymbol{\theta}^*)h(\boldsymbol{\theta}^*|\boldsymbol{\theta}) \tag{3.26}$$

where $h(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ is the proposal distribution and whose marginal distribution for $\boldsymbol{\theta}$ is the posterior of interest. The Metropolis-Hastings acceptance ratio becomes

$$a = \frac{q_{\boldsymbol{\theta}^*}(\boldsymbol{y})p(\boldsymbol{\theta}^*)}{q_{\boldsymbol{\theta}}(\boldsymbol{y})p(\boldsymbol{\theta})} \frac{h(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{h(\boldsymbol{\theta}^*|\boldsymbol{\theta})} \frac{q_{\boldsymbol{\theta}}(\boldsymbol{y}^*)}{q_{\boldsymbol{\theta}^*}(\boldsymbol{x}^*)} \frac{Z(\boldsymbol{\theta})Z(\boldsymbol{\theta}^*)}{Z(\boldsymbol{\theta})Z(\boldsymbol{\theta}^*)}$$

Specifically, for each MCMC update we first generate a new proposed value $\boldsymbol{\theta}^* \sim h(|\boldsymbol{\theta})$ and then draw auxiliary data $\boldsymbol{y}^* \sim p(|\boldsymbol{\theta}^*)$. We accept the newly

proposed value $\boldsymbol{\theta}^*$ with probability $\min\{1, a\}$ with

$$\begin{aligned} a &= \frac{p(\boldsymbol{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)p(\boldsymbol{y}^*|\boldsymbol{\theta})h(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{y}^*|\boldsymbol{\theta}^*)h(\boldsymbol{\theta}^*|\boldsymbol{\theta})}, \\ &= \frac{\left\{\prod_{y_i} \frac{q_{\boldsymbol{\theta}^*}(y_i)}{Z(\boldsymbol{\theta}^*)}\right\} p(\boldsymbol{\theta}^*)h(\boldsymbol{\theta}|\boldsymbol{\theta}^*) \left\{\prod_{y_i} \frac{q_{\boldsymbol{\theta}}(y_i^*)}{Z(\boldsymbol{\theta})}\right\}}{\left\{\prod_{y_i} \frac{q_{\boldsymbol{\theta}}(y_i)}{Z(\boldsymbol{\theta})}\right\} p(\boldsymbol{\theta})h(\boldsymbol{\theta}^*|\boldsymbol{\theta}) \left\{\prod_{y_i} \frac{q_{\boldsymbol{\theta}^*}(y_i^*)}{Z(\boldsymbol{\theta}^*)}\right\}}, \\ &= \frac{\left\{\prod_{y_i} q_{\boldsymbol{\theta}^*}(y_i)\right\} p(\boldsymbol{\theta}^*)h(\boldsymbol{\theta}|\boldsymbol{\theta}^*) \left\{\prod_{y_i} q_{\boldsymbol{\theta}}(y_i^*)\right\}}{\left\{\prod_{y_i} q_{\boldsymbol{\theta}}(y_i)\right\} p(\boldsymbol{\theta})h(\boldsymbol{\theta}^*|\boldsymbol{\theta}) \left\{\prod_{y_i} q_{\boldsymbol{\theta}^*}(y_i^*)\right\}}, \\ &= \frac{\left\{\prod_{y_i} q_{\boldsymbol{\theta}^*}(y_i)\right\} p(\boldsymbol{\theta}^*) \left\{\prod_{y_i} q_{\boldsymbol{\theta}}(y_i^*)\right\}}{\left\{\prod_{y_i} q_{\boldsymbol{\theta}}(y_i)\right\} p(\boldsymbol{\theta}) \left\{\prod_{y_i} q_{\boldsymbol{\theta}^*}(y_i^*)\right\}}, \end{aligned} \tag{3.27}$$

where the normalisation constants cancel out. The last line only holds when we choose a symmetric proposal distribution (e.g. $h(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = h(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$).

Looking at (3.4) and (3.27) we can see that $\frac{q_{\boldsymbol{\theta}}(y_i^*)}{q_{\boldsymbol{\theta}^*}(y_i^*)}$ can be thought of as an importance sampling estimate of $\frac{Z(\boldsymbol{\theta})}{Z(\boldsymbol{\theta}^*)}$.

We can interpret this algorithm as follows. Before every update we have the observed data y_i and the current state of the chain $\boldsymbol{\theta}$. We now simulate new data y_i^* using the newly proposed parameter value $\boldsymbol{\theta}^*$ and contemplate whether we should perform a swap, i.e. pair the observed data y_i with the candidate $\boldsymbol{\theta}^*$ and pair the auxiliary data y_i^* with the current $\boldsymbol{\theta}$.

In order to be able to use the algorithm in the next section, one has to be able to sample from the unnormalised density which, in the case of the COM-Poisson distribution can be done efficiently using rejection sampling.

3.3.2 Efficient sampling from the COM-Poisson distribution

Suppose we want to generate a random variable Y from the COM-Poisson distribution with probability mass function $P_{\boldsymbol{\theta}}(Y = y) = \frac{q_{\boldsymbol{\theta}}(y)}{Z(\boldsymbol{\theta})}$ where $q(y) = \left(\frac{\mu^y}{y!}\right)^\nu$ and $Z(\boldsymbol{\theta}) = \sum_y q_{\boldsymbol{\theta}}(y)$. We will now propose a rejection sampling algorithm which uses an instrumental distribution which is based on the piecewise geometric upper bound $\hat{Z}(\boldsymbol{\theta})$ discussed in section 3.2.

As before we cut the range of Y into $l+2$ segments with boundaries $k_1, \ldots, k_l$ such that all boundaries lie to the right of the mode. This idea can be extended to boundaries to the left of the mode. The latter can be beneficial if the mode is large.

We will now construct a discrete distribution whose unnormalised density is identical to the unnormalised density of the COM-Poisson distribution for $y \in \{0, \ldots, k_1\}$. For each segment $y \in \{k_j + 1, \ldots, k_{j+1}\}$ we use a shifted truncated geometric distribution and for $y \in \{k_l, k_l+1, \ldots, \}$ we use a shifted geometric distribution, i.e. using the notation from Section 3.2.

$$r_{\boldsymbol{\theta}}(y) = \begin{cases} q_{\boldsymbol{\theta}}(y) & \text{for } y \in \{0, \ldots, k_1 - 1\} \\ q_{\boldsymbol{\theta}}(k_1) \cdot (\hat{b}_{k_1,k_2-1})^{y-k_1} & \text{for } y \in \{k_1, \ldots, k_2 - 1\} \\ q_{\boldsymbol{\theta}}(k_2) \cdot (\hat{b}_{k_2,k_3-1})^{y-k_2} & \text{for } y \in \{k_2, \ldots, k_3 - 1\} \\ \ldots & \ldots \\ q_{\boldsymbol{\theta}}(k_{l-1}) \cdot (\hat{b}_{k_{l-1},k_l-1})^{y-k_{l-1}} & \text{for } y \in \{k_{l-1}, \ldots, k_l - 1\} \\ q_{\boldsymbol{\theta}}(k_l) \cdot (\hat{b}_{k_{l-1},\infty})^{y-k_l} & \text{for } y \in \{k_l, k_l + 1, \ldots, \} \end{cases} \tag{3.28}$$

The normalisation constant for $r_{\boldsymbol{\theta}}()$ can be computed efficiently, though its calculation is not necessary in this context. Except for the first segment, which can be chosen to be very small, the sum over $r_{\boldsymbol{\theta}}(y)$ is a geometric progression, making it very efficient to evaluate. Sampling from $r_{\boldsymbol{\theta}}(y)$ is also easy. We can first decide which segment to draw from and we can draw within each segment (except for the first) using a closed form formula as the c.d.f. of the shifted truncated geometric distribution has a closed form inverse[3].

By construction (cf. section 3.2) we have that $q_{\boldsymbol{\theta}}(y) \leq r_{\boldsymbol{\theta}}(y)$ for all y. If $g_{\boldsymbol{\theta}}(y) = P_{\boldsymbol{\theta}}(Y = y) = \frac{r_{\boldsymbol{\theta}}(y)}{Z_g(\boldsymbol{\theta})}$ denotes the normalised density corresponding to $r_{\boldsymbol{\theta}}()$, then also $Z(\boldsymbol{\theta}) \leq Z_g(\boldsymbol{\theta})$, and

$$\begin{aligned} p(y|\boldsymbol{\theta}) &= \frac{q_{\boldsymbol{\theta}}(y)}{Z(\boldsymbol{\theta})}, \\ &\leq \frac{r_{\boldsymbol{\theta}}(y)}{Z(\boldsymbol{\theta})}, \\ &= \frac{Z_g(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \frac{r_{\boldsymbol{\theta}}(y)}{Z_g(\boldsymbol{\theta})}, \\ &= M g_{\boldsymbol{\theta}}(y), \end{aligned} \tag{3.29}$$

with $M = \frac{Z_g(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}$. In addition,

$$\frac{p(y|\boldsymbol{\theta})}{M g_{\boldsymbol{\theta}}(y)} = \frac{\frac{q_{\boldsymbol{\theta}}(y)}{Z(\boldsymbol{\theta})}}{\frac{Z_g(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \frac{r_{\boldsymbol{\theta}}(y)}{Z_g(\boldsymbol{\theta})}} = \frac{q_{\boldsymbol{\theta}}(y)}{r_{\boldsymbol{\theta}}(y)} \tag{3.30}$$

i.e. we can decide upon the acceptance and rejection by only considering the unnormalised densities.

[3]Suppose we want to sample from a (truncated) geometric distribution, i.e. $g(y) \propto \alpha^y$ for $y \in \{0, \ldots, \delta\}$. The corresponding c.d.f. is $F(y) = \frac{1-\alpha^{y+1}}{C}$ with $C = 1 - \alpha^{\delta+1}$ ($C = 1$ for the untruncated geometric distribution). The inverse of the c.d.f. is $F^{-1}(u) = \lceil \frac{\log(1-Cu)}{\log(\alpha)} \rceil - 1$.

Finally, we can sample from a COM-Poisson distribution by

i. Draw a realisation $y^* \sim r_{\boldsymbol{\theta}}()$.

ii. Draw a random uniform $u \sim U(0,1)$.

iii. If $u \leq \frac{q_{\boldsymbol{\theta}}(y)}{r_{\boldsymbol{\theta}}(y)}$, then accept y^* as sample from $q_{\boldsymbol{\theta}}()$; otherwise go back to step i.

The proposed method has very high acceptance rates, usually in excess of 90%.

3.3.3 MCMC for exchange algorithm

We now present an overview of the MCMC algorithm for estimating the parameters μ and ν from a COM Poisson distribution which takes advantage of the procedure for sampling set out above and uses the exchange algorithm (Murray et al., 2006).

Let $y_1, y_2, \ldots, y_n$ be independent and identically distributed observations from the COM-Poisson distribution with unknown parameters μ and ν.

The full MCMC algorithm is the same as the one given on page 100, except that the decision on acceptance and rejection is carried out as follows. Suppose we have a new candidate $\boldsymbol{\theta}^*$, drawn from a proposal $h(|\boldsymbol{\theta})$.

i. Draw an auxiliary sample $y_1, \ldots, y_n$ by drawing each y_i from $p(|\theta^*)$ using the rejection sampling method from Section 3.3.2.

ii. Draw one realisation $u \sim \text{Unif}(0, 1)$.

iii. Compute the acceptance ratio a from equation (3.27).

iv. If $u \leq a$, accept $\boldsymbol{\theta}^*$.

If $u > a$ reject $\boldsymbol{\theta}^*$.

Note that in contrast to the retrospective algorithm described in Section 3.2, this algorithm does not lead to the same Markov chain as an exact Metropolis-Hastings sampler would. The simulation of the auxiliary sample $y_1, \ldots, y_n$ acts as an additional stochastic component and thus the performance of the proposed algorithm differs from the performance of an exact Metropolis-Hastings sampler in terms of acceptance rate and autocorrelation, with the former being typically lower and the latter being typically higher. We will assess this difference in more detail in the next section.

3.4 Simulation study comparing the algorithms

In this section we present a small simulation study comparing the two algorithms proposed in this chapter.

Let $y_1, y_2, \ldots, y_n$ be independent and identically distributed observations from the COM-Poisson distribution with parameters $\mu = 10$ and $\nu = 0.8$. We will estimate the parameters of the COM-Poisson distribution using a wide range of values for the number of observations ($n = 10, 100, 1000$). The

prior distributions for the parameters μ and ν are

$$\begin{aligned} \mu &\sim \text{gamma}(10, 1), \\ \nu &\sim \text{gamma}(10, 1), \end{aligned} \tag{3.31}$$

with the same mean and variance (equal to ten). The proposal distributions for the parameters are dependent on the current state of the MCMC. Specifically, the proposal distributions for μ and ν are

$$\begin{aligned} h_\mu &\sim \text{gamma}(\mu^2, \mu), \\ h_\nu &\sim \text{gamma}(\nu^2, \nu). \end{aligned} \tag{3.32}$$

These gamma distributions are centered at the current value of each parameter and have variance equal to 1.

We follow the procedure for both MCMC algorithms, retrospective and exchange, detailed on page 100 and 107 respectively. Figures 3.4 and 3.6 show the trace plots, density, and autocorrelation plots for the parameters when the retrospective MCMC is used (for $n = 100$). Figures 3.5 and 3.7 show the trace plots, density, and autocorrelation plots for the parameters when the exchange MCMC is used (for $n = 100$). Table 3.1 shows summary statistics, effective sample sizes and 95% highest posterior density intervals. The table and the aforementioned figures show that the results from the MCMC algorithms for $n = 100$ are similar. When using a smaller number of observations one can see that the acceptance probability for the retrospective MCMC is higher than the one for the exchange MCMC algorithm. Similarly, the auto-

correlation is smaller for the retrospective MCMC and as a result we get a larger effective sample size.

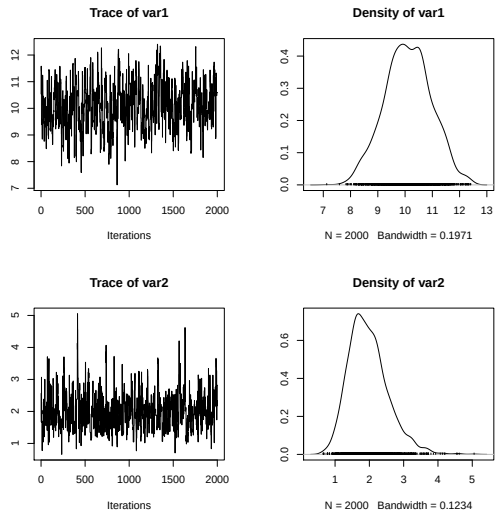


Figure 3.4: Trace plots and density plots for μ and ν using the retrospective MCMC for $n = 100$.

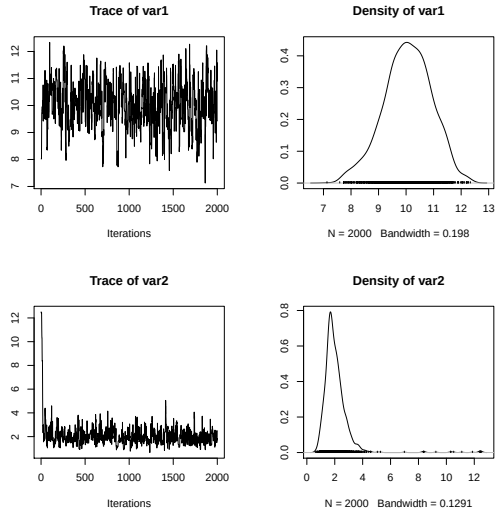


Figure 3.5: Trace plots and density plots for μ and ν using the exchange MCMC for $n = 100$.

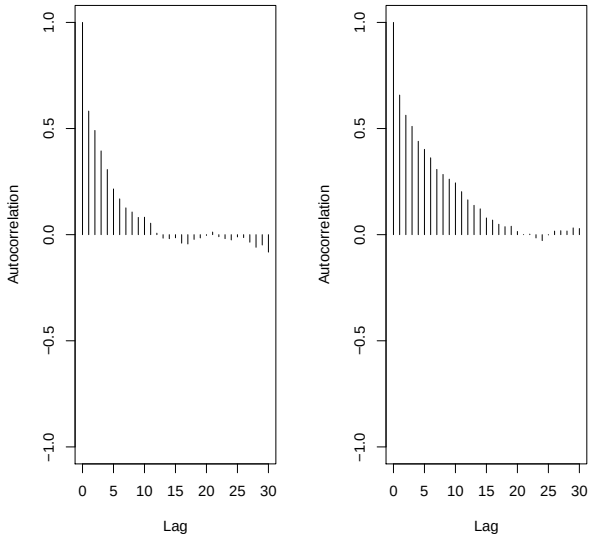


Figure 3.6: Autocorrelation plots for μ and ν using the retrospective MCMC for $n = 100$.

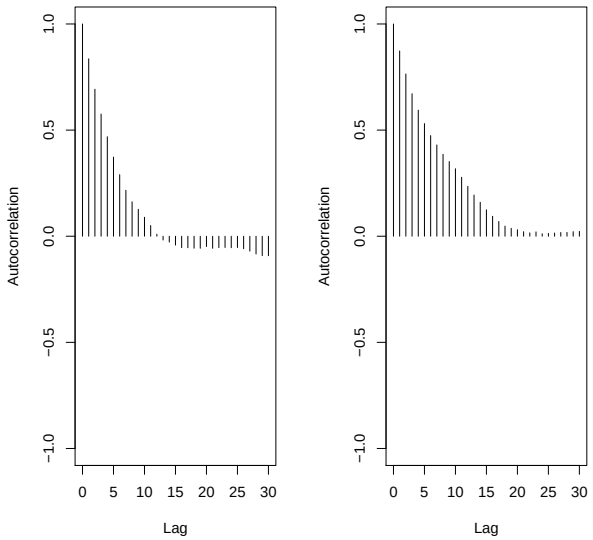


Figure 3.7: Autocorrelation plots for μ and ν using the exchange MCMC for $n = 100$.

Algorithm		Median	Std. deviation	HPD	ESS
Retrospective ($n = 10$)	μ	10.12	0.89	(8.31, 11.73)	198.82
	ν	1.76	1.33	(0.63, 3.22)	183.66
Exchange ($n = 10$)	μ	10.07	0.86	(8.29, 11.66)	178.65
	ν	1.88	1.21	(0.78, 3.29)	135.42
Retrospective ($n = 100$)	μ	10.26	0.36	(9.65, 10.91)	215.33
	ν	1.24	0.29	(0.75, 1.63)	210.12
Exchange ($n = 100$)	μ	10.22	0.29	(9.70, 10.82)	192.40
	ν	1.22	0.28	(0.76, 1.61)	163.53
Retrospective ($n = 1000$)	μ	10.02	0.13	(9.88, 10.34)	283.75
	ν	0.83	0.06	(0.72, 0.97)	256.92
Exchange ($n = 1000$)	μ	9.93	0.12	(9.65, 10.11)	240.02
	ν	0.79	0.04	(0.72, 0.84)	235.46

Table 3.1: Summary statistics for both parameters and both MCMC algorithms (for $n = 10, 100, 1000$).

Chapter 4

Flexible regression models for count data

In this chapter we present two flexible regression models for count data and propose MCMC algorithms based on the simulation techniques of Chapter 3. Regarding the first model, the COM-Poisson regression model, we give some background information on shrinkage priors which, besides being used for variable selection, allow us to have the Poisson regression model as the "baseline" model.

For the second model, the Bayesian density regression model, we show the added flexibility one gains when using a mixture of COM-Poisson regression models for fitting underdispersed and overdispersed distributions.

4.1 COM-Poisson regression

4.1.1 Model

We will implement a Bayesian approach for inference in the model (2.56),

$$
\begin{aligned}
P(Y_i = y_i|\mu_i, \nu_i) &= \left(\frac{\mu_i^{y_i}}{y_i!}\right)^{\nu_i} \frac{1}{Z(\mu_i, \nu_i)}, \\
Z(\mu_i, \nu_i) &= \sum_{j=0}^{\infty} \left(\frac{\mu_i^{j}}{j!}\right)^{\nu_i}, \\
\log\{\mu_i\} = \quad \boldsymbol{x}_i^\intercal \boldsymbol{\beta} \Rightarrow & \qquad \mathbb{E}[Y_i] \approx \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}\}, \\
\log\{\nu_i\} = -\boldsymbol{x}_i^\intercal \boldsymbol{c} \Rightarrow & \qquad \mathbb{V}[Y_i] \approx \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \boldsymbol{x}_i^\intercal \boldsymbol{c}\}.
\end{aligned}
$$

and propose two efficient and exact MCMC algorithms based on the simulation techniques of Chapter 3. For the regression coefficients of the COM-Poisson model we will use both vague and informative priors. We will use independent diffuse Gaussian priors with a mean of zero and a variance of 10^6 as a vague prior. As an informative prior we will use the Bayesian lasso (Tibshirani, 1996) and spike and slab priors (Mitchell and Beauchamp, 1988), which put a penalty on large values of the coefficients. We will focus on a combination of the two approaches by placing a non-informative prior on the "mean" coefficients $\boldsymbol{\beta}$ and an informative prior on the "dispersion" coefficients $\boldsymbol{c}$. This allows us to have the Poisson regression model as the "baseline" model since we put a higher probability on the ν_i being equal to one, compared to the use of vague priors for $\boldsymbol{c}$. Practically, this means that we want to see enough evidence to believe that the Poisson regression model is not appropriate. In addition, these priors can also be used for variable se-

lection, which can be important, especially in the presence of a large number of covariates. For both cases (informative and vague priors), the proposal distribution h is chosen to be a multivariate normal centered at the current value.

4.1.2 Shrinkage priors

In this section we will explain how shrinkage priors can be introduced into the model. We will focus on placing shrinkage priors on $\boldsymbol{c}$, i.e. we will assume a diffuse Gaussian prior for $\boldsymbol{\beta}$. The methods set out below however can also be applied to place a shrinkage prior on $\boldsymbol{\beta}$.

Tibshirani (1996) proposed a method for estimating regression coefficients in linear models. This technique is known as the lasso for "least absolute shrinkage and selection operator". This method minimises the sum of the squared residuals subject to a constraint for the regression coefficients. The constraint is that the sum of the absolute values of the coefficients must be less than a constant. As a result, it shrinks some coefficients and sets others to zero. This method improves the prediction accuracy of the model and can be used for interpretation since it can determine a smaller subset of coefficients with non-zero effects. Tibshirani (1996) suggested that lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e. double exponential) priors.

Park and Casella (2008) consider a fully Bayesian analysis for the lasso method by using a conditional (on the variance) Laplace prior for the regres-

sion coefficients. In the COM-Poisson regression model this can be specified as

$$\begin{aligned} \boldsymbol{c}|t_j^2 &\sim N(\mathbf{0}, \boldsymbol{D}_t), \\ t_j^2|\lambda^2 &\sim \text{exponential}\left(\frac{\lambda^2}{2}\right), \\ \lambda^2 &\sim \text{gamma}(a, b), \end{aligned} \tag{4.1}$$

where $\boldsymbol{D}_t = \text{diag}(t_1^2, \ldots, t_p^2)$. Park and Casella (2008) take advantage of the representation of the Laplace as a scale mixture of Gaussian distributions with an exponential mixing density and they update the regression coefficients and their variances in blocks.

The full conditional densities for the unknown regression coefficients $\boldsymbol{\beta}, \boldsymbol{c}$ do not have a closed form and thus require a Metropolis-Hastings update while on the other hand the posterior densities for the unknown parameters t_j^2, λ^2 in model (4.1) are

$$\begin{aligned} \frac{1}{t_j^2}|\lambda^2, c_j &\sim \text{inverse Gaussian}\left(\sqrt{\frac{\lambda^2}{c_j^2}}, \lambda^2\right), \\ \lambda^2|t_j^2 &\sim \text{gamma}\left(p + a, \frac{1}{2}\sum t_j^2 + b\right). \end{aligned} \tag{4.2}$$

The update of $\boldsymbol{c}$ is the same as in a model with a standard Gaussian prior, the only difference being that, conditionally on the t_j^2, the "prior" covariance of $\boldsymbol{c}$ is given by (4.1).

Mitchell and Beauchamp (1988) proposed another method for variable selection. The idea behind it is that the prior of every regression coefficient is a mixture of a point mass at zero and a diffuse uniform distribution elsewhere. Ishwaran and Rao (2005) proposed a continuous bimodal distribution for the

indicator variables ϕ_j. This form of prior is known as a spike and slab prior and can be specified as

$$\begin{aligned} \boldsymbol{c}|t_j^2, \phi_j &\sim N(\mathbf{0}, \boldsymbol{D}_t), \\ t_j^2|a, b &\sim \text{inverse gamma}(a, b), \\ \phi_j|v_0, v_1, \omega &\sim (1-\omega)\delta_{v_0}() + \omega\delta_{v_1}(), \\ \omega &\sim \text{uniform}(0, 1) \end{aligned} \tag{4.3}$$

where $\boldsymbol{D}_t = \text{diag}(t_1^2\phi_1, \ldots, t_p^2\phi_p)$. The parameter ω controls how likely the binary variable ϕ_j equals v_0 or v_1. Since it controls the size of the models it can be seen as a complexity parameter. It must be noted that one may use a beta prior for w to incorporate prior knowledge. The parameter v_0 should have a positive value equal to zero and the value of v_1 is set to one by default. The indicator ϕ_j takes the value 1 with probability w or some is equal to 0 with probability $1 - w$. The resulting prior for the variance $\boldsymbol{D}_t$ is a bimodal mixture of inverse gamma distributions, where one component is strongly concentrated on very small values (e.g. the spike with $\phi_j = 0$) and a second component with most mass on larger values (e.g. the slab with $\phi_j = 1$).

The full conditional densities for the unknown regression coefficients $\boldsymbol{c}$ in model 4.3 do not have a closed form, and thus require a Metropolis-Hastings update while the full conditional densities for the unknown binary indicator variables ϕ_j are Bernoulli distributions with probabilities

$$\begin{aligned} p_{1,j} &= \frac{1}{1 + \frac{A_{I,j}}{B_{I,j}}}, \\ \frac{A_{I,j}}{B_{I,j}} &= \frac{1-\omega}{\omega}\sqrt{\frac{v_1}{v_0}}\exp\left\{-\frac{1}{2v_0t_j^2}c_j^2 + \frac{1}{2v_1t_j^2}c_j^2\right\}. \end{aligned} \tag{4.4}$$

The full conditionals for the variance parameters t_j^2 are

$$t_j^2|c_j, \phi_j \sim \text{inverse gamma}\left(\frac{1}{2} + a, \frac{c_j^2}{2\phi_j} + b\right) \tag{4.5}$$

while for the complexity parameter ω

$$\begin{aligned} \omega|\phi_j &\sim \text{beta}(1 + n.v_1, 1 + n.v_0), \\ n.v_0 &= \#\{j : \phi_j = v_0\}, \\ n.v_1 &= \#\{j : \phi_j = v_1\}. \end{aligned} \tag{4.6}$$

More details on the Bayesian lasso and spike and slab models can be found in Ishwaran and Rao (2005); Belitz et al. (2009); Kneib et al. (2009).

4.1.3 MCMC for COM-Poisson regression

In Chapter 3 we presented MCMC algorithms for estimating the two parameters of the COM-Poisson distribution. In this section we will set out how these can be adapted to a regression setting. The algorithm stated below will be based on the exchange algorithm presented in Subsection 3.3.3, however one can construct a similar algorithm using the retrospective sampling approach from Subsection 3.2.2. For the sake of simplicity, we describe the algorithm for multivariate Gaussian priors of the coefficients. The modifications needed in order to include shrinkage priors have been described above.

We will start by stating how we can reach a decision on the acceptance or rejection of a set of parameters proposed by an MCMC move. The difference to the setting from Chapter 3 is that due to the regression setting each observation has its own pair of parameters $\boldsymbol{\theta}_i = (\mu_i, \nu_i)$.

To reduce correlation between successive states of the posterior sample, the MCMC consists of two different sets of moves for updating the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{c}$. We alternate between these moves at every sweep of the MCMC. The first proposes a move from $\boldsymbol{\beta}$ to $\boldsymbol{\beta}^*$ and afterwards from $\boldsymbol{c}$ to $\boldsymbol{c}^*$. The second proposes a move from (β_i, c_i) to (β_i^*, c_i^*) for $i = 1, 2, \ldots, p$, where p is the number of variables.

The two parts of the MCMC algorithm can be specified as

A. First part:

1. We draw $\boldsymbol{\beta}^* \sim h(|\beta)$ where the proposal $h()$ is a multivariate Gaussian centered at $\boldsymbol{\beta}$, and

$$\begin{aligned} \boldsymbol{\theta}_i &= (\mu_i, \nu_i), & \boldsymbol{\theta}_i^* &= (\mu_i^*, \nu_i), \\ \mu_i &= \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}\}, & \mu_i^* &= \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta^*}\}, \\ \nu_i &= \exp\{-\boldsymbol{x}_i^\intercal \boldsymbol{c}\}, & \nu_i^* &= \nu_i. \end{aligned} \tag{4.7}$$

We can now use one of the two methods set out below to decide on the acceptance or rejection of the newly proposed value $\boldsymbol{\beta}^*$.

Exchange algorithm

The decision of acceptance or rejection of a proposed value in the exchange algorithm is in principle carried out in the same way as set out on page 107. However we have to draw $y_i^* \sim p(|\boldsymbol{\theta}_i^*)$ and (3.27) becomes

$$a = \frac{\left\{\prod_{y_i} q_{\boldsymbol{\theta_i^*}}(y_i)\right\} p(\boldsymbol{\beta}^*) p(\boldsymbol{c}^*) \left\{\prod_{y_i} q_{\boldsymbol{\theta_i}}(y_i^*)\right\}}{\left\{\prod_{y_i} q_{\boldsymbol{\theta_i}}(y_i)\right\} p(\boldsymbol{\beta}) p(\boldsymbol{c}) \left\{\prod_{y_i} q_{\boldsymbol{\theta_i^*}}(y_i^*)\right\}}. \tag{4.8}$$

In this case we also have to compute

$$q_{\boldsymbol{\theta_i}}(y_i) = \left(\frac{\mu^y}{y!}\right)^{\nu} \qquad q_{\boldsymbol{\theta_i^*}}(y_i) = \left(\frac{(\mu^*)^y}{y!}\right)^{\nu}$$
$$q_{\boldsymbol{\theta_i}}(y_i^*) = \left(\frac{\mu^{y^*}}{y^*!}\right)^{\nu} \qquad q_{\boldsymbol{\theta_i^*}}(y_i^*) = \left(\frac{(\mu^*)^{y^*}}{y^*!}\right)^{\nu}. \tag{4.9}$$

Retrospective sampling

In the retrospective sampling framework we can use the algorithm stated on page 100 to decide on acceptance, rejection or refinement, with the only modification being that we need to compute bounds on the normalisation constant separately for each observation as $\boldsymbol{\theta_i}^*$ and $\boldsymbol{\theta_i}$ are (potentially) different for each observation. Specifically, we have to evaluate the bounds $\check{a}_n$ and $\hat{a}_n$ from (3.5) by computing the current bounds $\check{Z}_n(\boldsymbol{\theta})$, $\hat{Z}_n(\boldsymbol{\theta})$, $\check{Z}_n(\boldsymbol{\theta}^*)$ and $\hat{Z}_n(\boldsymbol{\theta}^*)$ for the current precision setting. If the uniform realisation $u \sim \text{Unif}(0,1)$ is smaller or greater than the acceptance ratio bounds $\check{a}_n$ and $\hat{a}_n$ respectively; then we can make a decision on accepting or rejecting the candidate value $\boldsymbol{\beta}^*$. In any other case we have to refine the bounds until we can make a decision.

2. We now draw $\boldsymbol{c}^* \sim h(|\boldsymbol{c})$ where the proposal $h()$ is a multivariate Gaussian centered at $\boldsymbol{c}$, and

$$\begin{aligned} \boldsymbol{\theta}_i &= (\mu_i, \nu_i), & \boldsymbol{\theta}_i^* &= (\mu_i, \nu_i^*), \\ \mu_i &= \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}\}, & \mu_i^* &= \mu_i, \\ \nu_i &= \exp\{-\boldsymbol{x}_i^\intercal \boldsymbol{c}\}, & \nu_i^* &= \exp\{-\boldsymbol{x}_i^\intercal \boldsymbol{c^*}\}. \end{aligned} \tag{4.10}$$

We can now use one of the two methods set out below to decide on the acceptance or rejection of the newly proposed value $\boldsymbol{c}^*$.

Exchange algorithm

In the case where we use the exchange algorithm we also have to compute

$$\begin{aligned} q_{\boldsymbol{\theta_i}}(y_i) &= \left(\frac{\mu^y}{y!}\right)^{\nu} & q_{\boldsymbol{\theta_i^*}}(y_i) &= \left(\frac{\mu^y}{y!}\right)^{\nu^*} \\ q_{\boldsymbol{\theta_i}}(y_i^*) &= \left(\frac{\mu^{y^*}}{y^*!}\right)^{\nu} & q_{\boldsymbol{\theta_i^*}}(y_i^*) &= \left(\frac{\mu^{y^*}}{y^*!}\right)^{\nu^*}. \end{aligned} \tag{4.11}$$

Retrospective sampling

We have to evaluate the bounds $\check{a}_n$ and $\hat{a}_n$ from (3.5) by computing the current bounds $\check{Z}_n(\boldsymbol{\theta})$, $\hat{Z}_n(\boldsymbol{\theta})$, $\check{Z}_n(\boldsymbol{\theta}^*)$ and $\hat{Z}_n(\boldsymbol{\theta}^*)$ for the current precision setting. If the uniform realisation $u \sim \text{Unif}(0,1)$ is smaller or greater than the acceptance ratio bounds $\check{a}_n$ and $\hat{a}_n$ respectively; then we can make a decision on accepting or rejecting the candidate value $\boldsymbol{\beta}^*$. In any other case we have to refine the bounds until we can make a decision.

B. Second part: For $j = 1, \ldots, p$:

We draw $\beta_j^* \sim h(|\beta_j)$ and $c_j^* \sim h(|c_j)$ where the proposal distribution $h()$ is a univariate Gaussian centered at β_j, c_j respectively and for $\iota \neq j$ copy $\beta_\iota^* = \beta_\iota$ and $c_\iota^* = c_\iota$. Furthermore,

$$\begin{aligned} \boldsymbol{\theta}_i &= (\mu_i, \nu_i), & \boldsymbol{\theta}_i^* &= (\mu_i^*, \nu_i^*), \\ \mu_i &= \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}\}, & \mu_i^* &= \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta^*}\}, \\ \nu_i &= \exp\{-\boldsymbol{x}_i^\intercal \boldsymbol{c}\}, & \nu_i^* &= \exp\{-\boldsymbol{x}_i^\intercal \boldsymbol{c^*}\}. \end{aligned} \tag{4.12}$$

We can now use one of the two methods set out below to decide

on the acceptance or rejection of the newly proposed values β_j^* and c_j^*.

Exchange algorithm

In the case where we use the exchange algorithm we also have to compute

$$\begin{aligned} q_{\boldsymbol{\theta_i}}(y_i) &= \left(\frac{\mu^y}{y!}\right)^{\nu} \qquad & q_{\boldsymbol{\theta_i^*}}(y_i) &= \left(\frac{(\mu^*)^y}{y!}\right)^{\nu^*}, \\ q_{\boldsymbol{\theta_i}}(y_i^*) &= \left(\frac{\mu^{y^*}}{y^*!}\right)^{\nu} \qquad & q_{\boldsymbol{\theta_i^*}}(y_i^*) &= \left(\frac{(\mu^*)^{y^*}}{y^*!}\right)^{\nu^*}. \end{aligned} \tag{4.13}$$

Retrospective sampling

We have to evaluate the bounds $\check{a}_n$ and $\hat{a}_n$ from (3.5) by computing the current bounds $\check{Z}_n(\boldsymbol{\theta})$, $\hat{Z}_n(\boldsymbol{\theta})$, $\check{Z}_n(\boldsymbol{\theta}^*)$ and $\hat{Z}_n(\boldsymbol{\theta}^*)$ for the current precision setting. If the uniform realisation $u \sim \text{Unif}(0,1)$ is smaller or greater than the acceptance ratio bounds $\check{a}_n$ and $\hat{a}_n$ respectively; then we can make a decision on accepting or rejecting the candidate value $\boldsymbol{\beta}^*$. In any other case we have to refine the bounds until we can make a decision.

4.2 Bayesian density regression for count data

Similar to Dunson et al. (2007), we focus on the following mixture of regression models:

$$f(y_i|\boldsymbol{x}_i) = \int f(y_i|\boldsymbol{x}_i, \boldsymbol{\phi}_i) G_{\boldsymbol{x}_i}(\mathrm{d}\boldsymbol{\phi}_i), \tag{4.14}$$

where

$$y_i|\boldsymbol{x}_i, \boldsymbol{\phi}_i \sim \text{COM-Poisson}\left(\exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}_i\}, \exp\{-\boldsymbol{x}_i^\intercal \boldsymbol{c}_i\}\right). \tag{4.15}$$

The conditional density of the response variable given the covariates is expressed as a mixture of COM-Poisson regression models with $\boldsymbol{\phi}_i = (\boldsymbol{\beta}_i, \boldsymbol{c}_i)^\intercal$ and $G_{\boldsymbol{x}_i}$ is an unknown mixture distribution that changes according to the location of $\boldsymbol{x}_i$.

Compared to the continuous model described in Section 2.6, three complications arise from the use of the COM-Poisson distribution.

- Rather than just having a set of mean regression parameters for each cluster, we now need regression parameters for both the mean and the excess variance.
- The full conditional distributions of these regression parameters are not available in closed from.
- Evaluating the normalisation constants of the COM-Poisson distribution is expensive.

These complications are dealt with on the first step of the MCMC algorithm presented in the next subsection. Even though for the first Step of the MCMC algorithm of Dunson et al. (2007), presented in Subsection 2.6.5, it was trivial to allocate observations to clusters with probabilities proportional to the posterior weights (2.124); in this case the normalisation constant c is unknown and this makes things more complicated. We overcome this problem by using the conditional prior of the weights, see (2.121), as the proposal

distribution, see algorithm 5 in Neal (2000), and applying a variation of the exchange algorithm for the first step of the proposed MCMC algorithm.

MCMC for Bayesian density regression

The MCMC algorithm alternates between the following steps:

Step 1: Update the cluster allocations S_i for $i = 1, \ldots, n$, using the conditional prior of the weights, see (2.121), as the proposal distribution. Afterwards we draw parameters $\boldsymbol{\theta}^*$ and an observation y_i^* for this proposed allocation. We accept, the parameters and the allocation, with probability $\min\{1, a\}$ with

$$\begin{aligned} a &= \frac{q_{\boldsymbol{\theta}^*}(y_i)p(\boldsymbol{\theta}^*)h(\boldsymbol{\theta}|\boldsymbol{\theta}^*)q_{\boldsymbol{\theta}}(y_i^*)}{q_{\boldsymbol{\theta}}(y_i)p(\boldsymbol{\theta})h(\boldsymbol{\theta}^*|\boldsymbol{\theta})q_{\boldsymbol{\theta}^*}(y_i^*)}, \\ &= \frac{q_{\boldsymbol{\theta}^*}(y_i)q_{\boldsymbol{\theta}}(y_i^*)}{q_{\boldsymbol{\theta}}(y_i)q_{\boldsymbol{\theta}^*}(y_i^*)}, \end{aligned} \tag{4.16}$$

which is the product of the unnormalised likelihoods. We can interpret this step as follows. We avoid the problem of computing the posterior probability of each observation belonging to a cluster by proposing a new allocation using the prior probability and sampling parameters for this allocation. We now simulate new data y^* using the newly proposed parameter value $\boldsymbol{\theta}^*$ and contemplate whether we should perform a swap. In the case that we do swap we also accept the proposed allocation S_i. Even though it is not trivial to update just the allocations, it is simple to update the allocations and the parameters together.

Specifically,

a) if the proposed move is to go to a new cluster we draw a parameter $\boldsymbol{\theta}^* = (\mu_0, \nu_o)$ for that cluster from G_0 and at the same time sample an observation y^* from the COM-Poisson(μ_0, ν_0). We accept the proposed value $\boldsymbol{\theta}^*$ with probability $\min\{1, a\}$ with a as in (4.16). In this case we accept the proposed move to the new cluster.

If the proposal is accepted, $C_{S_i} \sim$ multinomial $(\{1, \ldots, n\}, \boldsymbol{b}_i)$.

b) If the proposed move is to an already existing cluster h, we sample an observation y^* from the COM-Poisson(μ_h, ν_h) and accept with the same probability as in (4.16). If the proposal is accepted, $C_{S_i} = C_h$.

Step 2: Update the parameters $\boldsymbol{\theta}_h$, for $h = 1, \ldots, k$ by sampling from the conditional posterior distribution

$$(\boldsymbol{\theta}_h | \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{\theta}^{(h)}, k, \boldsymbol{y}, \boldsymbol{X}) \sim \prod_{i:S_i=h} f(y_i | \boldsymbol{x}_i, \boldsymbol{\theta}_h) G_0(\boldsymbol{\theta}_h), \tag{4.17}$$

using the Metropolis-Hasting algorithm with acceptance probability as in (3.27).

Step 3: Update C_h, for $h = 1, \ldots, k$, by sampling from the multinomial conditional with

$$(C_h | \boldsymbol{S}, \boldsymbol{C}^{(h)}, \boldsymbol{\theta}, k, \boldsymbol{y}, \boldsymbol{X}) \sim \frac{\prod_{i:S_i=h} b_{ij}}{\sum_{l=1}^{n} \prod_{i:S_i=h} b_{il}}, \; j = 1, \ldots, n. \tag{4.18}$$

Step 4: Update the location weights γ_j for $j = 1, 2, \ldots, n$, by using a data augmentation approach used in Dunson and Stanford (2005); Holmes et al. (2006). Letting $K_{ij} = \exp\{-\psi \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\}$ and $K^*_{ij} = \frac{K_{ij}}{\sum_{l \neq j} \gamma_l K_{il}}$, the conditional likelihood for γ_j is

$$L(\gamma_j) = \prod_{i=1}^{n} \left(\frac{\gamma_j K^*_{ij}}{1 + \gamma_j K^*_{ij}} \right)^{\mathbf{1}(C_{S_i}=j)} \left(\frac{1}{1 + \gamma_j K^*_{ij}} \right)^{\mathbf{1}(C_{S_i} \neq j)}. \tag{4.19}$$

This likelihood can be obtained by using $\mathbf{1}(C_{S_i} = j) = \mathbf{1}(Z_{ij}^* > 0)$, with $Z_{ij}^* \sim \text{Poisson}(\gamma_j \xi_{ij} K_{ij}^*)$ and $\xi_{ij} \sim \text{exponential}(1)$. Updating $\{Z_{ij}^*, \xi_{ij}\}$ and $\{\gamma_j\}$ in Gibbs steps:

1. let $Z_{ij}^* = 0$ if $\mathbf{1}(C_{S_i} \neq j)$ and otherwise $Z_{ij}^* \sim \text{Poisson}(\gamma_j \xi_{ij} K_{ij}^*)\mathbf{1}(Z_{ij}^* > 0)$, for all i and j;
2. $\xi_{ij} \sim \text{gamma}(1 + Z_{ij}^*, 1 + \gamma_j K_{ij}^*)$, for all i and j;
3. letting $\text{gamma}(a_\gamma, b_\gamma)$ denote the prior for γ_j,

$$\gamma_j \sim \text{gamma}(a_\gamma + \sum_{i=1}^{n} Z_{ij}^*, b_\gamma + \sum_{i=1}^{n} \xi_{ij} K_{ij}^*). \tag{4.20}$$

To compare with the MCMC algorithm for continuous data see Subsection 2.6.5.

Predictive density

Our goal is to estimate the predictive density of a future observation y_{new} from a new subject with predictors $\boldsymbol{x}_{\text{new}}$.

Following Dunson et al. (2007) we get

$$\begin{aligned}(y_{\text{new}}|\boldsymbol{x}_{\text{new}}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{\theta}, k, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta, \tau, \boldsymbol{X}) = \\ \omega_{n,0}(\boldsymbol{x}_{\text{new}})\text{COM-Poisson}(y_{\text{new}}; \exp\{\boldsymbol{x}_{\text{new}}^\intercal \boldsymbol{\beta}\}, \exp\{-\boldsymbol{x}_{\text{new}}^\intercal \boldsymbol{c}\}) \\ + \sum_{h=1}^{k} \omega_{n,h}(\boldsymbol{x}_{\text{new}})\text{COM-Poisson}(y_{\text{new}}; \exp\{\boldsymbol{x}_{\text{new}}^\intercal \boldsymbol{\beta}_h\}, \exp\{-\boldsymbol{x}_{\text{new}}^\intercal \boldsymbol{c}_h\})\end{aligned} \tag{4.21}$$

which is a finite mixture of COM-Poisson regression models. The probability weights depend on the location of $\boldsymbol{x}_{\text{new}}$, which allows deviations on the density across the space of the covariates.

To examine any changes in the conditional density across the covariate space we can calculate the expected predictive density by using a large number of iterations (after convergence) and averaging over them.

Chapter 5

Simulations and case studies

This chapter demonstrates the advantages of using the COM-Poisson regression model and Bayesian density regression models we have presented in Chapter 4.

Simulations have been carried out that illustrate the advantages of each model. The former models' ability to separate a covariate's effect on the mean of the response variable from that on the variance is a strength that other models (such as Poisson, negative binomial) do not have. For the latter model, the simulations show that it consistently outperforms the "jittering" method of (Machado and Santos Silva, 2005) in estimating the conditional quantiles of a distribution.

We illustrate the methods by analysing three different real-world data sets: emergency hospital admissions in Scotland for 2010, number of published papers by students during their Ph.D. studies, and the number of births for

women past childbearing age.

For the first two data sets, the advantages of using the COM-Poisson distribution are evident. The COM-Poisson regression model is able to identify places in Scotland with high variability in admissions, a sign of health inequalities. The COM-Poisson model also provides a clearer picture, compared to the regression models used in literature, of what the effect of a covariate on the response variable is. The COM-Poisson regression model shows that covariates that look as if they have an effect on the mean of the response (e.g. the prestige of a Ph.D. students' supervisor) in Poisson or negative binomial regression, actually have an effect on the variance of the response variable instead.

The third, and final, data set shows the strength of the Bayesian density regression model and the advantages over the "simple" COM-Poisson regression model. The Bayesian density regression model is able to fit distributions which are more complex than what can be represented by a single COM-Poisson distribution.

5.1 Simulations

5.1.1 COM-Poisson regression

As already mentioned, the COM-Poisson regression model is a flexible alternative to count data models mainly used in the literature, such as Poisson or negative binomial regression. The key strength of the COM-Poisson regres-

sion model is its ability to differentiate between a covariate's effect on the mean of the response variable and the effect on the (excess) variance. This can be seen if we simulate from the model (2.56),

$$
\begin{aligned}
P(Y_i = y_i|\mu_i, \nu_i) &= \left(\frac{\mu_i^{y_i}}{y_i!}\right)^{\nu_i} \frac{1}{Z(\mu_i, \nu_i)}, \\
Z(\mu_i, \nu_i) &= \sum_{j=0}^{\infty} \left(\frac{\mu_i^{j}}{j!}\right)^{\nu_i}, \\
\log\{\mu_i\} = \quad \boldsymbol{x}_i^\intercal \boldsymbol{\beta} \Rightarrow & \qquad \mathbb{E}[Y_i] \approx \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}\}, \\
\log\{\nu_i\} = -\boldsymbol{x}_i^\intercal \boldsymbol{c} \Rightarrow & \qquad \mathbb{V}[Y_i] \approx \exp\{\boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \boldsymbol{x}_i^\intercal \boldsymbol{c}\},
\end{aligned}
$$

with

$$
\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0.5 \end{pmatrix}, \boldsymbol{c} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \tag{5.1}
$$

where $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3})^\intercal$ and $x_{ij} \sim \mathrm{N}(0, 1)$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, 3$ where n is the number of the observations. The regression coefficients for the mean and the variance are zero for the first two covariates (and the intercept). The third covariate is statistically significant for both the variance and the mean of the distribution, but with opposite sign. Larger values (of the third covariate) correspond to larger mean of the response variable and smaller variance.

We simulate $n = 100$ observations, which have empirical mean and variance of 1.4 and 2.3, respectively. The 95% and 68% credible intervals for the coefficients for the Poisson, negative binomial, and COM-Poisson regression model can be seen in Figures 5.1 and 5.2. Figure 5.1 shows the credible

intervals for the regression coefficients of μ for all the models. The Poisson and negative binomial model falsely assume that the third covariate has a negative effect on the mean of the response variable. This happens due to the covariate having a negative effect on the variance of the response variable. On the other hand, the COM-Poisson regression model correctly identifies all regression coefficients for the mean of the response variable. The credible intervals for the regression coefficients of ν for the COM-Poisson model can be seen in Figure 5.2 where the only statistically significant variable is the last one. For this simulation we have not used any of the shrinkage priors.

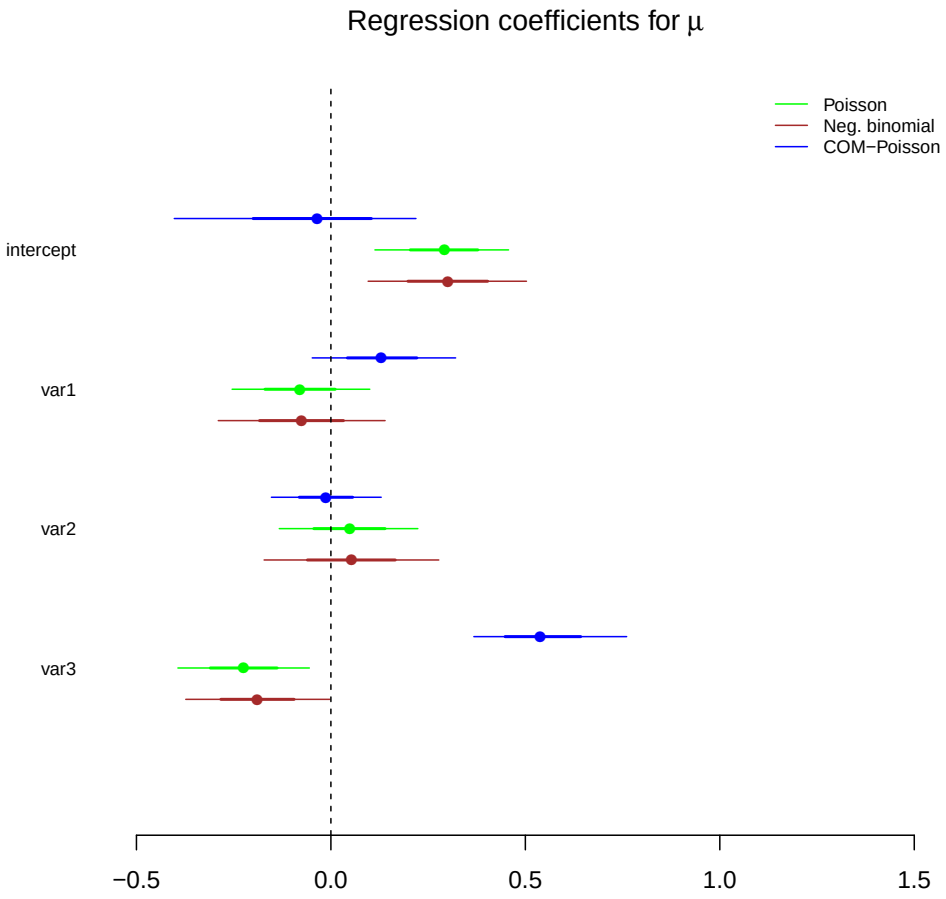


Figure 5.1: Simulation: 95% and 68% credible intervals for the regression coefficients of μ.

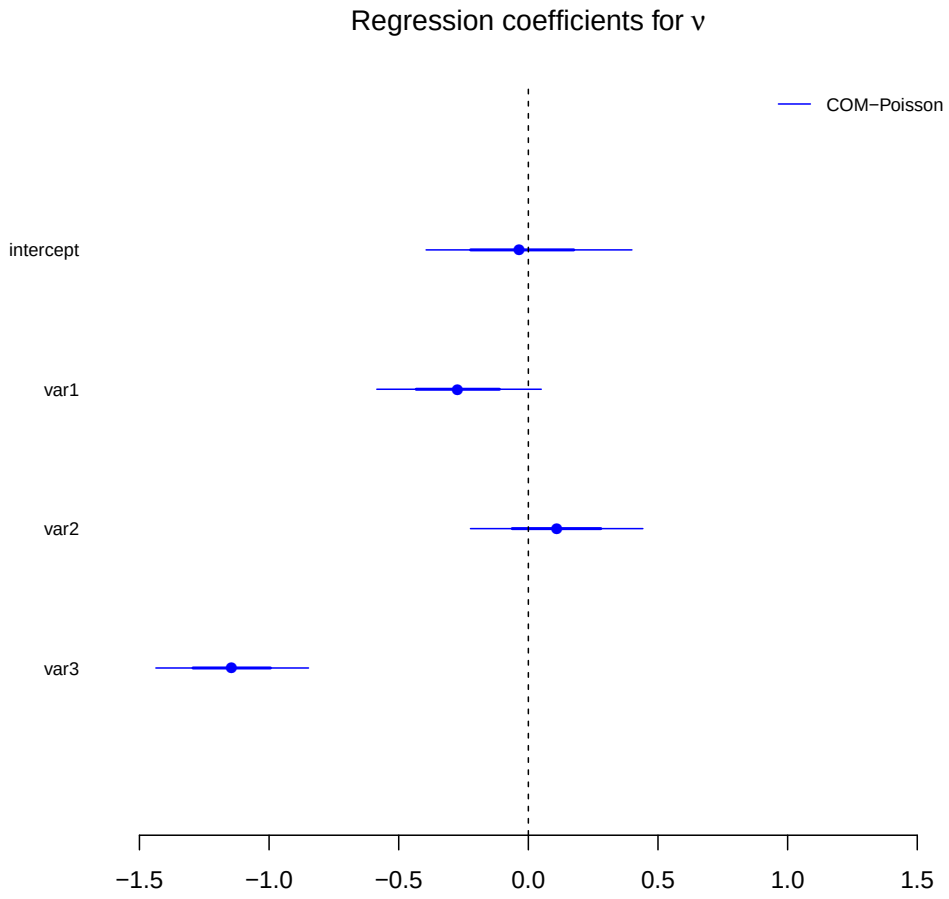


Figure 5.2: Simulation: 95% and 68% credible intervals for the regression coefficients of ν.

5.1.2 Bayesian density regression

For Bayesian density regression, we will use simulations to compare our estimation of the quantiles with the existing method of "jittering". We will simulate data from four different distributions[1]

$$
\begin{aligned}
Y_i|X_i = \boldsymbol{x}_i &\sim \text{COM-Poisson}(\exp\{1 - x_{i1}\}, \exp\{3 + x_{i1}\}),\\
Y_i|X_i = \boldsymbol{x}_i &\sim 0.3\text{COM-Poisson}(\exp\{x_{i1}\}, \exp\{x_{i1}\})\\
&\quad + 0.7\text{COM-Poisson}(\exp\{2 - 2x_{i1}\}, \exp\{1 + x_{i1}\}),\\
Y_i|X_i = \boldsymbol{x}_i &\sim \text{Binomial}(10, 0.3x_{i1}),\\
Y_i|X_i = \boldsymbol{x}_i &\sim 0.4\text{Poisson}(\exp\{1 + x_{i1}\}) + 0.2\text{Binomial}(10, 1 - x_{i1})\\
&\quad + 0.4\text{Geometric}(0.2),
\end{aligned} \tag{5.2}
$$

where $x_{i1} \sim \text{Unif}(0, 1)$ and $\boldsymbol{x}_i = (1, x_{i1})^\intercal$. We will implement the MCMC algorithm presented in Section 4.2, for 10000 iterations with the first 5000 used as a burn-in period. Each simulation has been run for ten different seeds for $n = 50, 100, 500$. Figure 5.3 shows the approximation for the probability mass function of the first simulation, for one of the ten different seeds. The top panel of Figure 5.4 shows the true quantiles (dotted lines) and the estimated quantiles (solid lines) using our approach while the bottom panel shows the true quantiles and the quantiles from using the "jittering" method of Machado and Santos Silva (2005). Figure 5.5 shows the sum of the absolute differences between the true and estimated quantiles accross the covariate space. It can be seen that our method outperforms jittering (with or without splines). Figures 5.6, 5.7 and 5.8 refer to the second simulation. Figures 5.9, 5.10 and 5.11 refer to the binomial distribution while Figures 5.12,

[1] More simulations can be found in Appendix B.

5.13 and 5.14 refer to the final simulation.

In order to compare the methods we will also use the estimated quantiles across the covariate space and compare them with the true quantiles. This will be done for the quantiles $p = (0.01, 0.05, 0.1, 0.15, \ldots, 0.95, 0.99)$ (e.g. 21 different quantiles) and covariate values $x = (0.01, 0.02, \ldots, 0.98, 0.99)$ (99 different covariate values). For each one of the ten simulations we will average over both the quantiles and the covariate values. Finally, we replicate this procedure for all the different realisations of the simulation and then average across them. Specifically, if $q_{p,x}$ is the true conditional quantile at quantile p and when the value of the covariate is x and $\hat{q}_{p,x}$ is an estimate of the conditional quantile then

$$\sum_{q}\sum_{x}|q_{p,x} - \hat{q}_{p,x}| \tag{5.3}$$

gives us the sum of the absolute error accross different quantiles and different values of the covariate. We average over the quantiles and the covariate values (by dividing with 21×99) and averaging accross ten different simulations (by dividing with 10). This can be seen as

$$\frac{1}{10}\frac{1}{99}\frac{1}{21}\sum_{m}\sum_{q}\sum_{x}|q_{p,x} - \hat{q}_{p,x,m}| \tag{5.4}$$

where $\hat{q}_{p,x,m}$ is an estimate of the conditional quantile at quantile p and when the value of the covariate is x for the simulation m. We also have included splines in the quantile regression model of Machado and Santos Silva (2005) and estimated the quantiles for this model.

Table 5.1 shows the integrated average absolute mean errors obtained using both methods for different number of observations. The discrete Bayesian

density regression (BDR) estimates outperform the "jittering" method and in almost all cases the "jittering" method leads to crossing quantiles (except when $N = 500$). Finally, it is important to note that we get similar results for different values of the parameter α of the Dirichlet process. We attribute this to the fact that the distance between covariates' values plays a more important role than the value of α, as far as the cluster assignment of an observation is concerned (cf. page 77).

Method	Number of Observations											
	COM-Poisson			COM-Poisson Mixture			Binomial			Mixture		
	50	100	500	50	100	500	50	100	500	50	100	500
Density Regression	**0.221**	**0.212**	**0.178**	**0.431**	**0.416**	**0.359**	**0.483**	**0.276**	0.245	**0.755**	**0.594**	**0.392**
Jittering (linear)	0.356	0.311	0.217	0.655	0.528	0.440	0.526	0.511	0.465	0.956	0.683	0.454
Jittering (splines)	0.351	0.298	0.205	0.482	0.433	0.388	0.789	0.511	**0.241**	0.947	0.894	0.422

Table 5.1: Integrated mean absolute error obtained using the different density/quantile regression methods.

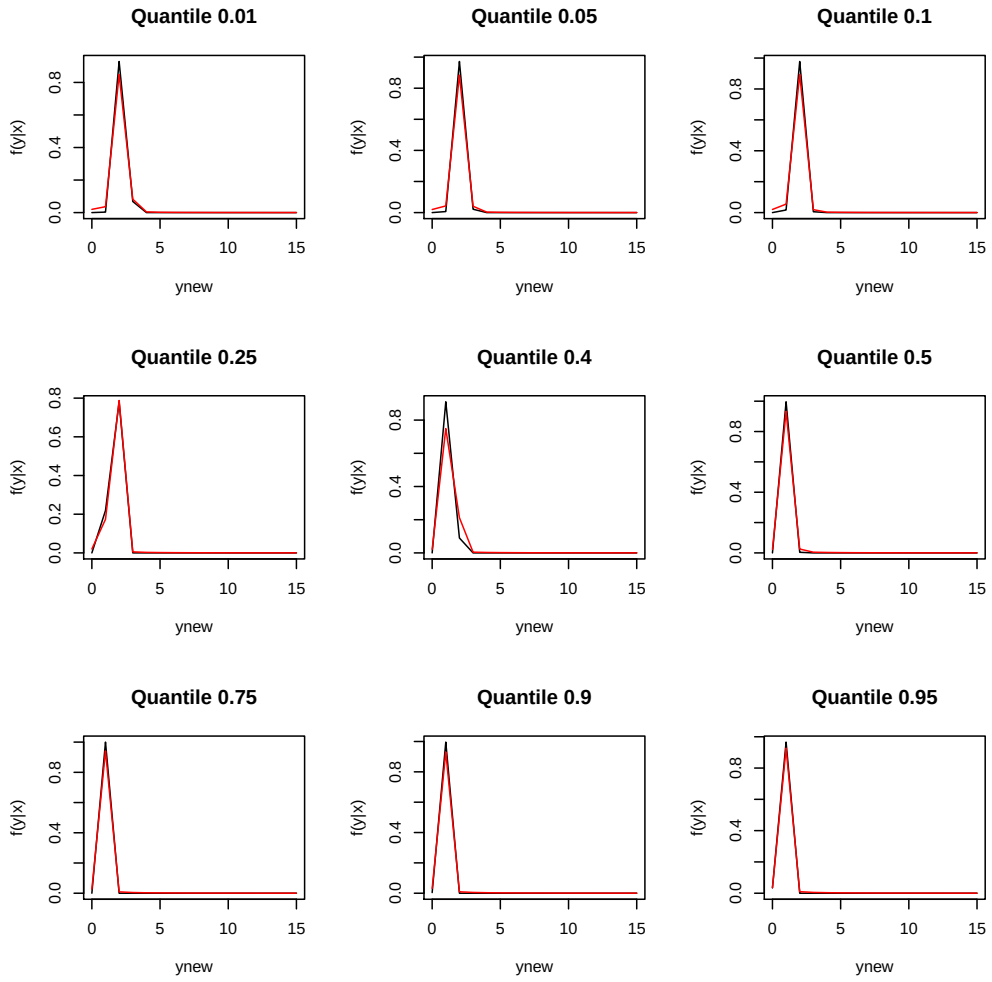


Figure 5.3: True probability mass function of the first distribution in (5.2) is in black and posterior mean estimates are in red. The plots refer to the quantiles $q = 0.01, 0.05, 0.1, 0.25, 0.40, 0.5, 0.75, 0.9, 0.95$ of the empirical distribution of x_{i1}.

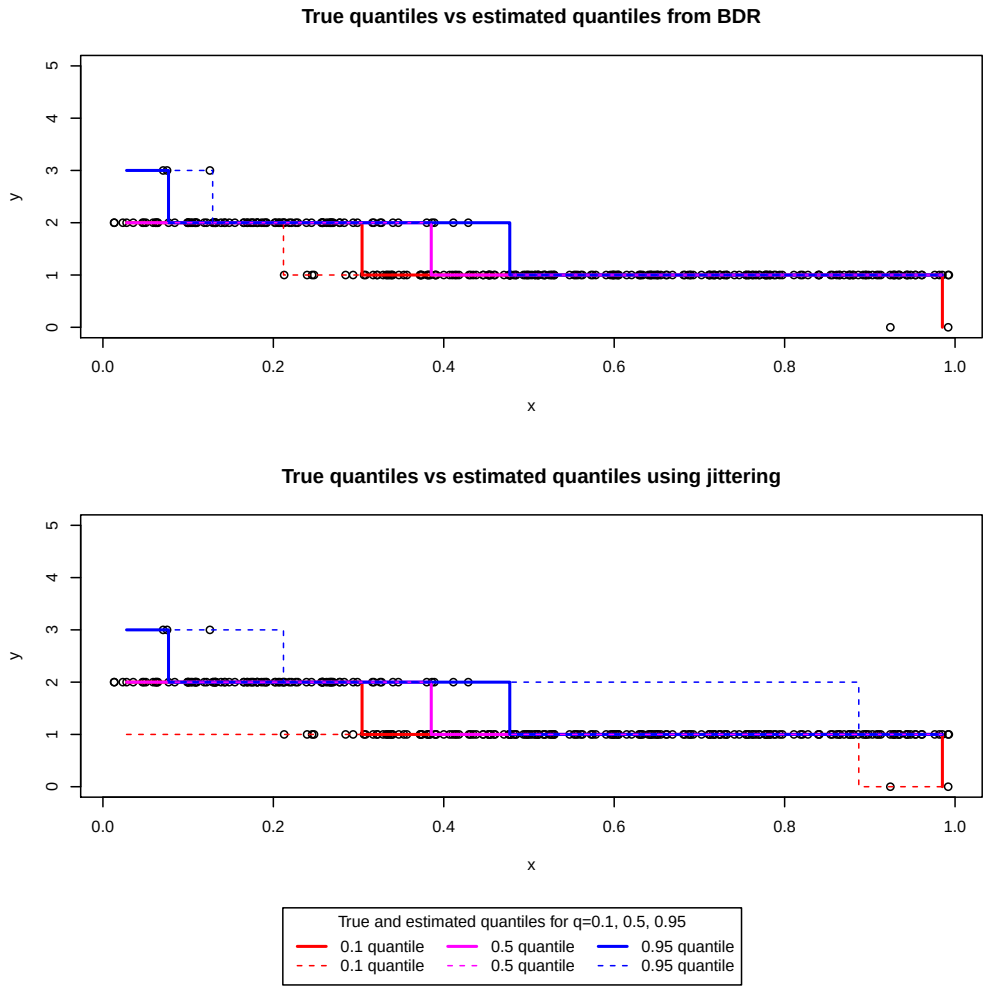


Figure 5.4: The data for the first simulated example, along with the true and estimated quantiles for $q = 0.1, 0.5, 0.95$ across the covariate space.

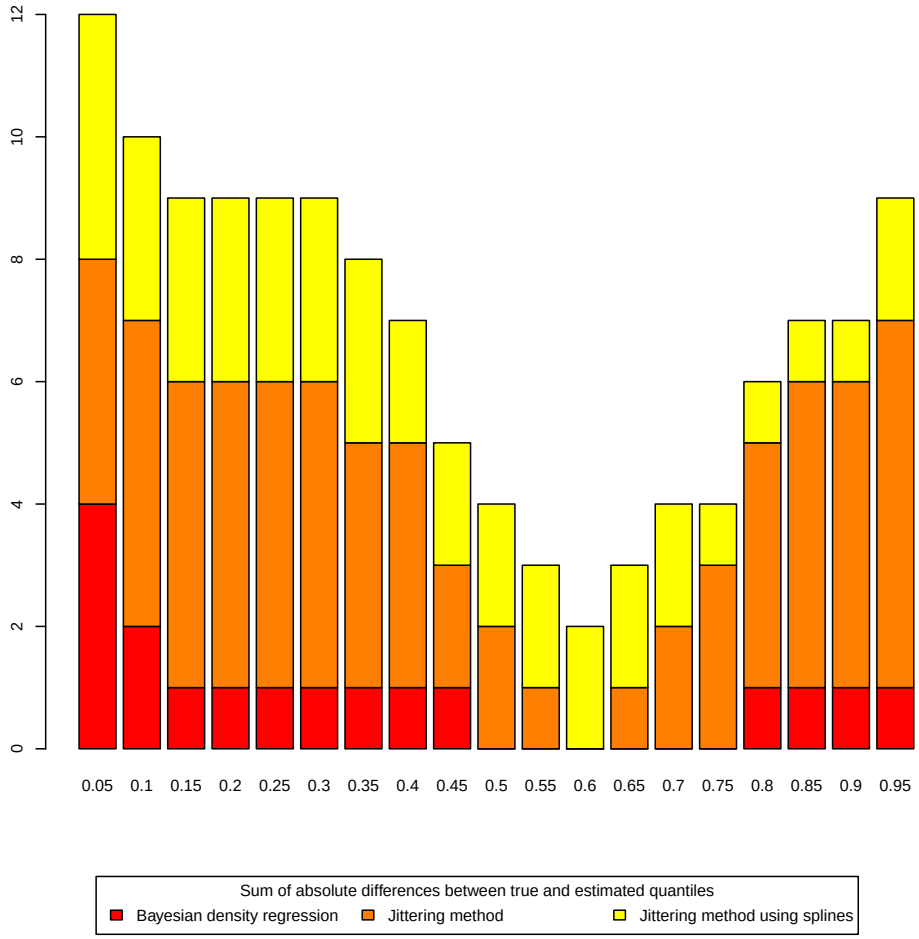


Figure 5.5: Sum of absolute differences between true and estimated quantiles across the covariate space.

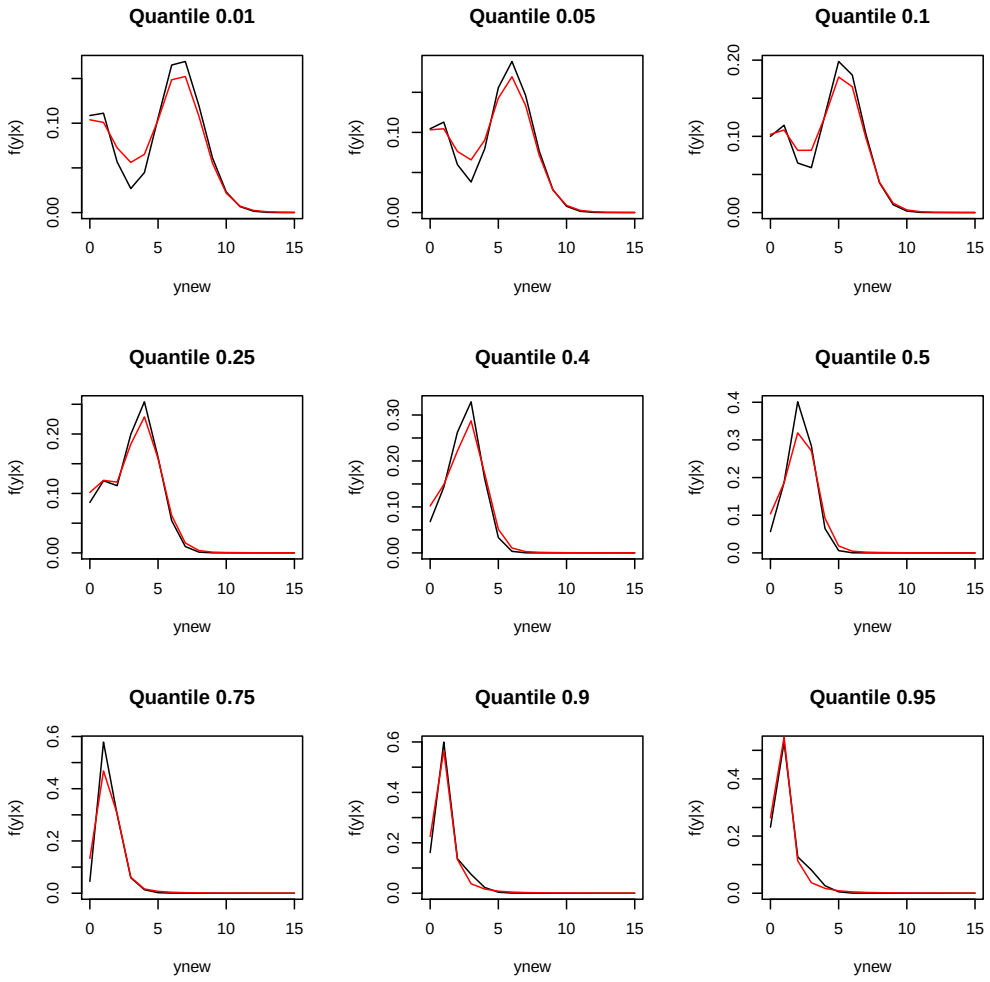


Figure 5.6: True probability mass function of the second distribution in (5.2) is in black and posterior mean estimates are in red. The plots refer to the quantiles $q = 0.01, 0.05, 0.1, 0.25, 0.40, 0.5, 0.75, 0.9, 0.95$ of the empirical distribution of x_{i1}.

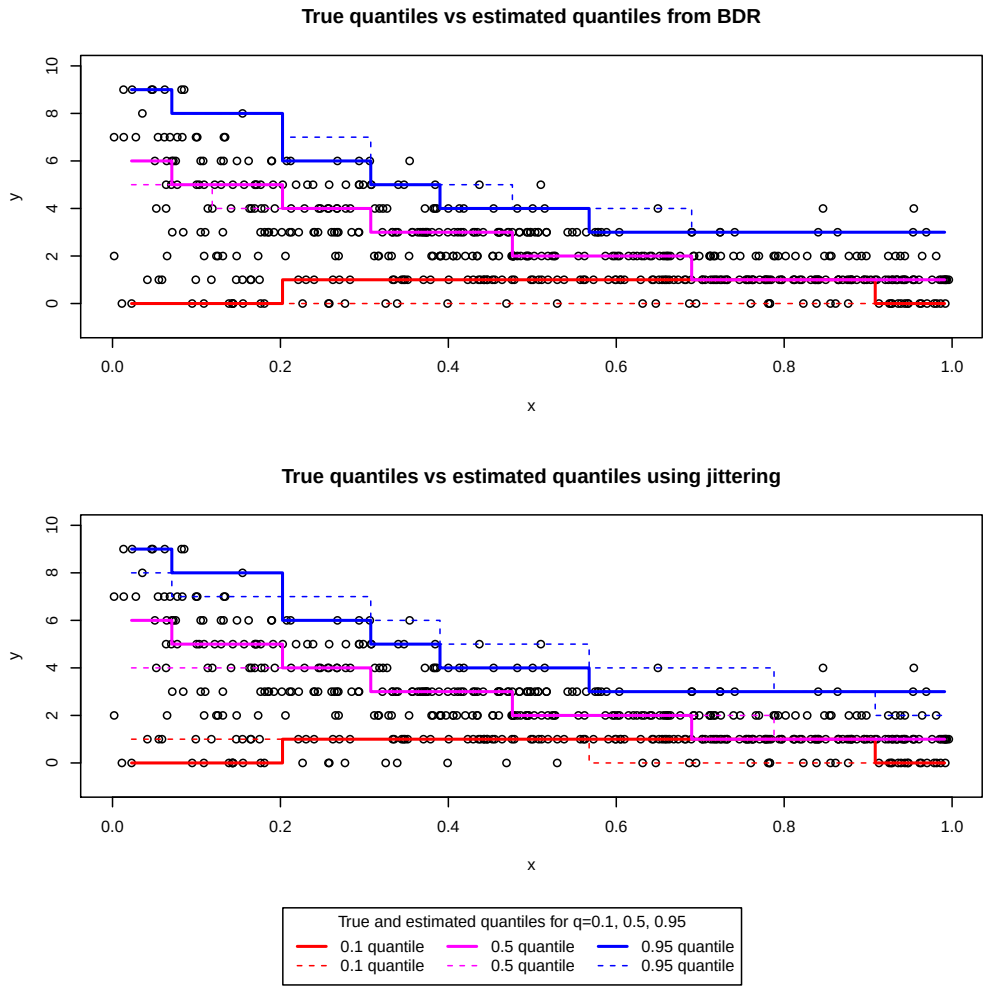


Figure 5.7: The data for the second simulated example, along with the true and estimated quantiles for $q = 0.1, 0.5, 0.95$ across the covariate space.

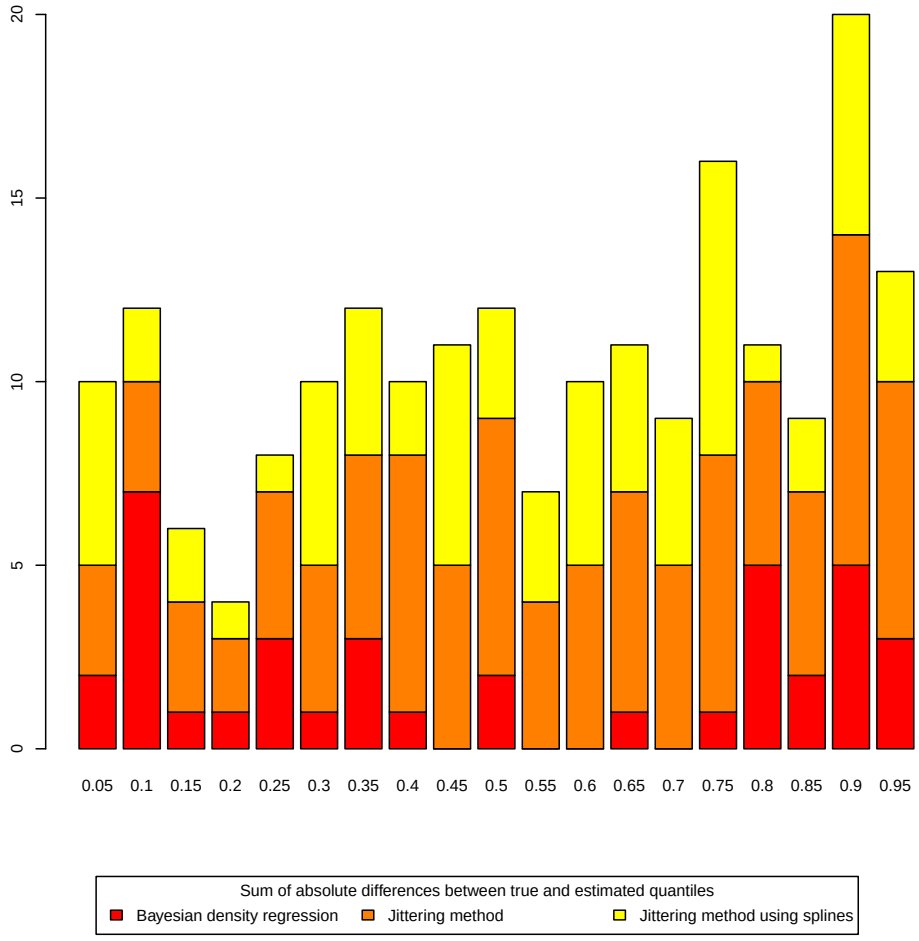


Figure 5.8: Sum of absolute differences between true and estimated quantiles across the covariate space.

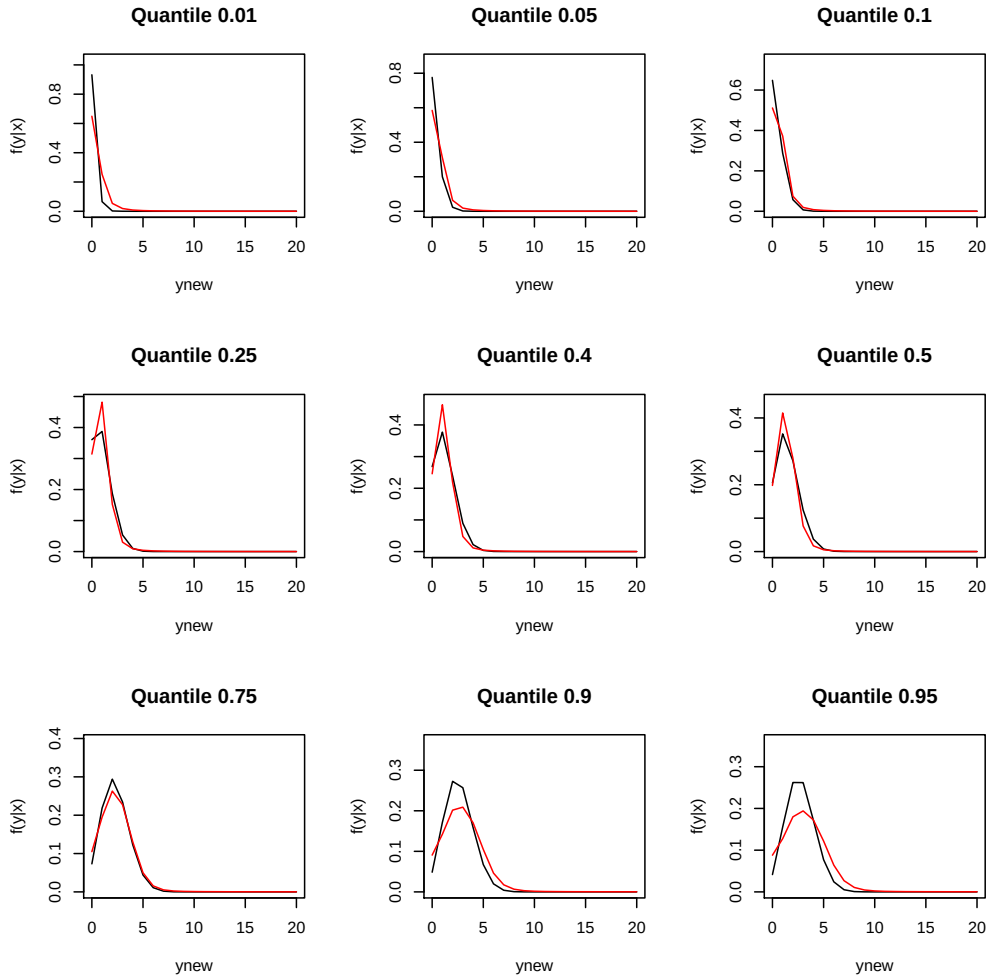


Figure 5.9: True probability mass function of the third distribution in (5.2) is in black and posterior mean estimates are in red. The plots refer to the quantiles $q = 0.01, 0.05, 0.1, 0.25, 0.40, 0.5, 0.75, 0.9, 0.95$ of the empirical distribution of x_{i1}.

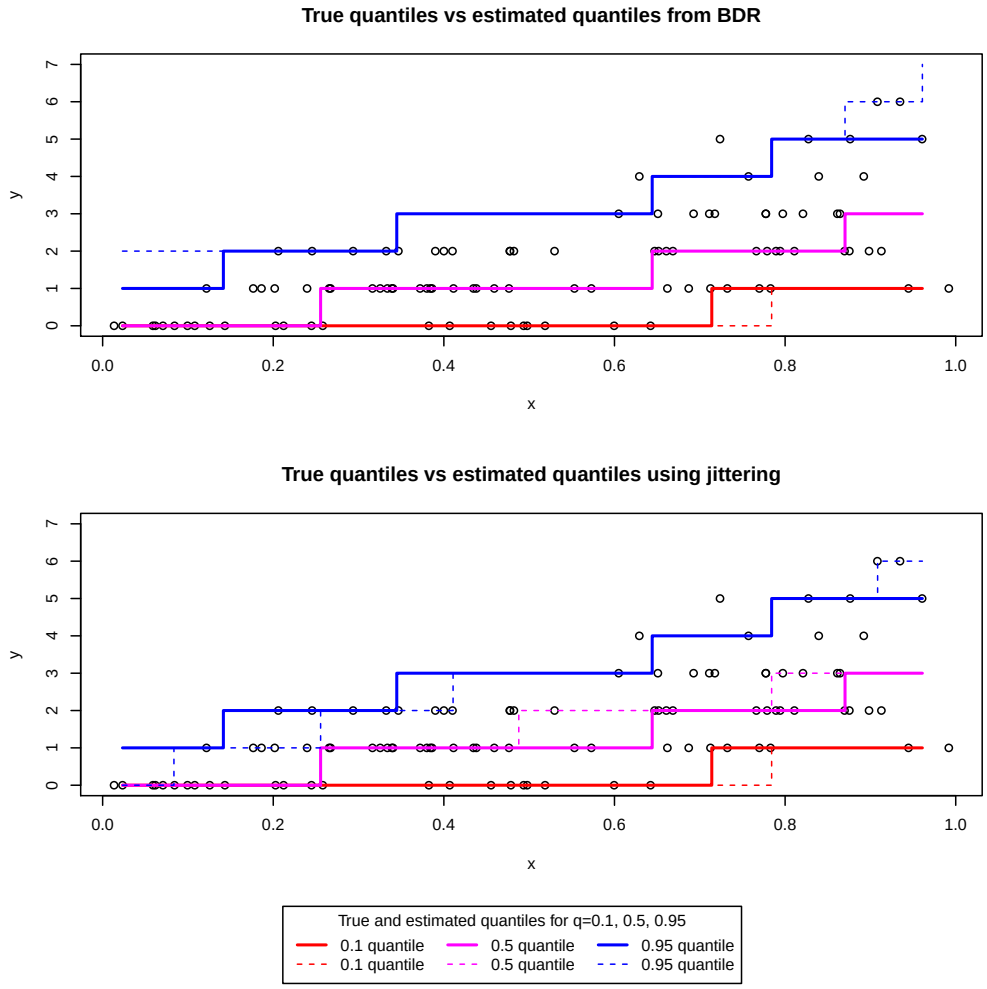


Figure 5.10: The data for the third simulated example, along with the quantiles for $q = 0.1, 0.5, 0.95$ across the covariate space.

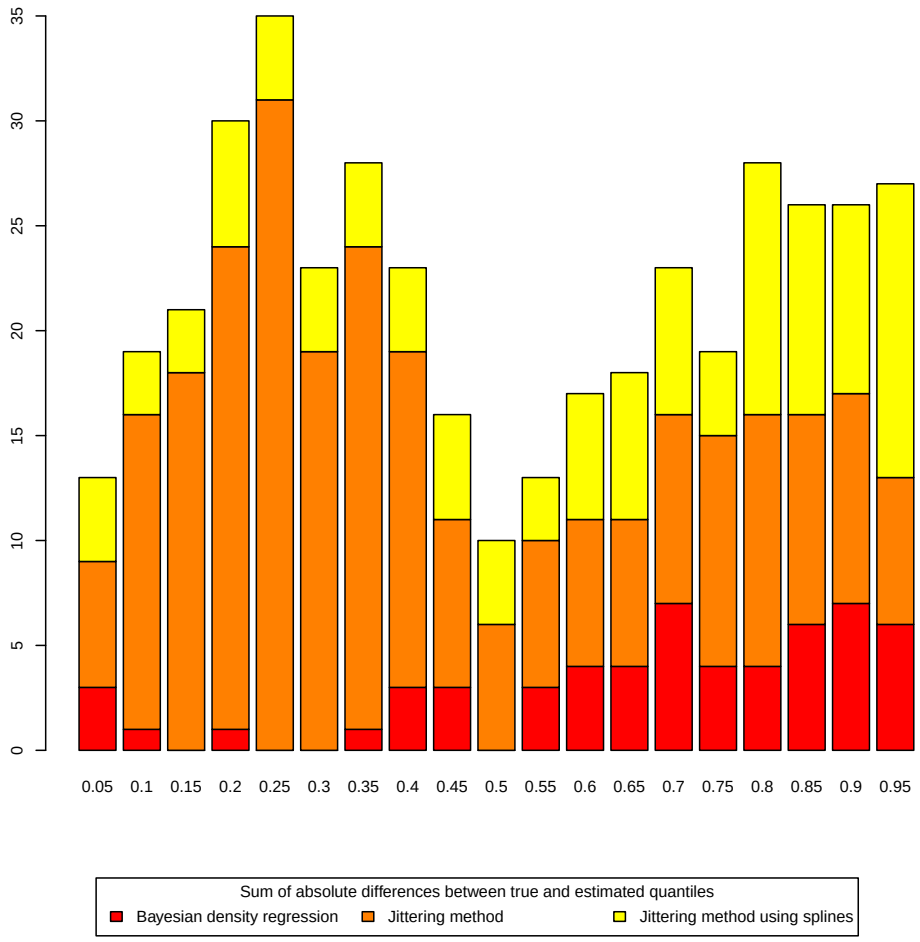


Figure 5.11: Sum of absolute differences between true and estimated quantiles across the covariate space.

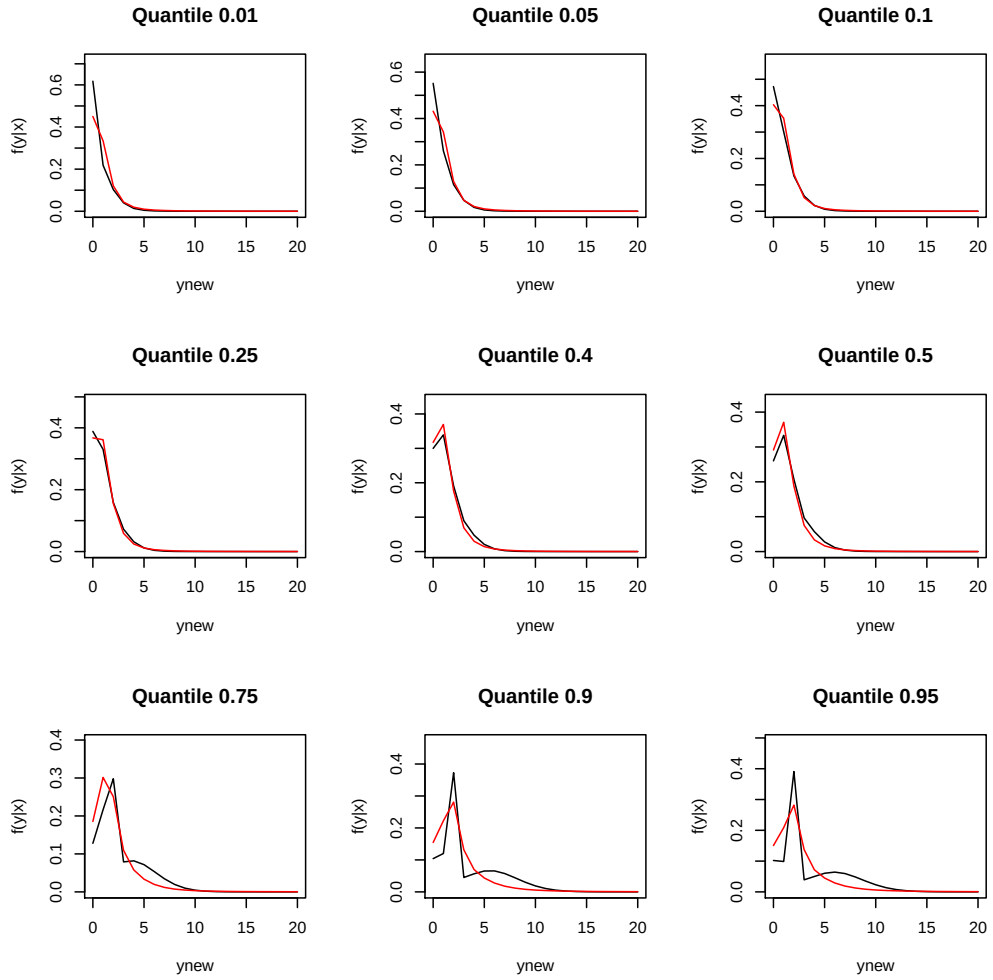


Figure 5.12: True probability mass function of the fourth distribution in (5.2) is in black and posterior mean estimates are in red. The plots refer to the quantiles $q = 0.01, 0.05, 0.1, 0.25, 0.40, 0.5, 0.75, 0.9, 0.95$ of the empirical distribution of x_{i1}.

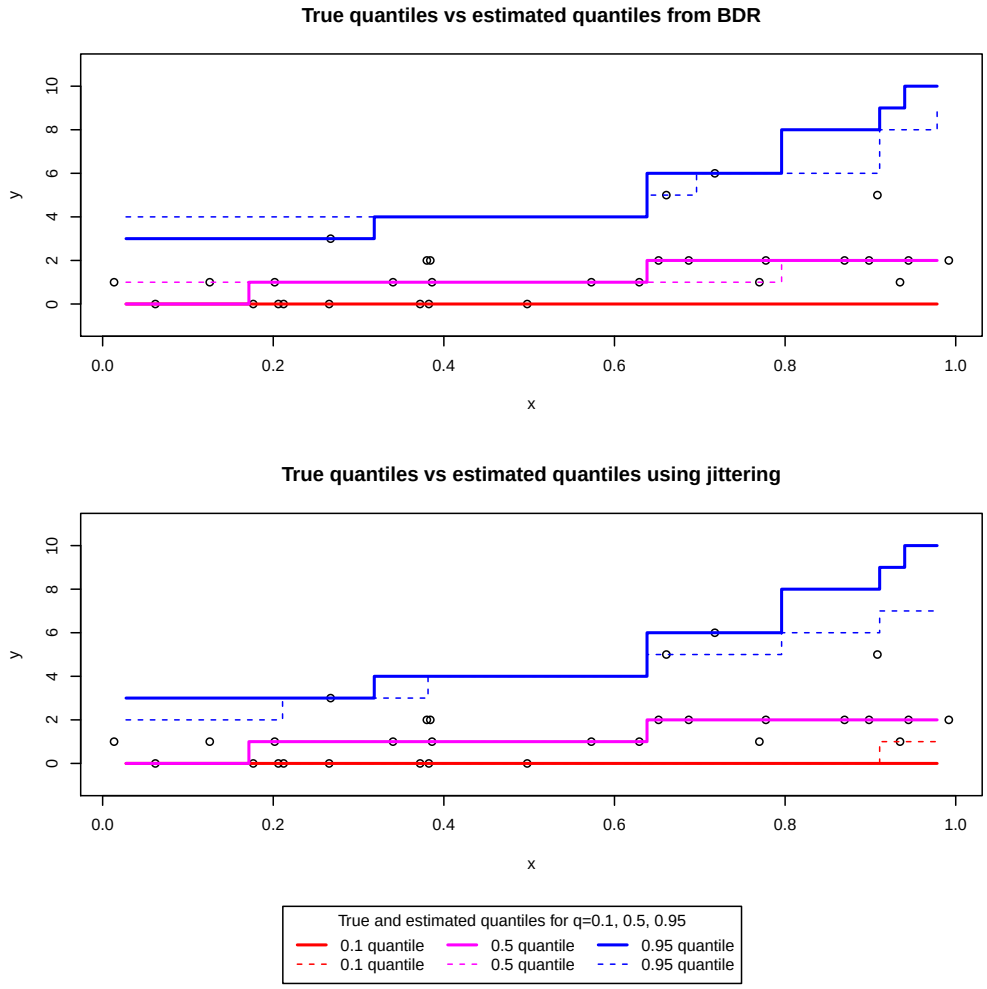


Figure 5.13: The data for the fourth simulated example, along with the quantiles for $q = 0.1, 0.5, 0.95$ across the covariate space.

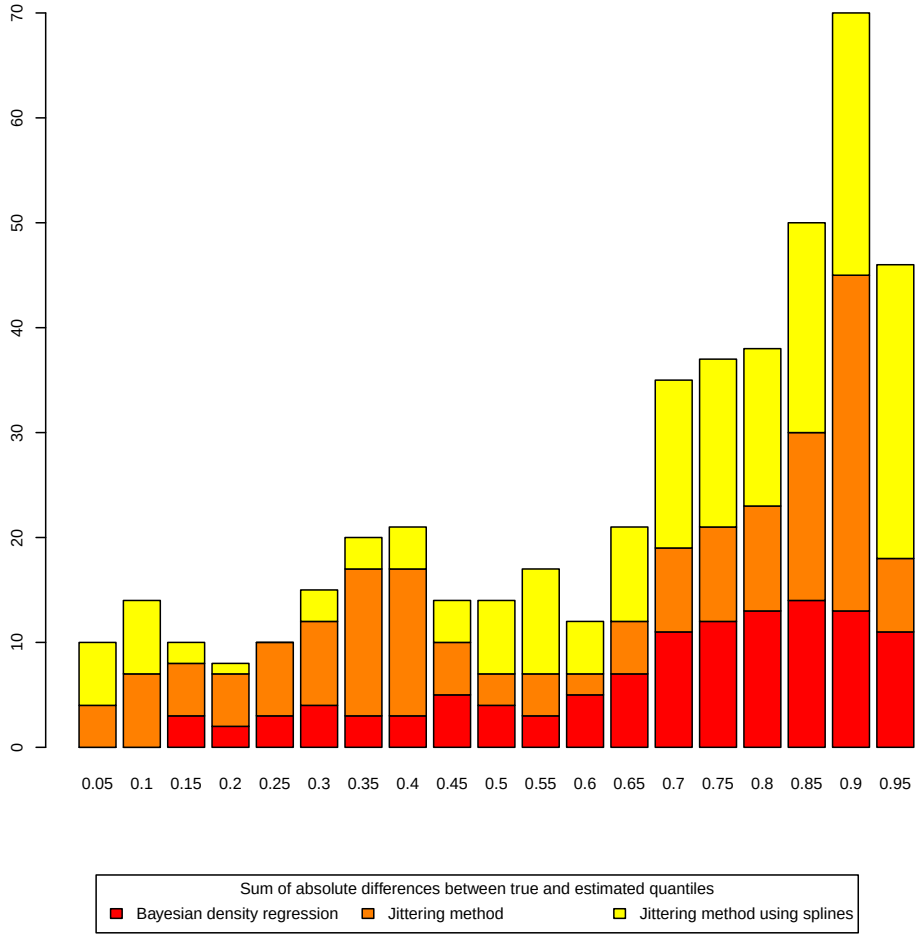


Figure 5.14: Sum of absolute differences between true and estimated quantiles across the covariate space.

5.2 Case studies

5.2.1 Emergency hospital admissions

As an illustration of the COM-Poisson regression method, we consider data on the hospital emergency admissions for each intermediate geography (1235 in total) in Scotland for the year 2010. Scotland is divided into 6505 small areas, called datazones, each containing around 350 households. An intermediate geography is comprised of neighbouring datazones. The Scottish Index of Multiple Deprivation (SIMD) is the Scottish government's official tool for identifying datazones suffering from deprivation. This index provides a relative ranking for each datazone, from 1 (most deprived) to 6505 (least deprived). The ranking is based on seven aspects of deprivation: income, employment, health, education, access to services, crime, and housing. Since the ranking itself does not provide any way of identifying areas that are "deprived" versus "not deprived", analysis of the SIMD requires the user to apply a cut-off to identify the most deprived areas. The Scottish government's cut-off for a datazone to be considered deprived is to belong in the 15% most deprived datazones in Scotland. Using the SIMD ranks for areas larger than datazones (such as intermediate geographies) one can consider the percentage of datazones within that intermediate geography that are in the 15% most deprived e.g. if an intermediate geography is comprised of 20 datazones and 10 of them are in the 15% most deprived then its local share is 50%. This can also be applied in larger areas such as local authorities.

Tables 5.3 and 5.4 refer to the local share of deprived datazones for each

local authority in Scotland. Local authorities in the west of Scotland such as Glasgow City, Inverclyde, North Ayrshire, North Lanarkshire, and West Dunbartonshire have a high local share of the most deprived datazones while at the same time their local share of least deprived datazones is small. Parts of the east of Scotland (Edinburgh City, East Lothian) show the opposite trend. It is important to note that the local share percentages of each local authority are not always showing which areas are deprived. Local authorities such as Eilean Siar, Orkney Islands, and Shetland Islands do not have any datazones in the 15% most deprived in the 2009 SIMD. This happens due to the small number of intermediate geographies they are comprised of and not because they are considered to be affluent areas.

This approach, of using a cut-off point for the datazones, has its drawbacks since datazones that just miss the 15% cut-off point are treated the same as the ones that are far away from it. A better approach, and the one followed in the thesis, would be to weight every datazone and average over all the datazones that belong to the same intermediate geography. Datazones with a small SIMD rank (most deprived) will have a higher weight and each datazone's SIMD rank contributes for the deprivation of the intermediate geography they belong to. The advantages of applying the above modifications on the SIMD rank are:

- Including information from all the datazones of an intermediate geography instead of looking only at the 15% cut-off point.
- Penalising intermediate geographies which are comprised of deprived datazones.

Table 5.2: Levels of income deprivation in Scotland's 15% most deprived areas.

	Number of income deprived people	Total population	% income deprived
15% most deprived areas	232050	742210	31.3%
Rest of Scotland	468430	4479890	10.5%
All of Scotland	700480	5222100	13.4%

It is important to remember that the SIMD identifies areas; not individuals. If our focus is on all deprived people then a different approach needs to be taken, e.g. using the underlying data from one of the domains, rather than the overall index. Table 5.2 shows that not everyone living in a deprived area is deprived, and not all deprived people live in deprived areas, even when looking at individual domains.

The Scottish Government classifies urban and rural areas across Scotland based on two criteria: population and accessibility to areas of contiguous high population density postcodes (that make up what is known as a settlement). The joint classification can be seen in table 5.5. Using this classification as an ordinal covariate is not appropriate due to how it is coded. For example the 6^{th} class (accessible rural area) is closer to an urban area than the previous two. Instead, we will use as covariates the percentages of those classes within each intermediate geography, e.g. if an intermediate geography is comprised of 6 datazones where 3 of them are coded as large urban areas and the other 3 as accessible small towns, the percentages of the first and the third class will be 50% and 0% for all the other classes. For ease of interpretation we center all covariates.

Table 5.3: The local share considers the percentage of a local authority's data-zones that are amongst the 15% most deprived in Scotland.

Local authorities	Local share	Local authorities	Local share
Aberdeen City	10.49	Highland	5.48
Aberdeenshire	1.33	Inverclyde	38.18
Angus	4.23	Midlothian	3.57
Argyll & Bute	8.20	Moray	0.86
Clackmannanshire	18.75	North Ayrshire	24.02
Dumfries & Galloway	5.70	North Lanarkshire	21.29
Dundee City	30.17	Orkney Islands	0.00
East Ayrshire	17.53	Perth & Kinross	3.43
East Dunbartonshire	3.15	Renfrewshire	20.09
East Lothian	2.50	Scottish Borders	3.85
East Renfrewshire	4.17	Shetland Islands	0.00
Edinburgh, City of	10.93	South Ayrshire	12.24
Eilean Siar	0.00	South Lanarkshire	14.57
Falkirk	8.63	Stirling	6.36
Fife	11.26	West Dunbartonshire	26.27
Glasgow City	43.52	West Lothian	9.00

Table 5.4: The local share considers the percentage of a local authority's data-zones that are amongst the 15% least deprived in Scotland.

Local authorities	Local share	Local authorities	Local share
Aberdeen City	35.58	Highland	4.79
Aberdeenshire	21.93	Inverclyde	2.73
Angus	9.86	Midlothian	14.29
Argyll & Bute	4.10	Moray	8.62
Clackmannanshire	9.38	North Ayrshire	2.79
Dumfries & Galloway	4.15	North Lanarkshire	6.22
Dundee City	11.17	Orkney Islands	0.00
East Ayrshire	6.49	Perth & Kinross	12.57
East Dunbartonshire	47.24	Renfrewshire	17.76
East Lothian	18.33	Scottish Borders	4.62
East Renfrewshire	57.50	Shetland Islands	0.00
Edinburgh, City of	39.53	South Ayrshire	16.33
Eilean Siar	0.00	South Lanarkshire	12.31
Falkirk	13.20	Stirling	18.18
Fife	12.36	West Dunbartonshire	2.54
Glasgow City	4.76	West Lothian	16.11

Table 5.5: Scottish Government joint urban/rural classification.

Class	Class name	Description
1	Large urban areas	settlements of over 125000 people.
2	Other urban areas	settlements of 10000 to 125000 people.
3	Accessible small towns	settlements of between 3000 and 10000 people, and within a 30 minute drive time to a settlement of 10000 or more.
4	Remote small towns	settlements of between 3000 and 10000 people, and with a drive time between 30 and 60 minutes to a settlement of 10000 or more.
5	Very remote small towns	settlements of between 3000 and 10000 people, and with a drive time of over 60 minutes to a settlement of 10000 or more.
6	Accessible rural areas	Areas with a population of less than 3000 people, and within a 30 minute drive time to a settlement of 10000 or more.
7	Remote rural areas	Areas with a population of less than 3000 people, and with a drive time between 30 and 60 minutes to a settlement of 10000 or more.
8	Very remote rural areas	Areas with a population of less than 3000 people, and with a drive time of over 60 minutes to a settlement of 10000 or more.

In order to be able to apply the COM-Poisson regression model in this context, we need to also take into account the population and age structure of each intermediate geography. This is achieved by reflecting expected counts (E_i) of hospital emergency admissions for each intermediate geography. The expected counts are computed using the age structure of each intermediate geography's population, together with estimates of the probabilities of hospitalisation in each age group. To account for the spatial autocorrelation of the data we use a conditional autoregressive model, which is specified as follows

$$\begin{aligned} P(Y_i = y_i|\mu_i,\nu_i) &= \left(\frac{\mu_i^{y_i}}{y_i!}\right)^{\nu_i} \frac{1}{Z(\mu_i,\nu_i)}, \\ Z(\mu_i,\nu_i) &= \sum_{j=0}^{\infty}\left(\frac{\mu_i^{j}}{j!}\right)^{\nu_i}, \\ \log\left\{\frac{\mu_i}{E_i}\right\} = \boldsymbol{x}_i^\intercal\boldsymbol{\beta} + \phi_i \Rightarrow & \qquad \mathbb{E}[Y_i] \approx E_i \exp\{\boldsymbol{x}_i^\intercal\boldsymbol{\beta} + \phi_i\}, \\ \log\{\nu_i\} = -\boldsymbol{x}_i^\intercal\boldsymbol{c} \Rightarrow & \qquad \mathbb{V}[Y_i] \approx E_i \exp\{\boldsymbol{x}_i^\intercal\boldsymbol{\beta} + \phi_i + \boldsymbol{x}_i^\intercal\boldsymbol{c}\}. \end{aligned} \tag{5.5}$$

Y is the dependent random variable being modelled (emergency hospital admissions), E_i is the expected emergency hospital admissions for the i^{th} intermediate geography, ϕ_i are the random effects for the parameter μ, while $\boldsymbol{\beta}$ and $\boldsymbol{c}$ are the regression coefficients for the centering link function and the shape link function. Finally, the covariates $\boldsymbol{x}_i$ are comprised of: the deprivation weight of the intermediate geography i, the percentages of each urban/rural class within the intermediate geography i (using large urban areas as the baseline model), and 32 dummy variables that relate the intermediate geography i to its local authority.

The conditional autoregressive prior (CAR) being used for the random effects ϕ_i in this model is given by

$$\phi_k|\phi_{-k} \sim \mathrm{N}\left(\frac{\rho\sum_{i=1}^{n} w_{ki}\phi_i}{\rho\sum_{i=1}^{n} w_{ki}+1-\rho}, \frac{\tau^2}{\rho\sum_{i=1}^{n} w_{ki}+1-\rho}\right) \tag{5.6}$$

and was proposed by Leroux et al. (2000) for modelling varying strengths of spatial autocorrelation. It can be seen as a generalisation of Besag et al. (1991) CAR prior where the first model can only represent strong spatial autocorrelation and produces smooth random effects. The random effects for non-neighbouring areas are conditionally independent given the values of the random effects of all the other areas. The parameter ρ can be seen as a spatial autocorrelation parameter, with $\rho = 0$ corresponding to independence, while $\rho = 1$ corresponds to strong spatial autocorrelation. In the first case there is an absence of spatial correlation in the data and the overdispersion is not caused by a spatial heterogeneity, while in the second case all the overdispersion is due to the spatial autocorrelation. When $0 < \rho < 1$, the random effects are correlated and the data present a combination of spatial structured and unstructured components. Lee (2011) compares four of the most common conditional autoregressive models and concludes that the model by Leroux et al. (2000) is the most appealing from both theoretical and practical standpoints.

In this formulation the coefficients have a direct link to either the mean or the variance, providing insight into the behaviour of the dependent variable. Larger values of $\boldsymbol{\beta}$ and $\boldsymbol{c}$ can be translated to higher mean and higher variance for the response variable, respectively. We implement a Bayesian approach for the previous model, and propose an efficient and exact MCMC algorithm based on the piecewise geometric bounds and the retrospective

sampling algorithm. We use diffuse multivariate normal priors for the regression coefficients with a mean of zero and a variance of 10^6. A uniform prior on the unit interval is specified for ρ, and a uniform prior on the interval $(0, 1000)$ is adopted for τ^2. In addition, the proposal distribution h is chosen to be a multivariate normal centered at the current value.

Table 5.6 shows the non-model-based regression coefficients for each local authority (32 local authorities in total). These coefficients refer to the intercepts of the 32 regression models (one for each local authority) each using an offset (e.g. $\log E_i$), but no other covariates. These 32 models are specified as follows

$$\begin{aligned} P(Y_i = y_i|\mu_i, \nu_i) &= \left(\frac{\mu_i^{y_i}}{y_i!}\right)^{\nu_i} \frac{1}{Z(\mu_i, \nu_i)}, \\ Z(\mu_i, \nu_i) &= \sum_{j=0}^{\infty} \left(\frac{\mu_i^{j}}{j!}\right)^{\nu_i}, \\ \log\left\{\frac{\mu_i}{E_i}\right\} = \beta_0 \Rightarrow & \qquad \mathbb{E}[Y_i] \approx E_i \exp\{\beta_0\}, \\ \log\{\nu_i\} = -c_0 \Rightarrow & \qquad \mathbb{V}[Y_i] \approx E_i \exp\{\beta_0 + c_0\}. \end{aligned} \tag{5.7}$$

Each of the 32 models includes only the intermediate geographies which belong to the same local authority. It must be noted that some of the local authorities are comprised of a small number of data points, for example Orkney Islands, Shetland Islands, and Eilean Siar include less than 10 intermediate geographies.

Table 5.7 shows the regression coefficients for the model in (5.5). The COM-Poisson coefficients for ν of most covariates are positive which is a sign of overdispersion. Table 5.7 shows that there is a wide range of values for the coefficients $\boldsymbol{c}$. They can take negative values (Orkney Islands) and up to

greater than 2 (Dumfries & Galloway, Scottish Borders). The regression coefficients b_1, c_1 for the deprivation weights have positive posterior median estimates, 0.87 and 0.05 respectively, with $(0.84, 0.90)$ and $(-0.36, 0.52)$ as their 95% credible intervals. This translates to higher emergency hospital admissions for intermediate geographies with high deprivation. This is not true for the variance, since the credible interval includes negative values. The data have a strong spatial autocorrelation as can be seen, in Table 5.8, from the credible intervals of the autocorrelation parameter ρ.

Figures 5.15 and 5.16 show the medians (plotted as diamonds) and the 95% credible intervals (plotted as lines) for the regression coefficients for both models. The black lines refer to the non-model-based coefficients whereas the red lines refer to the regression model including all covariates. In the top panel it can be seen that adjusting for the covariates (deprivation, urban/rural classification, and local authorities) shifts the regression coefficients towards zero. As we mentioned earlier, modelling the variance is the main interest in this application since it helps us identify areas with health inequalities. Comparing the panels in Figures 5.15 and 5.16 reveals a different pattern for the mean effects and variance effects of the local authorities. Local authorities with large μ coefficients (corresponding to poor health) do not necessarily have large ν coefficients (corresponding to large health inequalities). This can be seen in local authorities such as North Ayrshire, North Lanarkshire and South Ayrshire.

The coefficients for the percentages of each class are shown in Figures 5.17 and 5.18 where large urban areas are considered to be the baseline model. The remaining classes are plotted with regards to their distance from an

Table 5.6: Posterior medians of the non-model-based regression coefficients for each local authority.

Local authorities	β_0	c_0	Local authorities	β_0	c_0
Aberdeen City	-0.03	3.59	Highland	-0.06	3.41
Aberdeenshire	-0.26	2.10	Inverclyde	0.18	3.62
Angus	-0.16	2.15	Midlothian	-0.13	2.31
Argyll & Bute	-0.07	3.05	Moray	-0.25	2.27
Clackmannanshire	-0.18	2.57	North Ayrshire	0.18	3.06
Dumfries & Galloway	-0.18	3.23	North Lanarkshire	0.18	2.93
Dundee City	0.04	3.14	Orkney Islands	-0.17	2.29
East Ayrshire	0.16	3.24	Perth & Kinross	-0.13	3.02
East Dunbartonshire	-0.15	3.14	Renfrewshire	0.06	3.59
East Lothian	-0.25	2.48	Scottish Borders	0.01	3.10
East Renfrewshire	-0.26	2.99	Shetland Islands	-0.18	3.23
Edinburgh, City of	-0.31	3.69	South Ayrshire	0.11	3.10
Eilean Siar	-0.02	2.26	South Lanarkshire	0.01	2.54
Falkirk	-0.15	2.26	Stirling	-0.21	3.43
Fife	-0.14	2.75	West Dunbartonshire	0.13	2.57
Glasgow City	0.20	3.59	West Lothian	0.09	3.08

Table 5.7: Posterior medians for the regression coefficients of the full model.

Covariates	β_i	c_i	Covariates	β_i	c_i
Deprivation weight	0.87	0.05	Eilean Siar	0.00	0.39
Other urban area	0.00	-0.31	Falkirk	-0.19	1.48
Accesible small town	-0.04	0.06	Fife	-0.14	1.40
Remote small town	-0.04	-0.12	Glasgow City	0.03	1.72
Very remote small town	0.15	0.91	Highland	0.00	1.87
Accesible rural area	-0.08	-0.01	Inverclyde	0.04	1.39
Remote rural area	-0.19	0.11	Midlothian	-0.11	1.39
Very remote rural area	-0.09	0.51	Moray	-0.12	1.44
Aberdeen City	0.05	1.26	North Ayrshire	0.07	1.19
Aberdeenshire	-0.05	1.49	North Lanarkshire	0.06	1.08
Angus	-0.11	1.93	Orkney Islands	-0.11	-0.33
Argyll & Bute	-0.02	1.62	Perth & Kinross	0.03	1.26
Clackmannanshire	-0.21	1.68	Renfrewshire	0.04	1.70
Dumfries & Galloway	-0.10	2.42	Scottish Borders	0.12	2.38
Dundee City	-0.07	1.58	Shetland Islands	-0.13	1.04
East Ayrshire	0.07	1.87	South Ayrshire	0.12	1.17
East Dunbartonshire	0.05	1.57	South Lanarkshire	-0.01	1.45
East Lothian	-0.16	0.69	Stirling	-0.08	1.75
East Renfrewshire	-0.03	1.29	West Dunbartonshire	-0.03	1.05
Edinburgh, City of	-0.20	1.92	West Lothian	0.05	1.68

Table 5.8: Posterior medians for the variance and spatial autocorrelation of the random effects.

	Median	2.5%	97.5%
τ^2	0.004	0.002	0.008
ρ	0.927	0.783	0.971

urban area. The black circle represents the large urban area class whereas the blue, brown and violet lines represent the urban area, small town and rural area classes respectively. It can be seen that very remote small towns have higher (on average) emergency hospital admissions (see Panel 5.17) and higher excess variance (see Panel 5.18) compared to large urban areas.

Finally, Figure 5.19 shows the standardised incidence ratio of the average emergency hospital admissions using the non-model-based coefficients (on the left) and the coefficients of the full model (on the right).

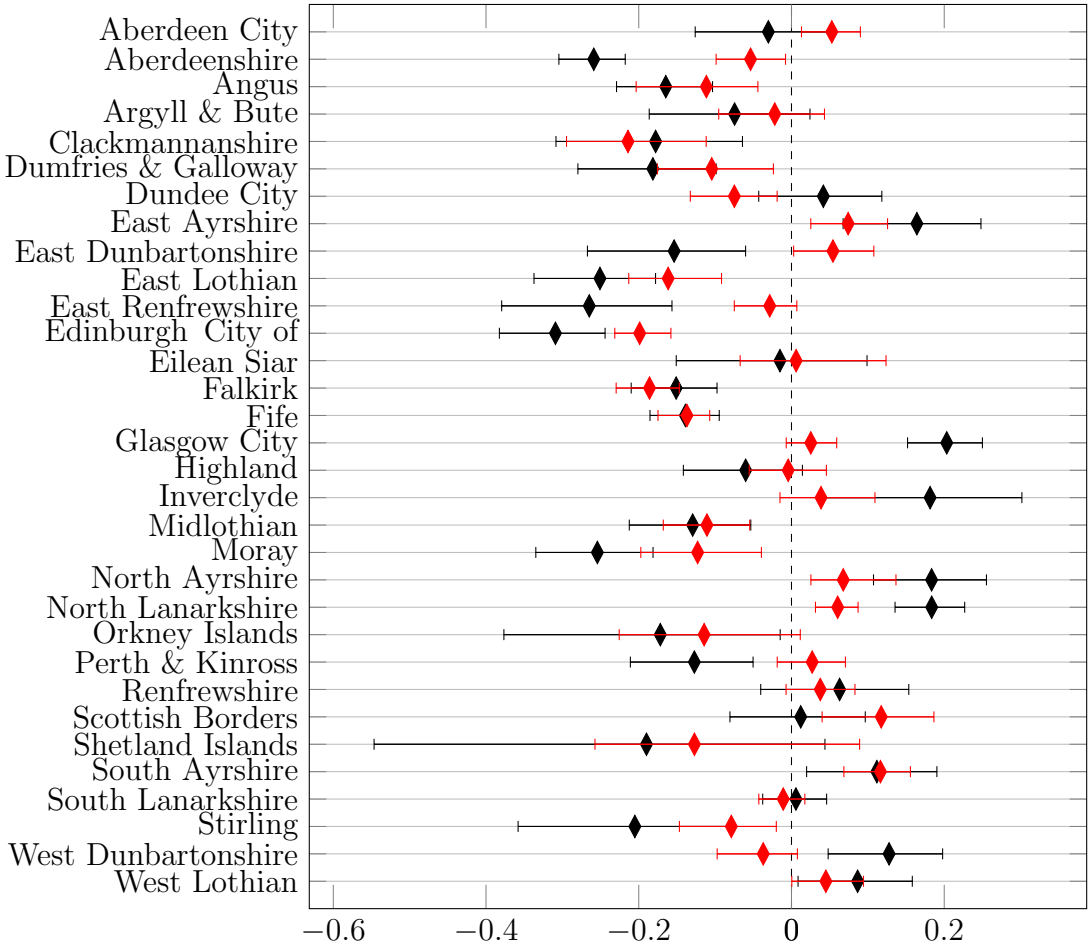


Figure 5.15: Credible intervals for the regression coefficients of μ for the Local authorities. The non-model-based regression coefficients of Table 5.6 are shown in black and the full model of Table 5.7 in red.

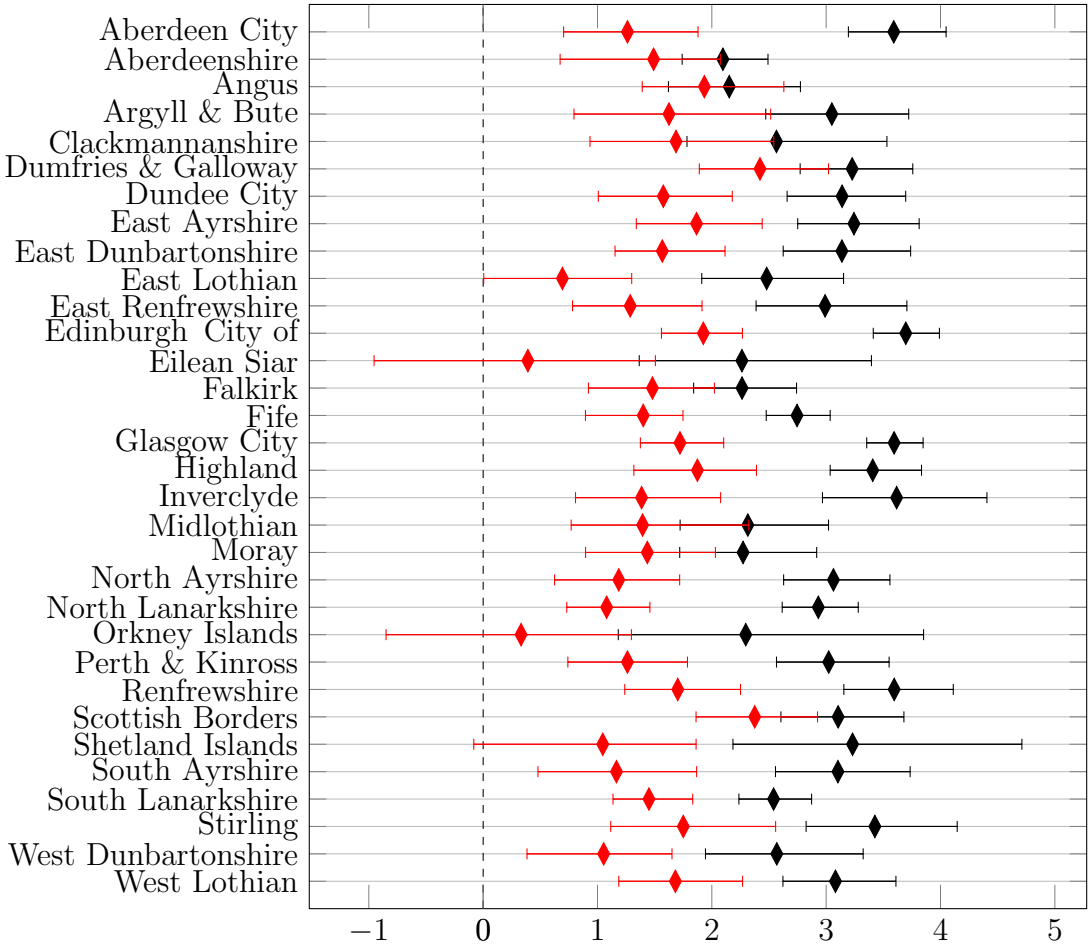


Figure 5.16: Credible intervals for the regression coefficients of ν for the local authorities. The non-model-based regression coefficients of Table 5.6 are shown in black and the full model of Table 5.7 in red.

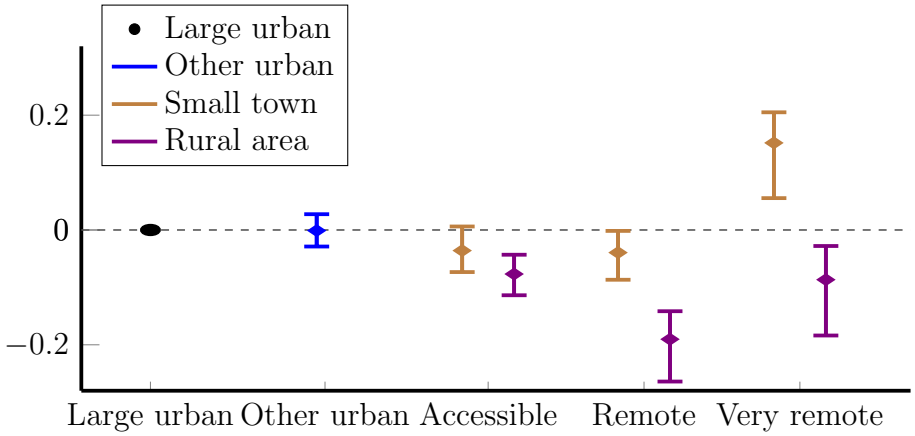


Figure 5.17: Credible intervals for the regression coefficients for μ for each class in Table 5.5.

Range of bounds for the acceptance probability

This section discusses the computational aspects of the retrospective algorithm used for inference in this example. The computational speed of the retrospective technique depends on which strategy is chosen to refine the bounds. We chose to increase the number of terms that are computed exactly for the estimation of the normalisation constant and use the piecewise geometric bounds for the remaining terms. We start by computing exactly 240 terms for every $Z(\mu_i, \nu_i)$ and every time the bounds need to be refined we compute 100 more terms for the observations that have a large difference between the upper and lower bound. Tables 5.9 and 5.10 show the percentages of every possible outcome (acceptance, rejection, or further refinement) for different number of refinements. Almost half of the time, for both parameters, there is no need to refine the bounds and we can make an accept/reject decision on just computing 240 terms. Tables 5.11 and 5.12 show the mean values for the difference of the log bounds. One can see that, even when a

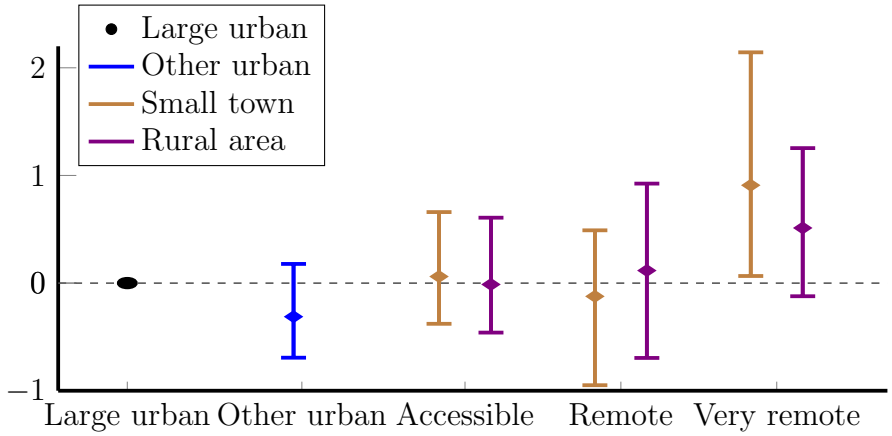


Figure 5.18: Credible intervals for the regression coefficients for ν for each class in Table 5.5.

large number of refinements is needed in order to make a decision, the difference of the log bounds is still large. This shows that there is no need to be very precise in our bounds. The weighted averages for the log differences and the computed terms are also shown. The weighted average for the log difference of the bounds when the MCMC rejects the candidate value for the first parameter is 2.51 which means that a rejection decision can often be reached with very loose upper and lower bounds.

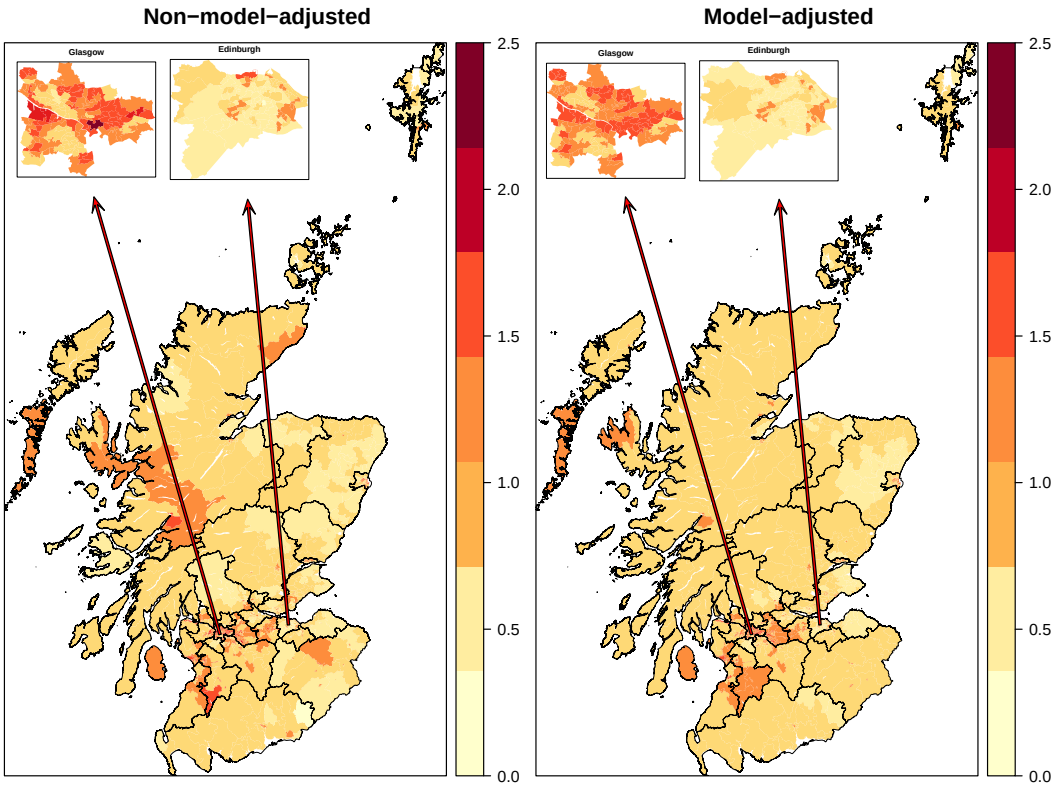


Figure 5.19: SIR for emergency hospital admissions.

Table 5.9: Percentages for refinements when updating the parameter μ.

Refinements	Accepted	Rejected	Still need refinement
0	9.6%	43.2%	47.2%
1	11.5%	20.1%	15.6%
≥ 2	7.3%	8.3%	0%
Total	28.4%	71.6%	

Table 5.10: Percentages for refinements when updating the parameter ν.

Refinements	Accepted	Rejected	Still need refinement
0	11.9%	27.2%	60.9%
1	16.3%	23.8%	20.8%
≥ 2	9.6%	11.2%	0%
Total	37.8%	62.2%	

Table 5.11: Mean values for the difference of the log bounds and the computed terms when updating the parameter μ.

Refinements	Accepted	Acceptance $\log\{\hat{a}_n\} - \log\{\check{a}_n\}$	Computed terms	Rejected	Rejection $\log\{\hat{a}_n\} - \log\{\check{a}_n\}$	Computed terms
0	33.9%	3.51	240	60.3%	3.55	240
1	40.5%	1.11	264.71	28.1%	1.13	264.66
≥ 2	25.6%	0.40	426.15	11.6%	0.43	420.55
Weighted average		1.74	297.84		2.51	267.75

Table 5.12: Mean values for the difference of the log bounds and the computed terms when updating the parameter ν.

Refinements	Accepted	Acceptance $\log\{\hat{a}_n\} - \log\{\check{a}_n\}$	Computed terms	Rejected	Rejection $\log\{\hat{a}_n\} - \log\{\check{a}_n\}$	Computed terms
0	31.5%	3.52	240	43.7%	3.55	240
1	43.1%	1.11	264.72	38.2%	1.13	264.70
≥ 2	25.4%	0.41	425.16	18.1%	0.43	420.06
Weighted average		1.69	297.82		2.07	281.52

5.2.2 Publications of Ph.D. students

Long (1990) examined the effect of education, marriage, family, and the mentor on gender differences in the number of published papers during the Ph.D. studies of 915 individuals. The population was defined as all male biochemists who received their Ph.D.'s during the periods 1956-1958 and 1961-1963 and all female biochemists who obtained their Ph.D.'s during the period 1950-1967. Some of the variables that were used in the paper are shown in Table 5.13. For ease of interpretation we standardise all covariates by subtracting their mean and dividing by their standard deviation.

Table 5.13: Description of variables.

Variable	Description
Gender of student	Equals 1 if the student is female; else 0.
Married at Ph.D.	Equals 1 if the student was married by the year of the Ph.D.; else 0.
Children under 6 years old	Number of children less than 6 years old at the year of the students Ph.D.
Ph.D. prestige	Prestige of the Ph.D. program in biochemistry based on studies. Unranked institutions were assigned a score of 0.75 while ranked institutions had scores ranging from 1 to 5.
Mentor	Number of articles produced by Ph.D. mentor during the last 3 years.

The study found, amongst other things, that females and Ph.D. students having children publish fewer (on average) papers during their Ph.D. studies. In addition, having a mentor with a large number of publications in the last three years has a positive effect on the number of publications of the Ph.D.

student.

We will focus on the students with at least one publication (640 individuals) with empirical mean and variance of 1.42 and 3.54 respectively, a sign of overdispersion. We compare the Poisson, negative binomial, and the COM-Poisson regression models. For the COM-Poisson regression model we will also use the shrinkage priors for the regression coefficients discussed in Chapter 4. In this example, the exchange algorithm will be used for posterior simulation. We prefer using the exchange algorithm due to its simplicity compared to the retrospective sampling algorithm since there is no need to estimate the normalisation constant at all or estimate bounds for the acceptance probability. As a result, te exchange algorithm is usually faster, in computational time, than the retrospective algorithm.

Figure 5.20 shows the 95% and 68% credible intervals for the regression coefficients of μ for all the regression models. The Poisson and negative binomial models have similar results. The only difference between them is that the effect of having children is not statistically significant using the latter model. The gender of a Ph.D. student and the number of articles by the Ph.D. mentor are the only covariates that are statistically significant for both the Poisson and negative binomial models. Specifically, these models conclude that female Ph.D. students publish less on average than male Ph.D. students and that a mentor who has published a lot of articles has a positive effect on the number of articles of the Ph.D. student. On the other hand for the COM-Poisson models, the previous two covariates seem to not have a statistically significant effect on the mean of the number of articles published by a Ph.D. student.

Regarding the *gender* covariate it must be noted that there are four male Ph.D. students with a large number of articles published $(11, 11, 15, 18)$ that could be considered as outliers. If these four students are taken out of the dataset, the gender covariate does not have a significant effect for the Poisson and negative binomial models. In addition, the empirical means of the male and female Ph.D. students are 1.5 and 1.2 respectively while the empirical median is 1 for both genders. Thus the COM-Poisson regression model seems to be doing a better job at not concluding that there is an effect of the gender covariate.

Figure 5.21 shows the 95% and 68% credible intervals for the regression coefficients of ν for the COM-Poisson regression models. This figure shows that there seems to be a positive effect of the *mentor* covariate on the variance of the articles of the Ph.D. student. The more articles a mentor publishes (during the last 3 years) the larger the variance for the number of articles published by a Ph.D. student. This seems to be reinforced further when we look at the empirical variances of students having mentors with an above average number of articles published versus students having mentors with less than average number of articles published. The empirical variances for the former group is 5.8, with the latter group having a variance of 2.1 respectively (ratio of around 2.8). The corresponding empirical means are 1.9 and 1.2 (ratio of around 1.6). In Poisson-distributed data one would expect the ratios to be roughly equal.

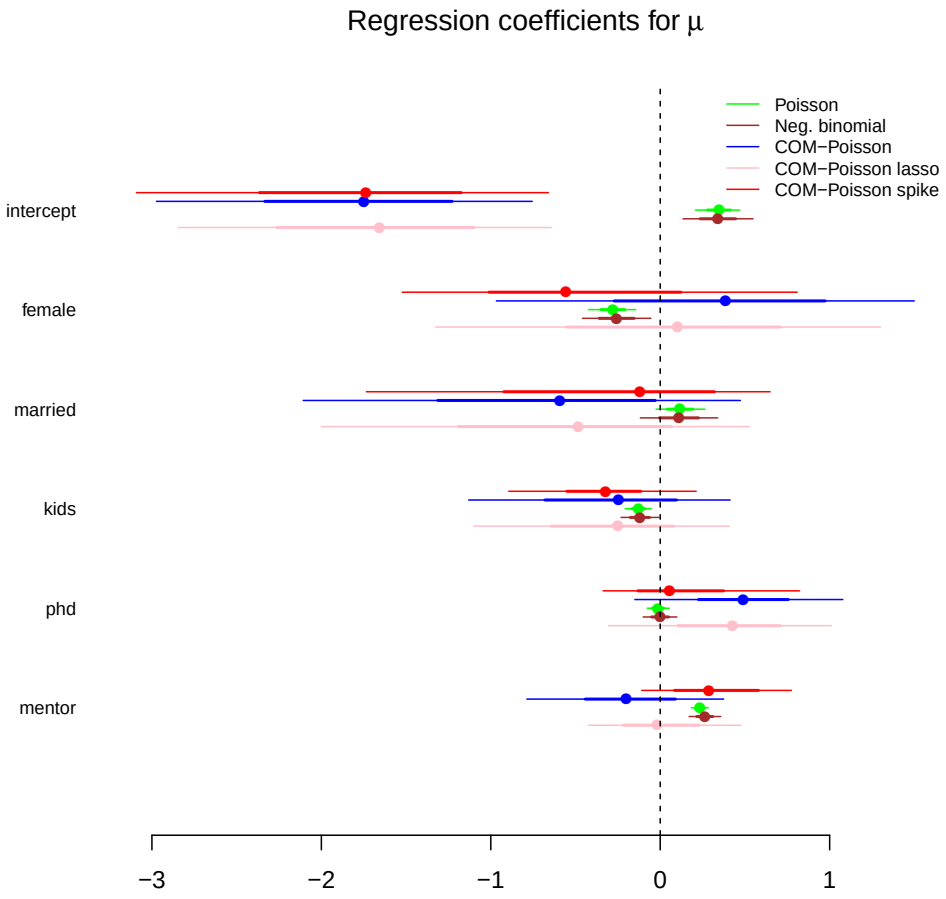


Figure 5.20: Publication data: 95% and 68% credible intervals for the regression coefficients of μ.

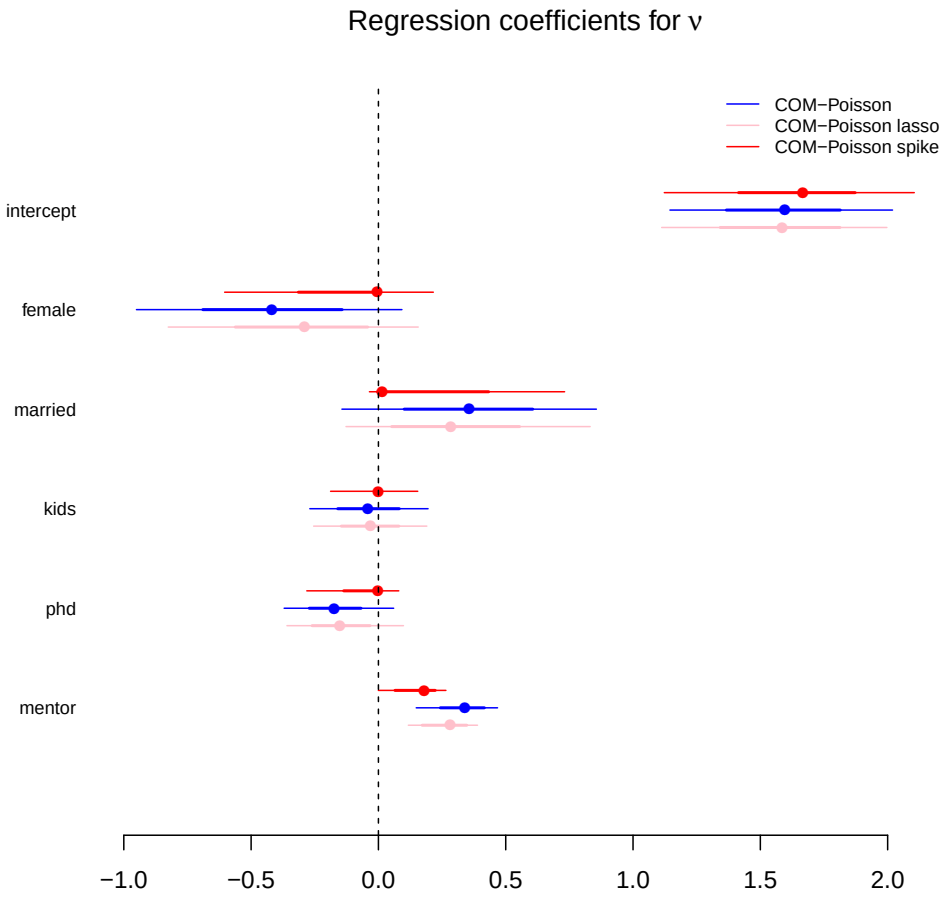


Figure 5.21: Publication data: 95% and 68% credible intervals for the regression coefficients of ν.

5.2.3 Fertility data

We finally use a data set from Winkelmann (1995) based on data from the second (1985) wave of the German Socio-Economic Panel. The data consist of 1243 women over 44 in 1985, who are in first marriages and who answered the questions relevant to the analysis. The variables that were used in the paper can be seen in Table 5.14. For ease of interpretation we standardise all covariates by subtracting their mean and dividing by their standard deviation.

Table 5.14: Description of variables.

Variable	Description
Nationality	Equals 1 if the woman is German; else 0.
General education	Measured as years of schooling.
Post-secondary education (vocational training)	Equals 1 if the woman had vocational training; else 0.
Post-secondary education (university)	Equals 1 if the woman had a university degree; else 0.
Religion	The woman's religious denomination (Catholic, Protestant, Muslim) with other or none as the baseline group.
Year of birth	Year that the woman was born.
Age at marriage	Year that the woman was married.

The empirical mean and variance of the response are 2.39 and 2.33 respectively. The unconditional variance is already slightly smaller than the unconditional mean. Including covariates the conditional variance will reduce further, thus suggesting that the data show underdispersion. For this reason, the negative binomial model was not used in this context.

Parametric regression

The results for the different parametric models for count data are shown in Figures 5.22 and 5.23. The credible intervals for the coefficients of μ are similar across all the models. Looking at Figure 5.23 we can see that vocational education, age, and age at marriage are statistically significant.

Even though the estimated regression coefficients seem similar one might want to ask the question which model describes the data the best.

Model selection

Spiegelhalter et al. (2002) defined the deviance information criterion (DIC) as a (more) Bayesian alternative to model assessment tool like AIC and BIC. The DIC can be applied to non-nested models and its calculation does not require maximisation over the parameter space, like the AIC and BIC. It is comprised of two terms, one representing goodness-of-fit and another for model complexity. It is defined as

$$\begin{aligned} \text{DIC} &= p_D + \overline{D}, \\ &= 2\overline{D} - D(\overline{\boldsymbol{\theta}}), \\ &= 2p_D + D(\overline{\boldsymbol{\theta}}), \end{aligned} \tag{5.8}$$

where $\overline{D} = \mathbb{E}[-2\log f(y|\boldsymbol{\theta})]$ is the expectation of the deviance, $p_D = \overline{D} - D(\overline{\boldsymbol{\theta}})$ is known as the effective number of parameters and $\overline{\boldsymbol{\theta}}$ is the posterior estimate of the parameters (mean, median, etc). Computing the second term of the DIC is done by estimating the deviance at each iteration of the MCMC chain and then finding the average. A smaller DIC indicates a better fit to

the data set. For more information one can see Spiegelhalter et al. (1998, 2002).

The results can be seen in Table 5.15. The COM-Poisson model always outperforms the Poisson and the negative binomial models. Using the models with informative priors for the regression coefficients of ν gives similar results as the model with diffuse priors.

Figures 5.24-5.29 show the sample and predicted relative frequencies (for the number of children) for the six biggest groups in the data set. Information about the groups and the number of women within each group are presented in the figures. The figures reinforce the notion that the COM-Poisson regression model provides a better fit to the data, compared to the Poisson alternative.

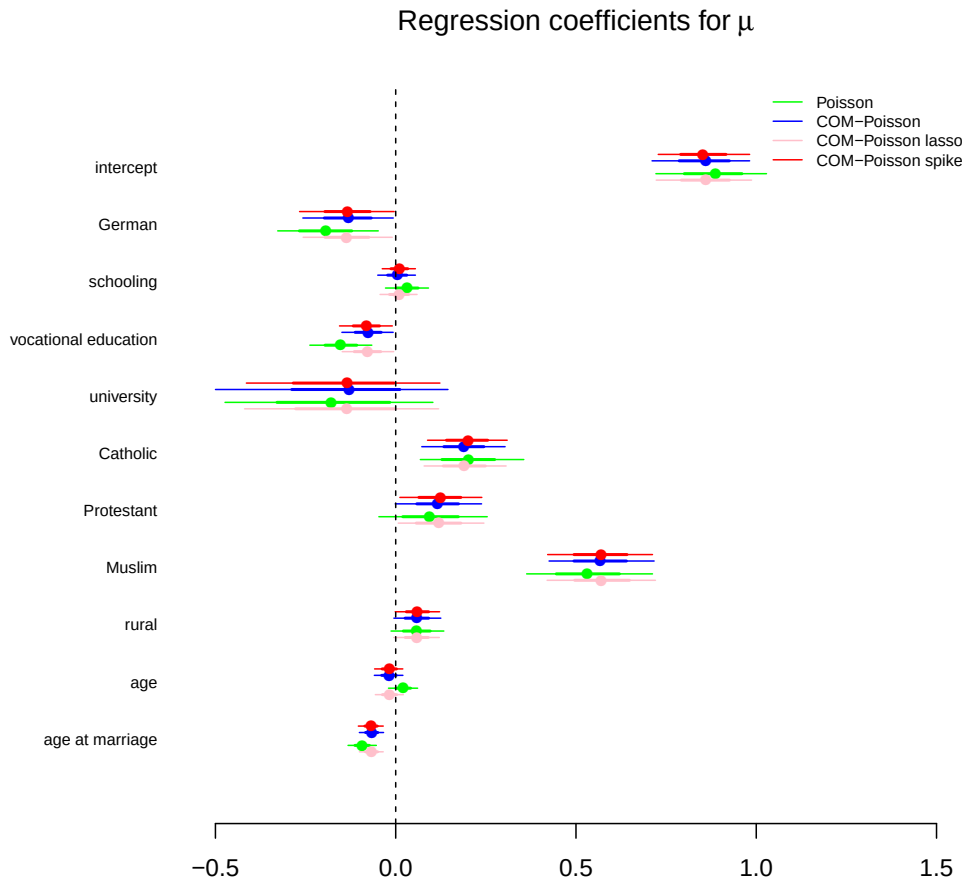


Figure 5.22: Fertility data: 95% and 68% credible intervals for the regression coefficients of μ.

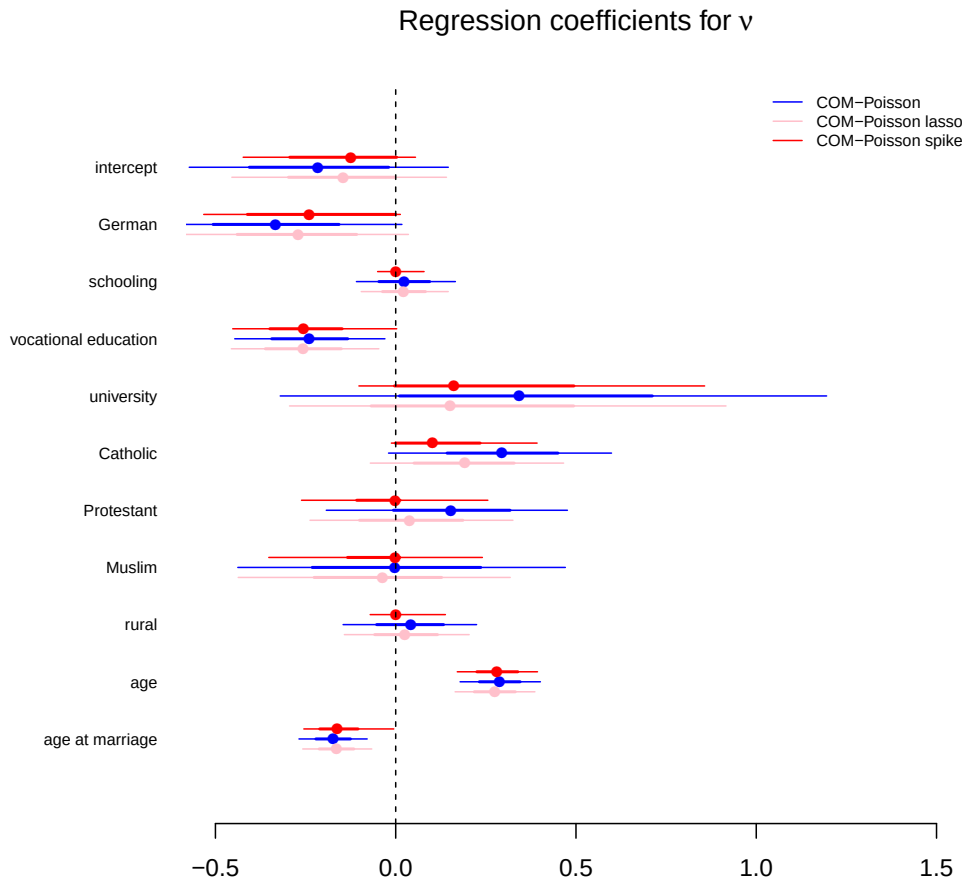


Figure 5.23: Fertility data: 95% and 68% credible intervals for the regression coefficients of ν.

Table 5.15: Deviance information criterion for all models (minimum DIC is in bold).

	Poisson	Negative binomial	COM-Poisson	COM-Poisson (lasso)	COM-Poisson (spike and slab)
Ph.D. data	2251.09	2108.05	**2056.77**	2058.05	2062.23
Fertility data	4214.55	-	4121.92	**4121.43**	4121.74

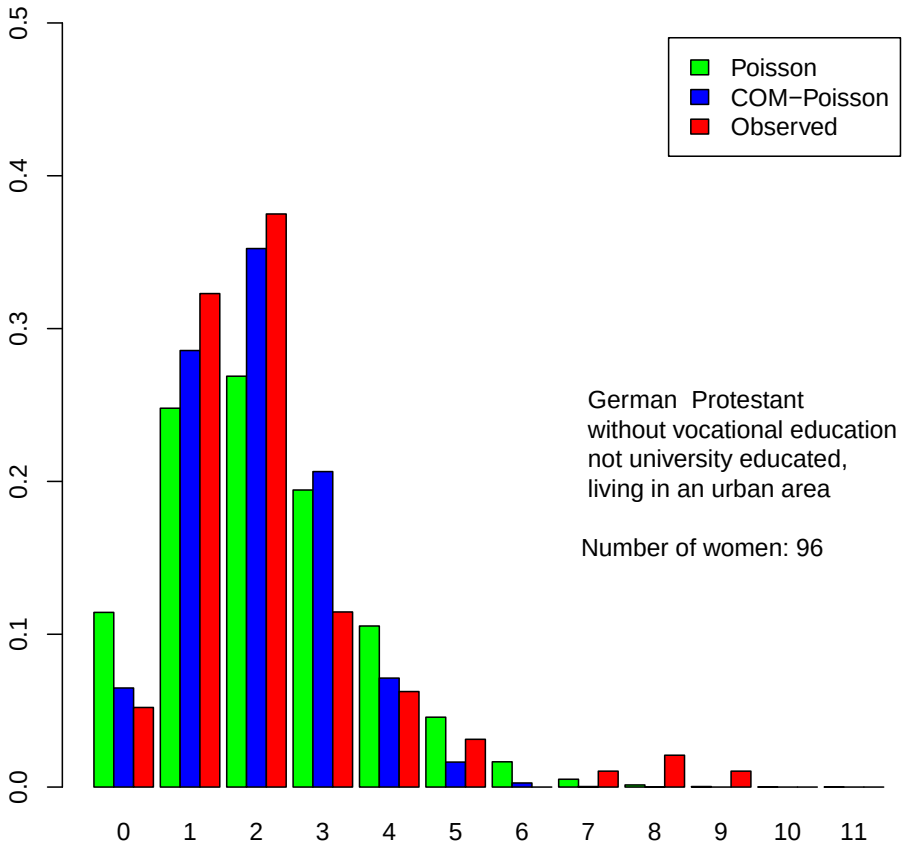


Figure 5.24: Sample and predicted relative frequencies for the number of children.

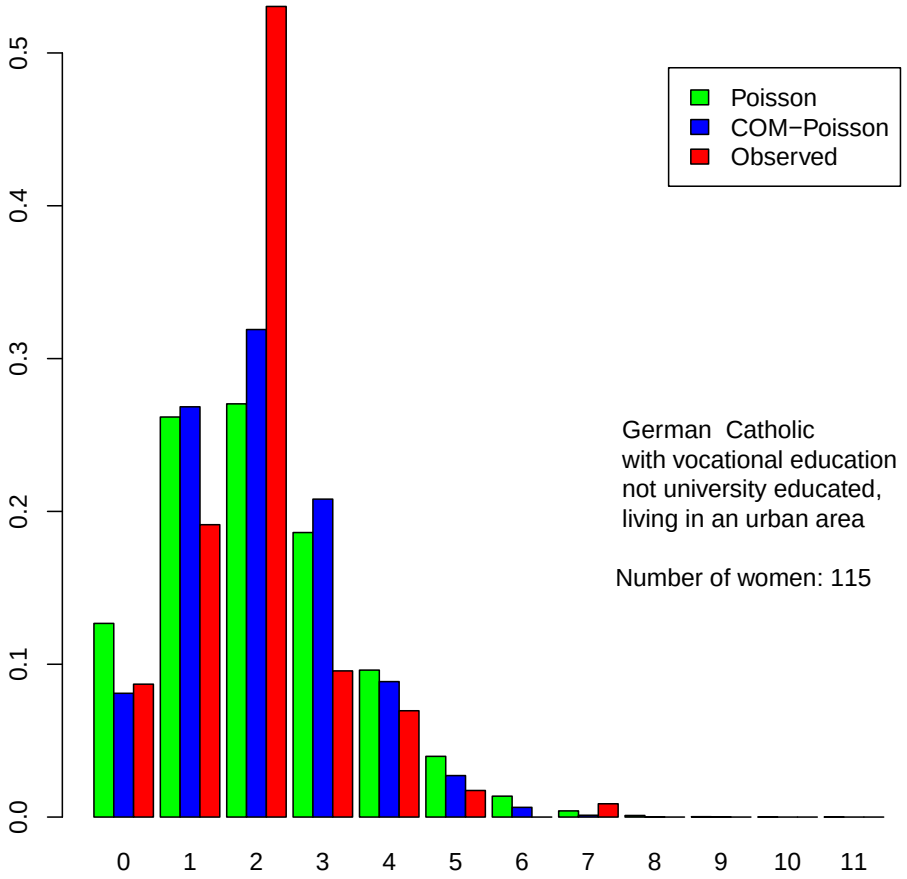


Figure 5.25: Sample and predicted relative frequencies for the number of children.

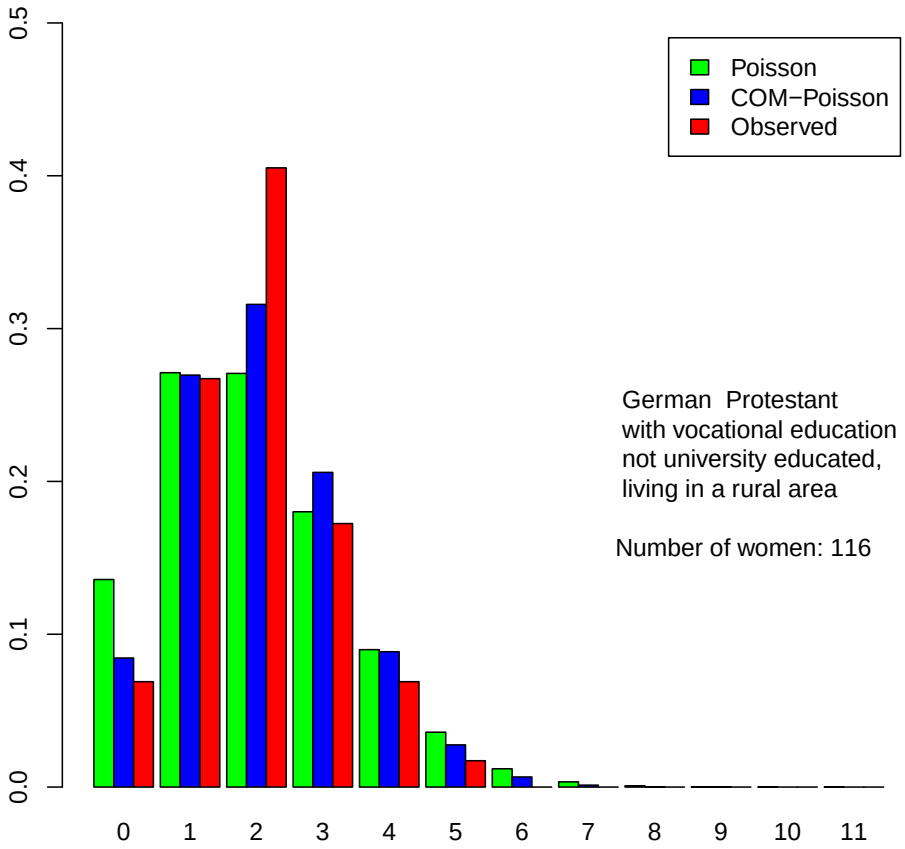


Figure 5.26: Sample and predicted relative frequencies for the number of children.

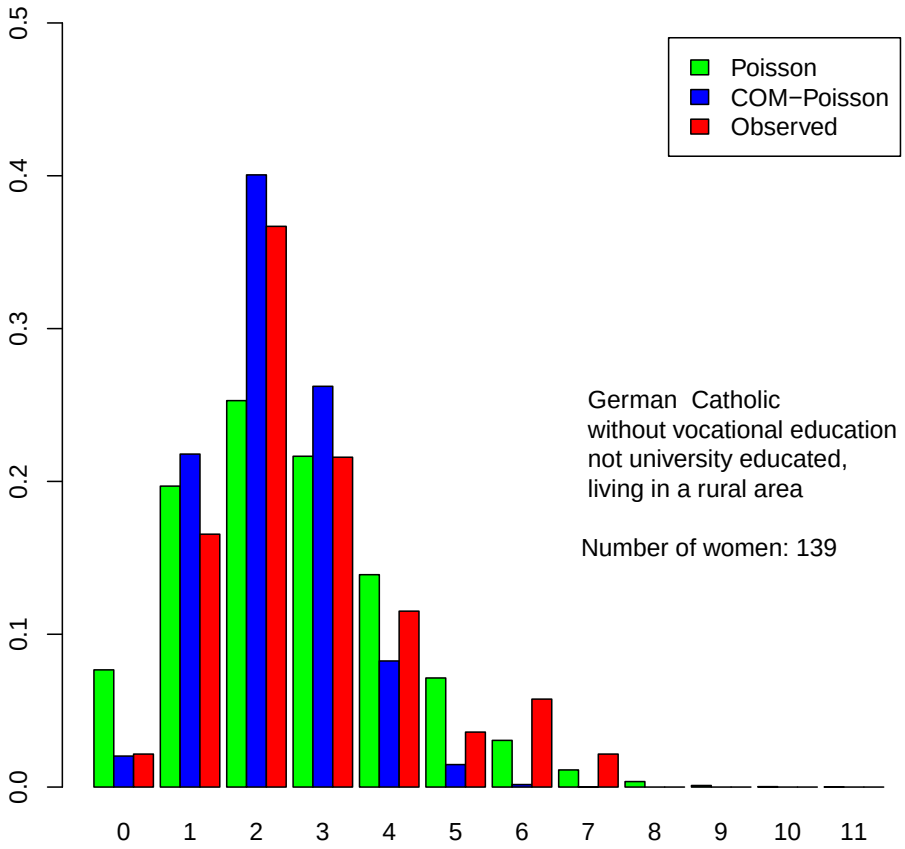


Figure 5.27: Sample and predicted relative frequencies for the number of children.

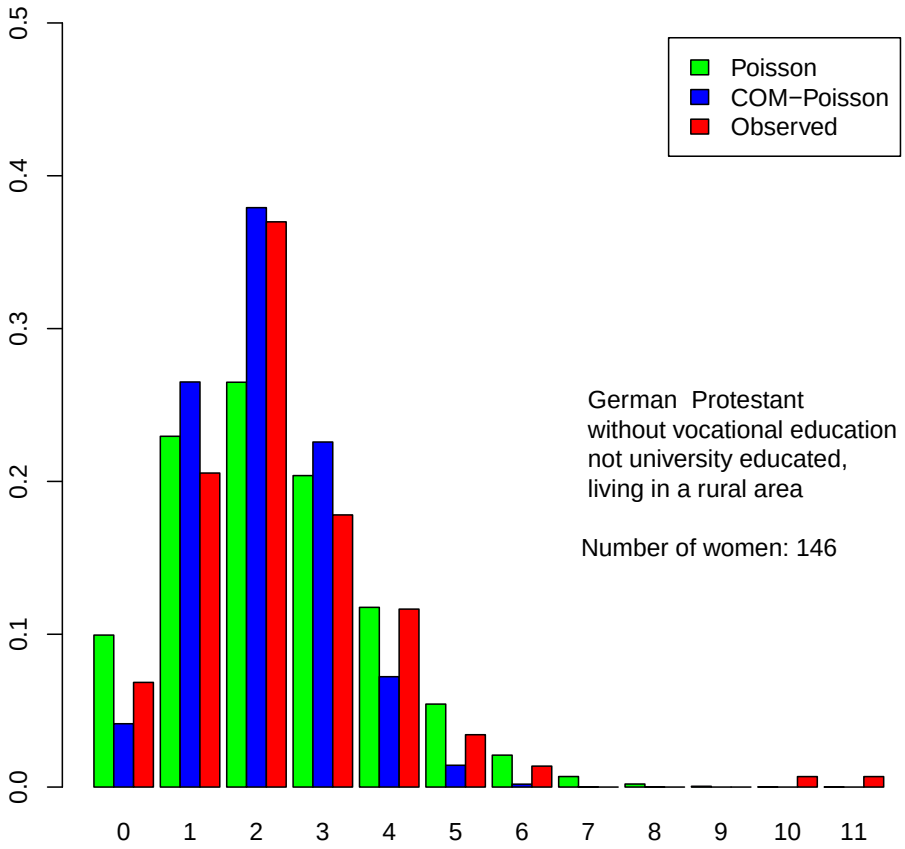


Figure 5.28: Sample and predicted relative frequencies for the number of children.

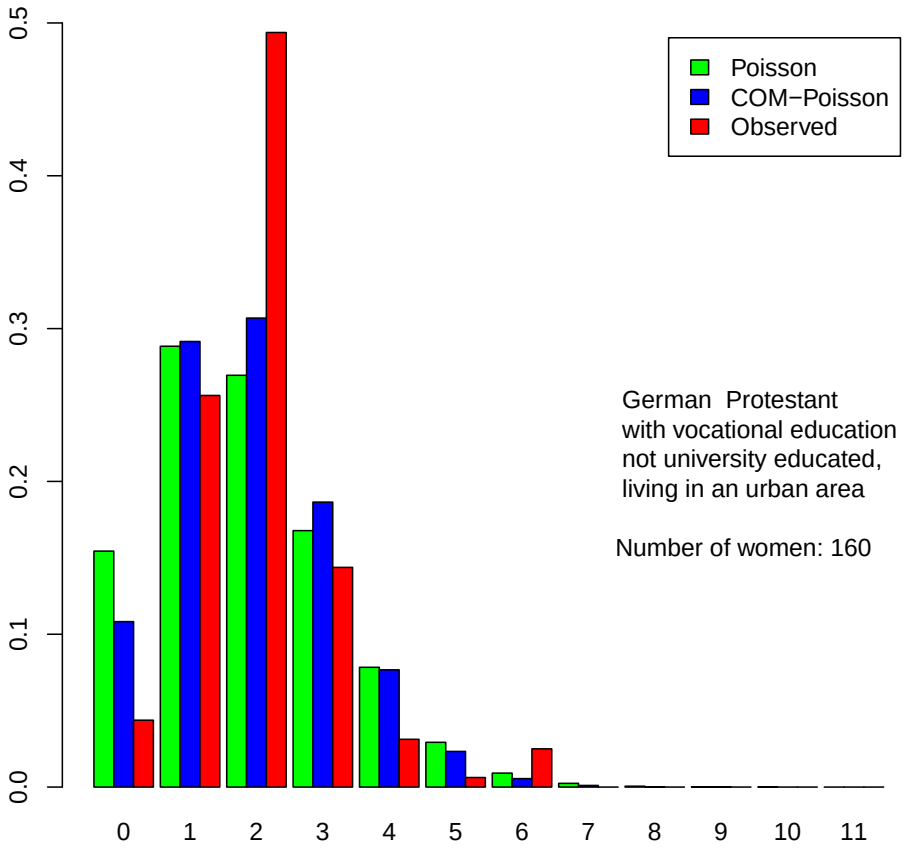


Figure 5.29: Sample and predicted relative frequencies for the number of children.

Bayesian density regression

Despite the COM-Poisson model providing a better fit to the data than its parametric competitors, it does not seem to fully capture the conditional distribution of the response. The COM-Poisson regression model is able to identify the mode correctly across all six groups but cannot always estimate the density accurately (cf. Figures 5.25, 5.29). The mode of the distribution of the sample relative frequencies for both these groups is at two and all other values, even the ones close to two, have small relative frequencies. This is difficult to capture even for the COM-Poisson model which has higher predicted relative frequencies to values close to two. This can be seen in Figure 5.25 looking at the, sample and predictive, relative frequencies at values one and three.

As a further improvement we will now employ the Bayesian density regression model presented in Section 4.2. The Bayesian density regression model has the COM-Poisson regression model as its "baseline" model, but also has the ability to use an adaptive mixture of these models. Figures 5.30, 5.31 show the sample and predicted relative frequencies of the previous models along with the predictive relative frequency of the Bayesian density regression. Both figures show that the proposed Bayesian density regression model provides a better fit to the data. This is due to the new model assigning most of the observations to an underdispersed COM-Poisson regression model with mode at two, and the rest of them to an overdispersed COM-Poisson regression model with mode at two. When we use the "simple" COM-Poisson regression model the second COM-Poisson regression model is "merged" with

the first one and as a result we lose valuable information.

Model selection

Watanabe (2010) introduced the widely applicable information ctiterion, also known as WAIC or the Watanabe-Akaike criterion, that also works with singular models and thus is particularly helpful for models with hierarchical and mixture structures in which the number of parameters increases with sample size and where point estimates do not make sense. This criterion can also be considered as a generalisation of the DIC seen in 5.2.3. Like the DIC, WAIC estimates the effective number of parameters to adjus for overfitting. Instead of p_D (in the case of DIC), two adjustments have been proposed. These are

$$\begin{aligned} p_{\text{WAIC1}} &= 2\sum_{i=1}^{n}\left(\log \mathbb{E}[f(y_i|\boldsymbol{\theta})] - \mathbb{E}[\log f(y_i|\boldsymbol{\theta})]\right), \\ p_{\text{WAIC2}} &= \sum_{i=1}^{n} \mathbb{V}[\log f(y_i|\boldsymbol{\theta})], \end{aligned} \tag{5.9}$$

Computing these terms can be done by replacing the expectations with averages over the number of posterior draws, similar to computing p_D. We will focus on p_{WAIC1} which is also similar to p_D. Gelman et al. (2013) scale the WAIC of by a factor of 2 so that it is comparable to DIC since in the original definition WAIC is the negative of the average log pointwise redictive density and thus is divided by n). As a result, the WAIC can be defined as

$$\text{WAIC} = -2\left(\text{lppd} - p_{\text{WAIC}}\right) \tag{5.10}$$

where $\text{lppd} = \sum_{i=1}^{n} \log \int f(y_i|\boldsymbol{\theta}) f_{\text{post}}(\boldsymbol{\theta})$ is the log pointwise predictive density (where $f_{\text{post}}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|y)$). This equation shows another advantage of

the WAIC, instead of conditioning on a point estimate (mle, posterior mean, posterior median) it averages over the posterior distribution. This can be very important especially in a predictive context. For more information on the WAIC see Vehtari and Gelman (2014); Gelman et al. (2014). In the programming language R, the package *LaplacesDemon* (Hall, 2012) includes a function which takes as an argument a $n \times s$ matrix of log-likelihoods (n data points and s samples) and calculates the WAIC. Table 5.16 shows the DIC and WAIC for the fertility data.

Table 5.16: Minimum DIC for the "simple" COM-Poisson regression model (with lasso priors) and WAIC for the Bayesian density regression model for the fertility data (minimum criterion is in bold).

	DIC	WAIC
Fertility data	4121.43	**4116.54**

Table 5.17 shows the Kullback-Leibler divergence of the predicted distributions from the observed distribution for each of the six groups, further supporting the claim that the Bayesian density regression model provides the best fit to the data.

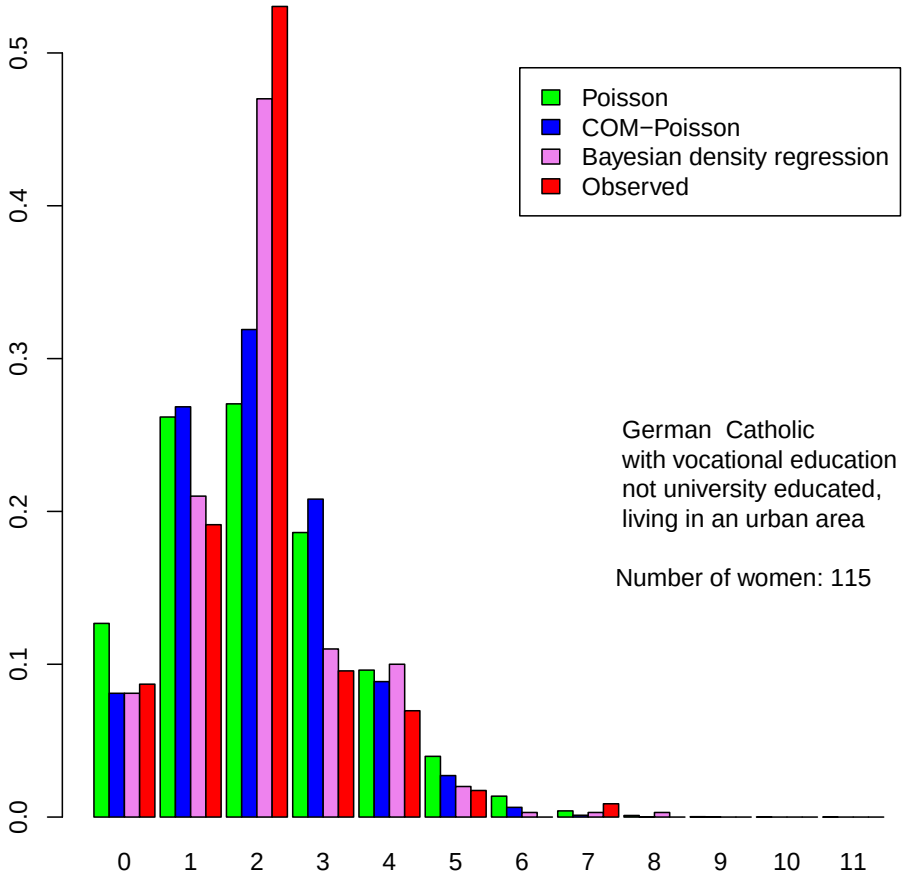


Figure 5.30: Sample and predicted relative frequencies for the number of children.

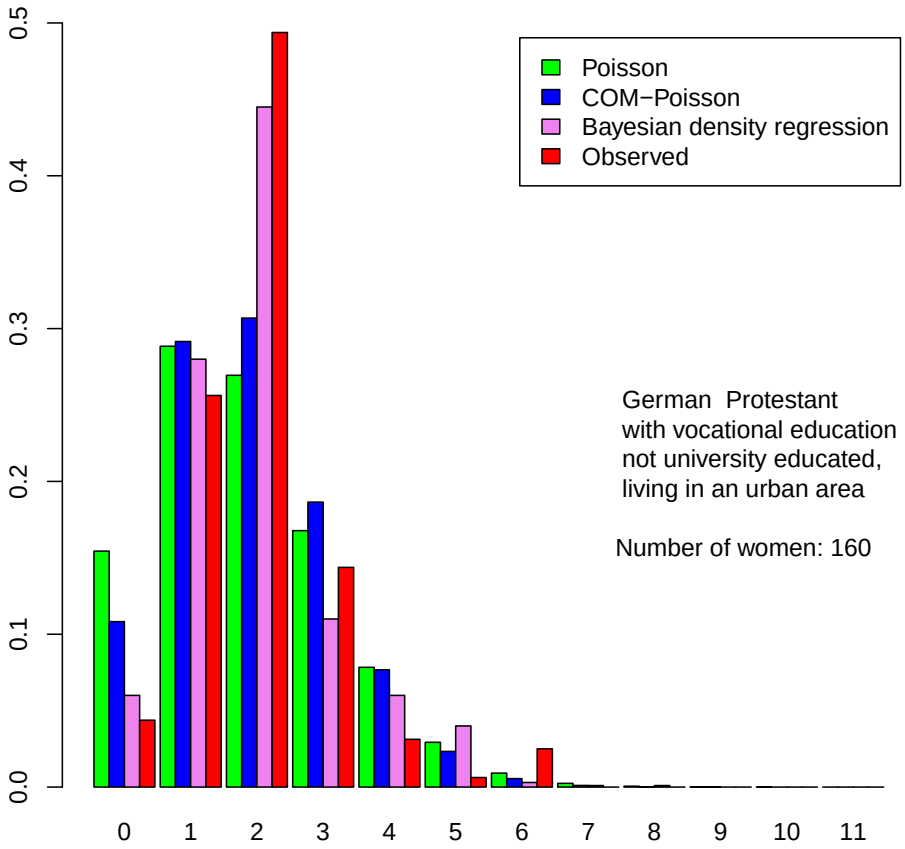


Figure 5.31: Sample and predicted relative frequencies for the number of children.

Table 5.17: Kullback-Leibler divergence between the predicted distributions and the observed distribution.

Group	Poisson regression	COM-Poisson regression	Bayesian density regression
1	0.264	0.253	**0.223**
2	0.171	0.129	**0.061**
3	0.074	0.028	**0.026**
4	0.184	0.167	**0.128**
5	0.194	0.234	**0.166**
6	0.178	0.126	**0.056**

Chapter 6

Conclusions and future work

Conclusions

In this thesis we proposed and presented

1. a parametric Bayesian regression models for count data which is more flexible than frequently used methods such as Poisson regression or negative binomial regression;
2. two different simulation techniques for intractable likelihoods; and
3. a much more flexible Bayesian density regression model for count data,

The first regression model is the "simple" COM-Poisson Bayesian regression model. We showed through simulations and case studies that its ability to differentiate between a covariate's effect on the mean of the response and the one on its variance can give a more complete picture of the effect a covariate

has on the conditional distribution of the response variable. The regression models for count data that are mainly used in the literature (Poisson and negative binomial) can give a false picture of the effect of a covariate on the response variable. The COM-Poisson, on the other hand, is able to detect the true effects of a covariate on the response variable (cf. Subsection 5.1.1). We have used informative (shrinkage) priors for the regression coefficients of this model and argued that the usage of these priors corresponds to using the Poisson regression model as the "baseline" model. This means that using this model we can, at a small additional computational cost, represent underdispersion or overdispersion, if necessary, but otherwise fall back to classical Poisson regression. Three different data sets have been analysed and we have shown that the COM-Poisson regression model provides a better fit to the data compared to the alternative models (Poisson, negative binomial) in terms of DIC.

We have presented an extensive case study for COM-Poisson regression in which we modelled emergency hospital admissions data in Scotland. By using a COM-Poisson distribution we were able to model the mean and the variance explicitly. As a result, we were able to identify areas with a high level of health inequalities (Chanialidis et al., 2014).

One challenge of fitting models involving the COM-Poisson distribution is that its normalisation constant is not known in closed form. We have proposed two simulation techniques which avoid having to compute this normalisation constant exactly. The first approach, retrospective sampling, was based on sequentially computing lower and upper bounds on the normalisation constant for distributions for which the ratios of consecutive probabilities

can be bounded over ranges of the random variable. Using these bounds we are often able to make a decision of the acceptance or rejection of a proposed value before the calculation of the normalisation constant gets too expensive. Indeed, the results have shown that in order for the MCMC algorithm to make a decision between accepting or rejecting a candidate move, the bounds on the acceptance probability do not need to be tight.

The second simulation technique was based on the exchange algorithm (Murray et al., 2006). Its key idea is that the inclusion of an auxiliary sample in the acceptance ratio allows for cancelling out the normalisation constants, which are expensive to compute. This however requires being able to sample from the COM-Poisson distribution efficiently. In Subsection 3.3.2 we have showed how one can use rejection sampling to draw efficiently from the COM-Poisson distribution.

The second regression model is based on an adaptive Dirichlet process mixture of COM-Poisson regression models. This model can be thought of as an extension to the "simple" COM-Poisson regression model in the same way that the COM-Poisson regression model can be thought of as an extension to the Poisson regression model. It provides more flexibility compared to the "simple" COM-Poisson model and is flexible enough to model underdispersed and overdispersed distributions (Section 2.4.5). We take advantage of the flexibility of this Bayesian density regression model and we use it to determine the conditional quantiles by estimating the density of the data, thus eliminating the problems of applying quantile regression to count data (Section 2.3).

Future work

We mentioned on page 104 that if one compares the acceptance ratios of the exchange algorithm, in equation (3.27), and the Metropolis Hastings, in equation (3.4), the only difference is that the ratio of the normalisation constants $\frac{Z(\boldsymbol{\theta})}{Z(\boldsymbol{\theta}^*)}$ is replaced by $\frac{q_{\boldsymbol{\theta}}(y^*)}{q_{\boldsymbol{\theta}^*}(y^*)}$. This can be seen as an importance sampling ratio of $\frac{Z(\boldsymbol{\theta})}{Z(\boldsymbol{\theta}^*)}$. An alternative approach that takes advantage of the exchange algorithm has been proposed by Alquier et al. (2014). Instead of simulating a single auxiliary vector $\boldsymbol{y}^*$ they use an unbiased estimator of $\frac{Z(\boldsymbol{\theta})}{Z(\boldsymbol{\theta}^*)}$ at each step of the exchange algorithm by simulating a number of auxiliary vectors $\boldsymbol{y}^{*1}, \boldsymbol{y}^{*2}, \ldots, \boldsymbol{y}^{*M}$ from $p(|\boldsymbol{\theta}^*)$ and then approximate the ratio of normalisation constants by

$$\frac{1}{M}\sum_{m=1}^{M}\frac{q_{\boldsymbol{\theta}}(\boldsymbol{y}^{*m})}{q_{\boldsymbol{\theta}^*}(\boldsymbol{y}^{*m})} \approx \frac{Z(\boldsymbol{\theta})}{Z(\boldsymbol{\theta}^*)}. \tag{6.1}$$

As a result an approximation, $\tilde{a}$ of the acceptance ratio, a, is computed. For $M = 1$ the new MCMC will be the same as the exchange algorithm, and when $M \to \infty$ the new MCMC will be equivalent to the Metropolis-Hastings algorithm. Alquier et al. (2014) demonstrate the method on a simple single parameter model and then apply their methodology to more challenging models (e.g. network data). Even though for $1 < M < \infty$ the new algorithm is not guaranteed to sample from the target distribution, Alquier et al. (2014) show that the new method, termed as "noisy" Monte Carlo, performs better compared to the exchange algorithm in terms of bias (on the posterior means). This is due to the improved mixing in the approximate "noisy" algorithm.

Another possible way to improve mixing for the exchange algorithm would

be to combine it with a modified delayed rejection scheme (Mira, 2001). Instead of proposing a parameter $\boldsymbol{\theta}^*$ and sampling a single observation y^* from a COM-Poisson with parameter $\boldsymbol{\theta}^*$; one could sample more auxiliary variables $\boldsymbol{y}^{*1}, \boldsymbol{y}^{*2}, \ldots, \boldsymbol{y}^{*M}$, use the first one along with $\boldsymbol{\theta}^*$ as the new candidate values of the Metropolis-Hastings algorithm (e.g.$(y^{*1}, \boldsymbol{\theta}^*)$); in case of rejection the new candidate pair will be $(\boldsymbol{y}^{*2}, \boldsymbol{\theta}^*)$ and so on. It will be interesting to implement both ideas ("noisy" algorithm and delayed rejection algorithm) along with the exchange algorithm for both the proposed models (COM-Poisson and Bayesian density regression) and draw comparisons between them.

Appendices

Appendix A

MCMC diagnostics

R (R Core Team, 2014) was used for all the computations in this paper. Trace plots, density plots, autocorrelation plots (for every regression coefficient) and results for the Gelman and Rubin diagnostic, (Gelman and Rubin, 1992), were employed to assess convergence of the MCMC sampler to the posterior distribution, using the coda package (Plummer et al., 2006). Regarding the COM-Poisson regression model, we present MCMC diagnostics for the emergency hospital admissions case study in Section 5.2.1. The MCMC was run for 60000 iterations with the first 20000 as the burn-in period. Three simulations were run, with different starting values, with a multivariate potential scale reduction factor $\hat{R} = 1.08$ (Gelman et al., 2004). In the next pages we present the trace plots and autocorrelation for the first eight regression coefficients. The first regression coefficient (e.g. β_1, c_1) corresponds to the *deprivation weight* while the rest correspond to the classes of Table 5.5 (using large urban areas as the baseline model). The trace plots of all the

other regression coefficients which are not in the appendix, can be found on my website.[1]

For the Bayesian density regression model we cannot apply similar diagnostics since the number of clusters change for every iteration, thus the number of regression parameters are not constant across the MCMC algorithm. One can plot the number of clusters across iterations and find the empirical probability of having K clusters (Dunson et al., 2007). We will focus on the second simulated example on page 134.

Similar diagnostics have been applied to all simulations and case studies.

[1] http://www.chanialidis.com/diagnosticsMCMC

A.1 COM-Poisson regression

Emergency hospital admissions

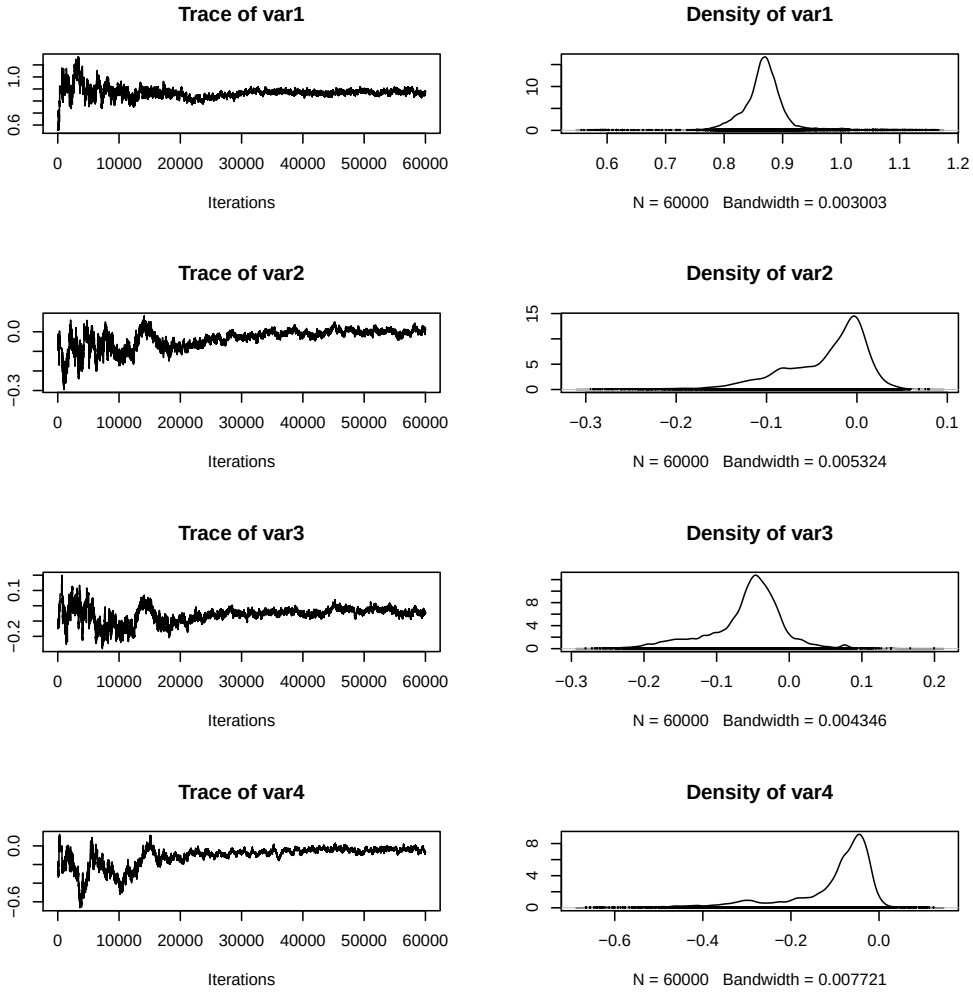


Figure A.1: Trace plots for $\beta_1, \beta_2, \beta_3, \beta_4$.

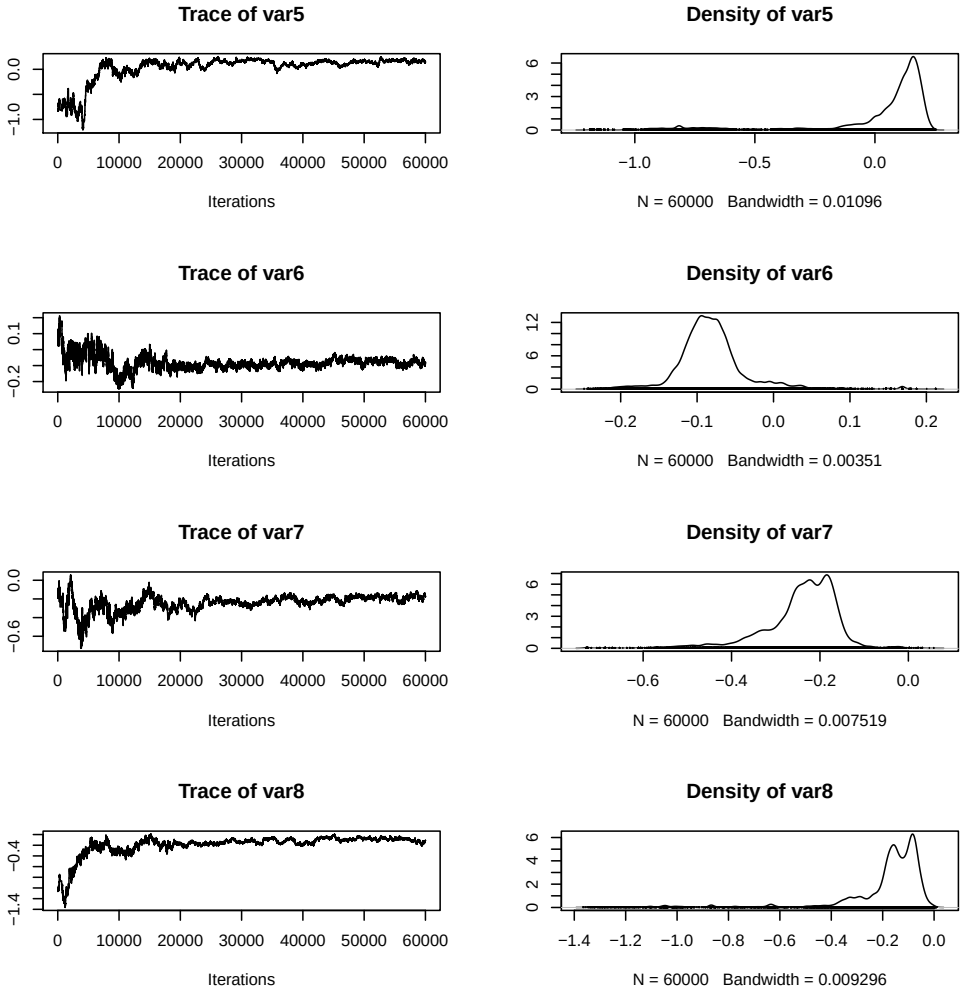


Figure A.2: Trace plots for $\beta_5, \beta_6, \beta_7, \beta_8$.

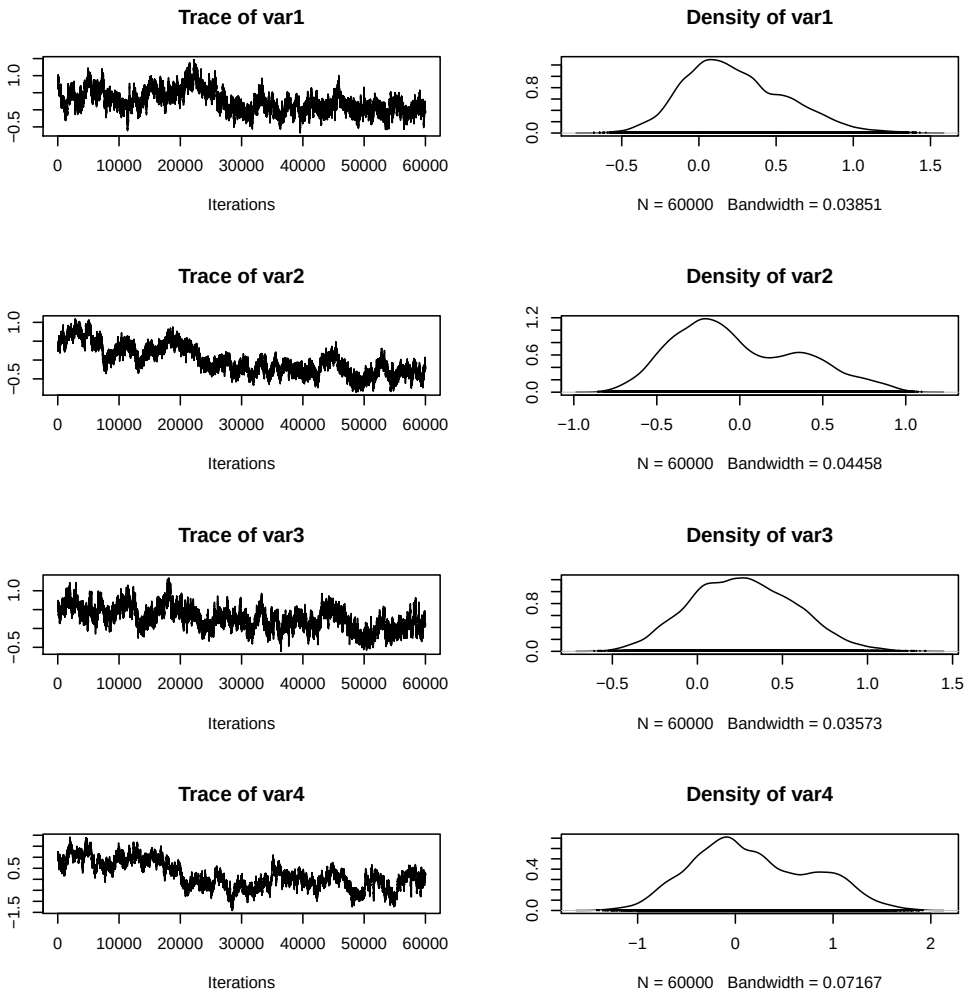


Figure A.3: Trace plots for c_1, c_2, c_3, c_4.

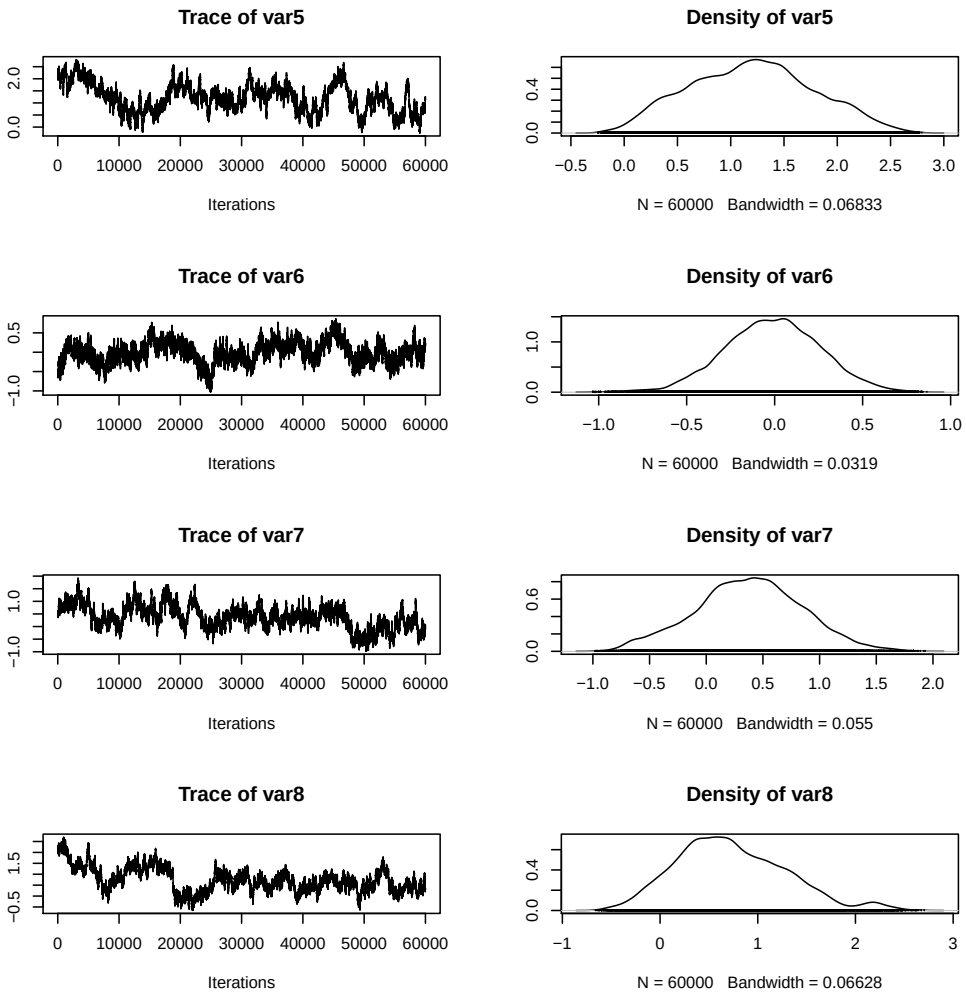


Figure A.4: Trace plots for c_5, c_6, c_7, c_8.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Lag 1	0.68	0.86	0.70	0.71	0.85	0.62	0.81	0.88	0.66	0.79	0.77	0.76	0.76	0.72	0.68	0.80	0.64	0.71	0.77	0.65
Lag 5	0.54	0.67	0.52	0.23	0.42	0.29	0.59	0.59	0.40	0.59	0.49	0.23	0.38	0.41	0.40	0.65	0.48	0.41	0.60	0.29
Lag 10	0.52	0.54	0.36	0.33	0.16	0.27	0.34	0.36	0.35	0.48	0.37	0.24	0.20	0.37	0.27	0.49	0.28	0.21	0.51	0.04
Lag 50	0.01	0.20	0.10	0.14	-0.19	0.23	0.15	0.09	0.11	-0.06	0.03	-0.06	0.03	-0.23	-0.24	0.16	-0.09	0.02	-0.07	0.01

	21	22	23	24	25	26	27	28	29	30	31	32	33	44	35	36	37	38	39	40
Lag 1	0.88	0.61	0.65	0.77	0.60	0.74	0.77	0.83	0.76	0.71	0.88	0.73	0.64	0.70	0.93	0.68	0.69	0.73	0.73	0.67
Lag 5	0.50	0.38	0.34	0.67	0.32	0.44	0.39	0.64	0.68	0.48	0.56	0.52	0.38	0.21	0.71	0.19	0.53	0.34	0.51	0.36
Lag 10	0.22	0.25	0.17	0.49	0.08	0.13	0.19	0.48	0.57	0.36	0.49	0.45	0.24	0.08	0.47	0.08	0.55	0.22	0.37	0.16
Lag 50	0.17	-0.01	0.06	0.08	0.13	0.01	-0.02	-0.03	0.06	0.07	-0.15	-0.28	-0.01	0.01	0.13	-0.00	0.29	-0.16	0.07	-0.05

Table A.1: Autocorrelation for coefficients of β at lags$= 1, 5, 10, 50$.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Lag 1	0.73	0.75	0.56	0.66	0.80	0.67	0.69	0.69	0.77	0.84	0.67	0.86	0.80	0.52	0.71	0.53	0.75	0.75	0.72	0.73
Lag 5	0.56	0.48	0.36	0.37	0.46	0.53	0.44	0.44	0.66	0.77	0.42	0.68	0.46	0.19	0.51	0.24	0.55	0.58	0.49	0.63
Lag 10	0.42	0.31	0.23	0.16	0.28	0.25	0.21	0.17	0.51	0.64	0.42	0.56	0.34	0.15	0.41	0.13	0.41	0.50	0.43	0.51
Lag 50	-0.08	0.13	-0.05	-0.09	-0.00	0.06	-0.00	-0.01	0.16	0.12	-0.05	-0.15	0.02	-0.06	0.01	0.06	-0.15	0.33	0.09	-0.08

	21	22	23	24	25	26	27	28	29	30	31	32	33	44	35	36	37	38	39	40
Lag 1	0.85	0.56	0.68	0.78	0.46	0.75	0.76	0.58	0.85	0.81	0.90	0.64	0.56	0.50	0.76	0.81	0.68	0.66	0.74	0.58
Lag 5	0.53	0.44	0.58	0.60	0.22	0.53	0.49	0.34	0.75	0.72	0.59	0.37	0.32	0.19	0.41	0.65	0.51	0.52	0.52	0.22
Lag 10	0.44	0.27	0.49	0.43	0.03	0.43	0.39	0.16	0.70	0.64	0.38	0.28	0.13	0.06	0.12	0.58	0.40	0.42	0.44	0.14
Lag 50	0.06	-0.05	0.12	0.16	0.01	-0.01	-0.10	-0.02	0.18	0.12	-0.07	-0.02	-0.08	0.09	0.17	0.11	0.11	-0.04	0.09	0.03

Table A.2: Autocorrelation for coefficients of c at lags= $1, 5, 10, 50$.

A.2 Bayesian density regression

Simulation

One of the simulations on which we applied the Bayesian density regression model was

$$\begin{aligned} Y_i|X_i = \boldsymbol{x}_i \sim\ & 0.3\text{COM-Poisson}(\exp\{x_{i1}\}, \exp\{x_{i1}\}) \\ & + 0.7\text{COM-Poisson}(\exp\{2 - 2x_{i1}\}, \exp\{1 + x_{i1}\}), \end{aligned}$$

where $x_{i1} \sim \text{Unif}(0, 1)$ and $\boldsymbol{x}_i = (1, x_{i1})^\intercal$. For more information one can see page 134. Figure A.5 shows the number of "active" clusters for each iteration, after the burn-in period. Table A.3 shows the empirical probabilities of having K clusters, after the burn-in period. Figure A.6 shows the cumulative mean probabilities for each value of y (in colour) along with the true probabilities (in grey) for $y = 0, 1, \ldots, 15$ and $x_{i1} = 0.25$ while Figures A.7 and A.8 show the 95% highest posterior density intervals for the estimated probabilities and the KL divergence between the true probability distribution and the cumulative means of the estimated probabilities respectively (for $x_{i1} = 0.25$).

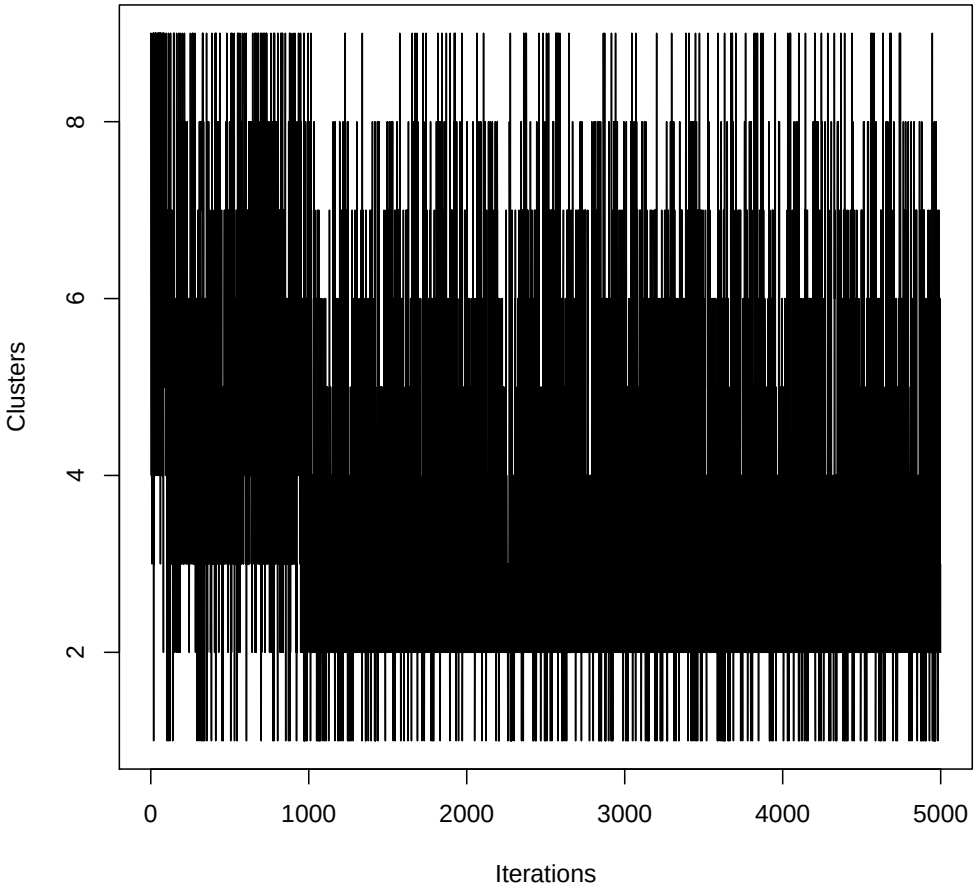


Figure A.5: Number of clusters across iterations.

Number of clusters	Probability
1	0.05
2	0.34
3	0.15
4	0.12
5	0.10
6	0.11
7	0.06
8	0.04
9	0.03

Table A.3: Empirical probabilities of having K clusters.

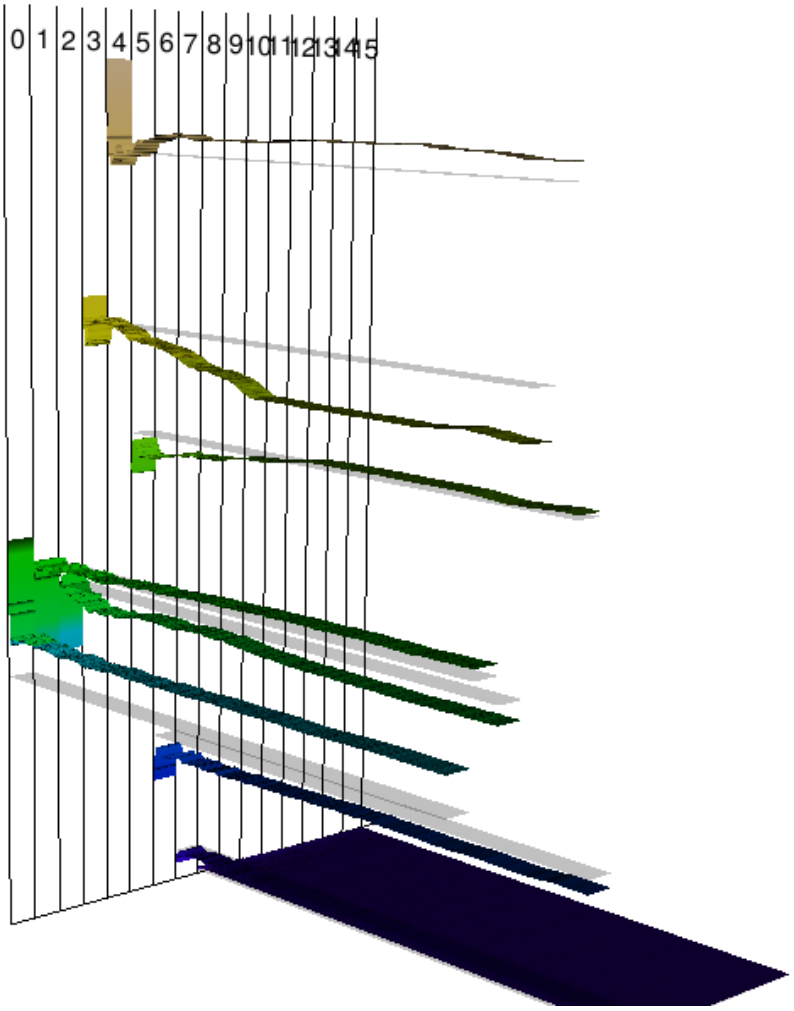


Figure A.6: Cumulative mean probabilities for each value of y (in colour) along with the true probabilities (in grey) for $y = 0, 1, \ldots, 15$, for $x_{i1} = 0.25$.

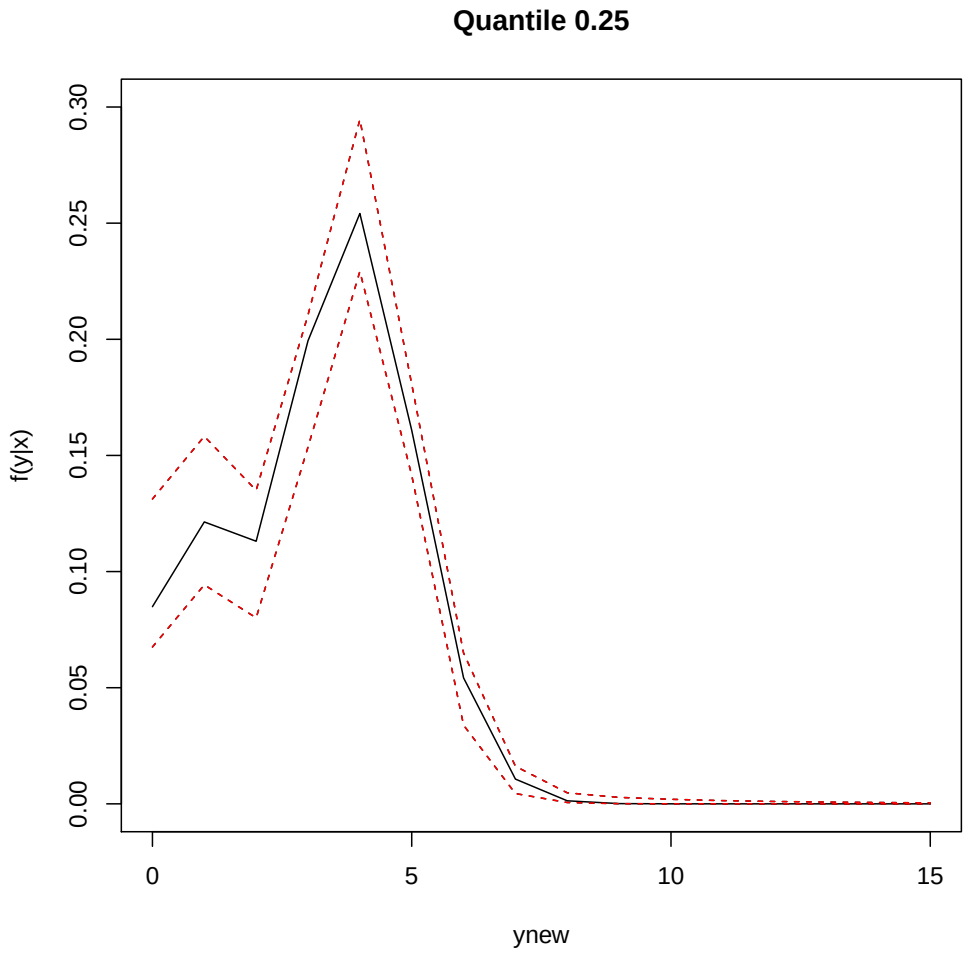


Figure A.7: The 95% highest posterior density intervals for the estimated probabilities along with the true probabilities, for $x_{i1} = 0.25$.

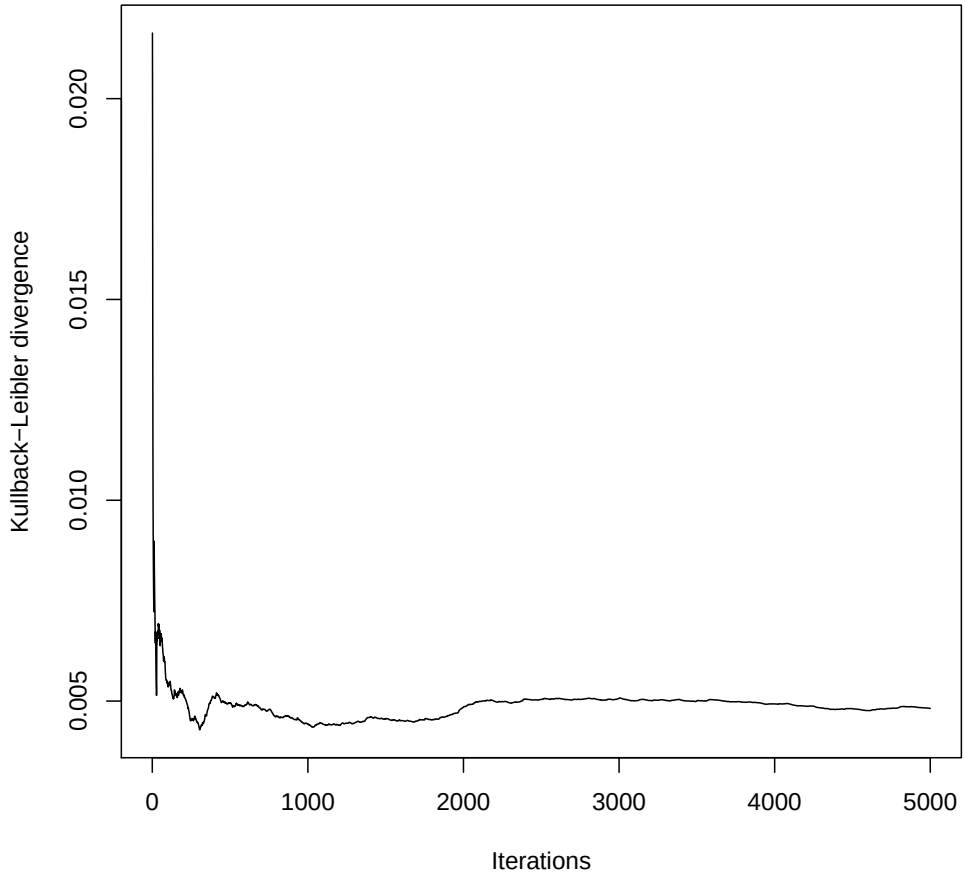


Figure A.8: KL divergence between the true probability distribution and the cumulative means of the estimated probabilities for $x_{i1} = 0.25$.

Appendix B

(More) Simulations

Bayesian density regression

In addition to the simulations in Chapter 5 we also simulate data from the following distributions:

$$\begin{aligned} Y_i|X_i = x_i &\sim 0.9\text{COM-Poisson}(\exp\{1 - x_{i1}\}, \exp\{3 + x_{i1}\}) \\ &\quad + 0.1\text{COM-Poisson}(\exp\{1 + 3x_{i1}\}, \exp\{4 + 4x_{i1}\}), \\ Y_i|X_i = x_i &\sim \text{Binomial}(20, 0.5), \end{aligned}$$

where $x_{i1} \sim \text{Unif}(0, 1)$ and $\boldsymbol{x}_i = (1, x_{i1})^\intercal$. We will implement the MCMC algorithm seen in 4.2, for 10000 iterations with the first 5000 used as a burn-in period. In the first simulation the second mixture component has small probability weight and its mean changes radically across the covariate space. Figure B.1 shows the approximation for the probability mass function of the first simulation (for $N = 100$). The top panel of Figure B.2 shows

the true quantiles (dotted lines) and the estimated quantiles (solid lines) using our approach while the bottom panel shows the true quantiles and the quantiles from using the "jittering" method of Machado and Santos Silva (2005). Figures B.3 and B.4 refer to the second simulation.

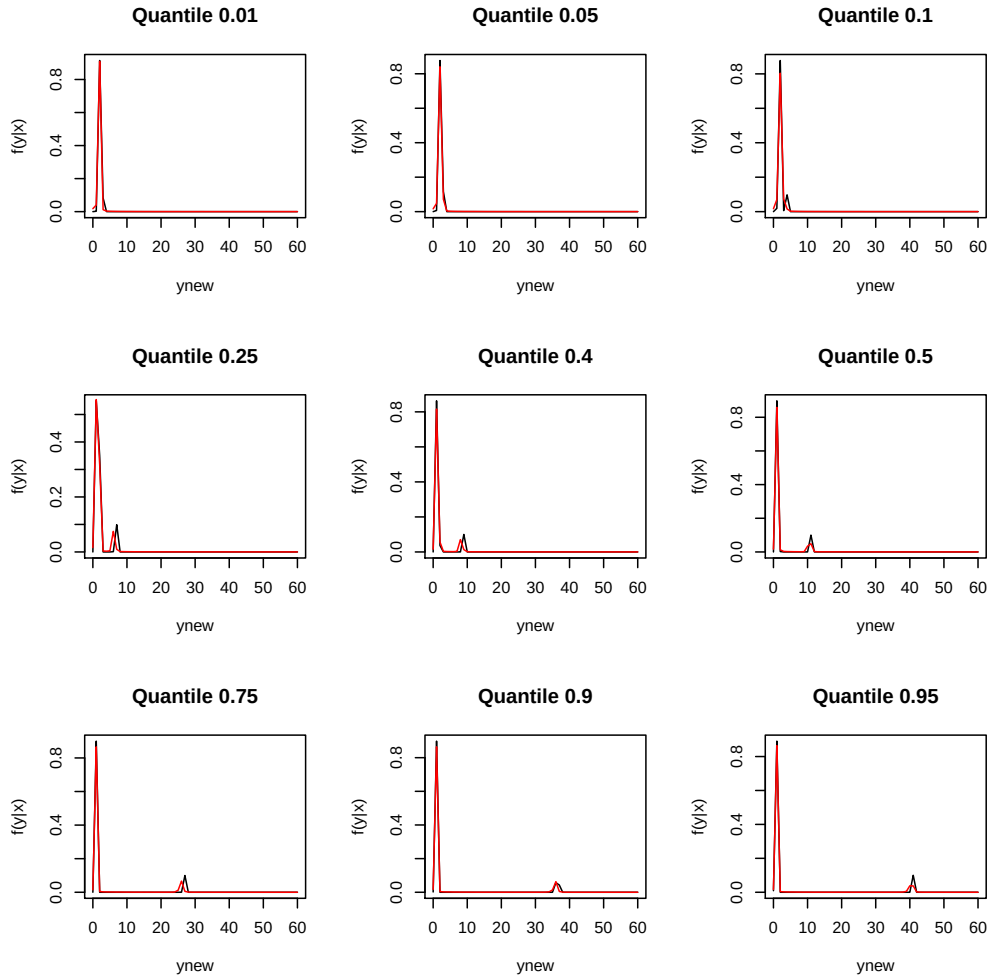


Figure B.1: True probability mass function is in black and posterior mean estimates are in red. The plots refer to the quantiles $q = 0.01, 0.05, 0.1, 0.25, 0.40, 0.5, 0.75, 0.9, 0.95$ of the empirical distribution of x_{i1}.

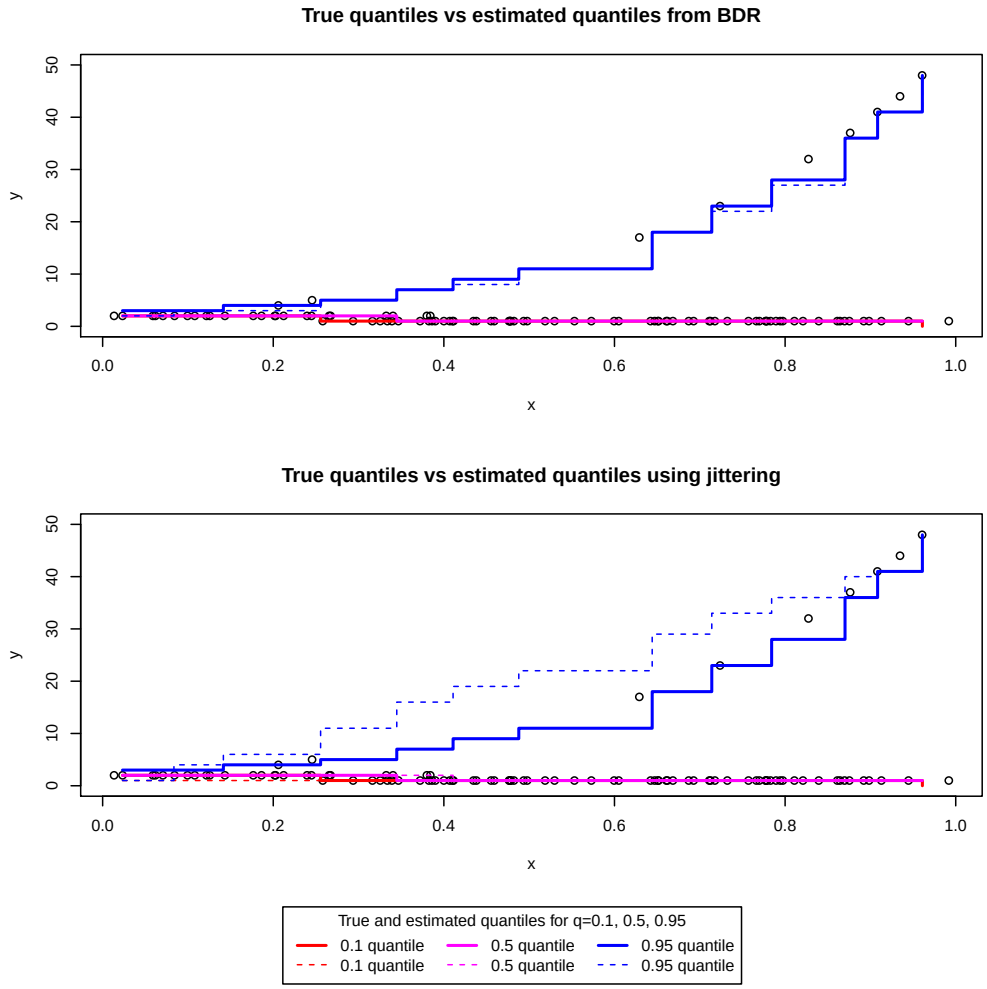


Figure B.2: The data for the first, in the appendix, simulated example, along with the quantiles for $q = 0.1, 0.5, 0.95$ across the covariate space.

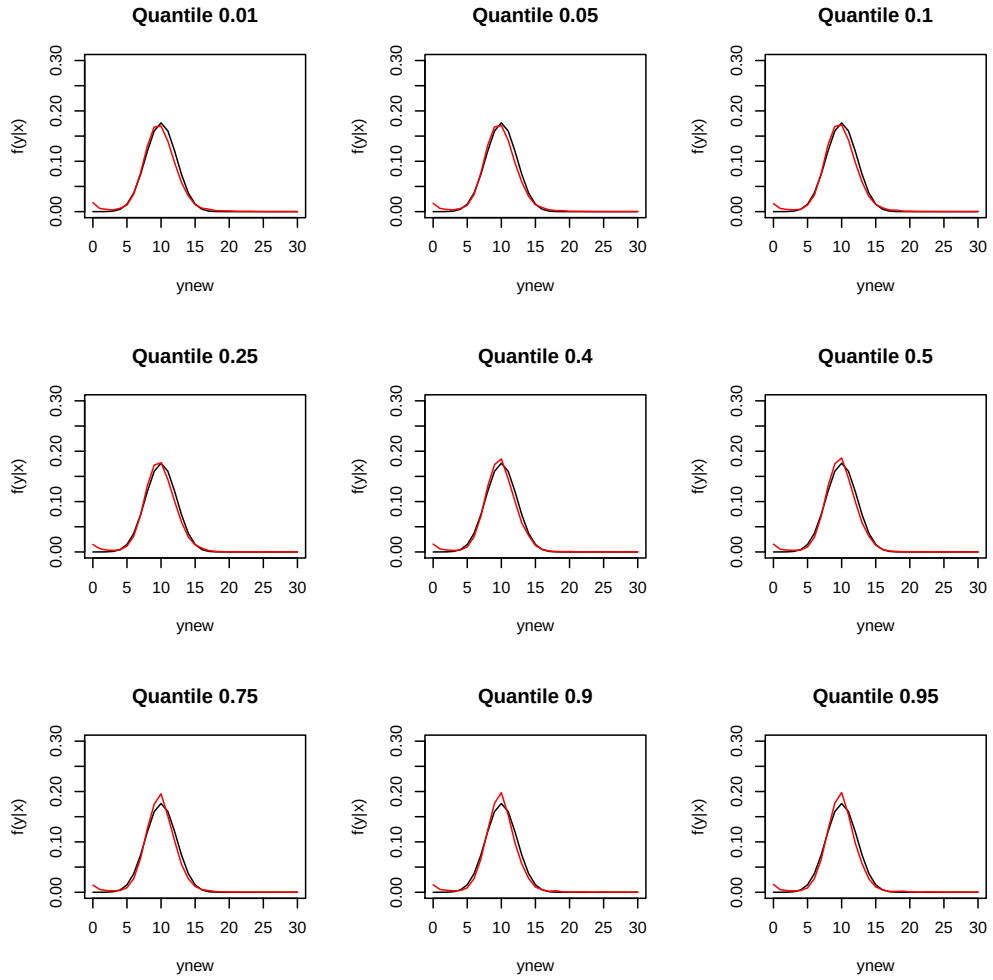


Figure B.3: True probability mass function is in black and posterior mean estimates are in red. The plots refer to the quantiles $q = 0.01, 0.05, 0.1, 0.25, 0.40, 0.5, 0.75, 0.9, 0.95$ of the empirical distribution of x_{i1}.

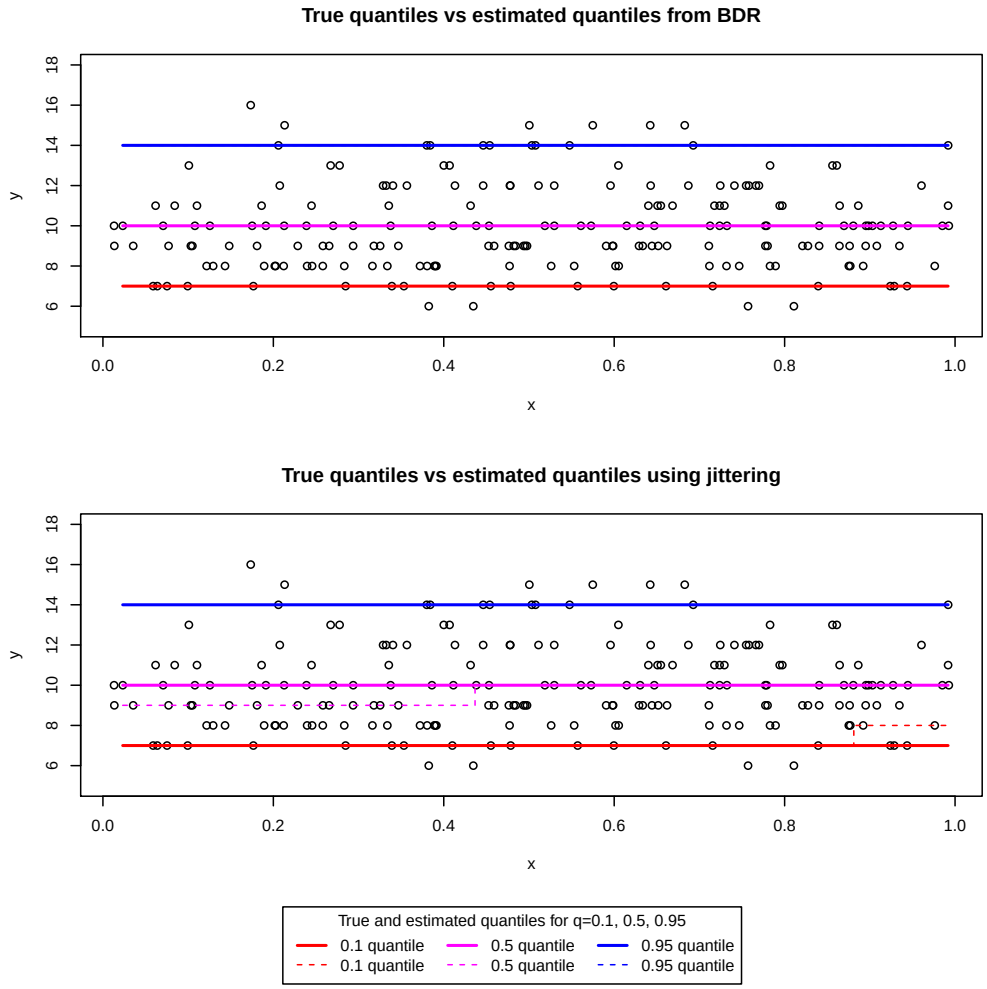


Figure B.4: The data for the second, in the appendix, simulated example, along with the quantiles for $q = 0.1, 0.5, 0.95$ across the covariate space.

Bibliography

Alquier, Pierre; Friel, Nial; Everitt, Richard, and Boland, Aidan. Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *arXiv preprint arXiv:1403.5496*, 2014.

Antoniak, Charles E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152–1174, 11 1974. doi: 10.1214/aos/1176342871. URL http://dx.doi.org/10.1214/aos/1176342871.

Belitz, Christiane; Brezger, Andreas; Kneib, Thomas, and Lang, Stefan. *BayesX: Software for Bayesian Inference in Structured Additive Regression Models (Methodology manual)*, 2009. URL http://www.BayesX.org/. Version 2.0.1.

Besag, Julian; York, Jeremy, and Mollié, Annie. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, 1991. ISSN 0020-3157. doi: 10.1007/BF00116466. URL http://dx.doi.org/10.1007/BF00116466.

Beskos, Alexandros; Papaspiliopoulos, Omiros; Roberts, Gareth O., and

Fearnhead, Paul. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B*, 68(3):333–382, 2006. doi: 10.1111/j.1467-9868.2006.00552.x.

Biernacki, Christophe; Celeux, Gilles, and Govaert, Gérard. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3):561–575, 2003.

Bondell, Howard D.; Reich, Brian J., and Wang, Huixia. Noncrossing quantile regression curve estimation. *Biometrika*, 97(4):825–838, 2010. doi: 10.1093/biomet/asq048. URL http://biomet.oxfordjournals.org/content/97/4/825.abstract.

Booth, Alison and Kee, Hiau Joo. Intergenerational transmission of fertility patterns. *Oxford Bulletin of Economics and Statistics*, 71 (2):183–208, 04 2009. URL http://ideas.repec.org/a/bla/obuest/v71y2009i2p183-208.html.

Buchinsky, Moshe. Recent advances in quantile regression models: A practical guideline for empirical research. *The Journal of Human Resources*, 33 (1):88–126, 1998. ISSN 0022166X. doi: 10.2307/146316.

Buuren, Stef van and Fredriks, Miranda. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, 20 (8):1259–1277, 2001.

Cameron, Colin A. and Johansson, Per. Count data regression using series expansions: with applications. *Journal of Applied Econometrics*, 12

(3):203–223, 1997. ISSN 1099-1255. URL http://dx.doi.org/10.1002/(SICI)1099-1255(199705)12:3<203::AID-JAE446>3.0.CO;2-2.

Cameron, Colin A. and Trivedi, Pravin K. *Regression analysis of count data*. Number 53. Cambridge university press, 2013.

Caron, Francois; Davy, Manuel; Doucet, Arnaud; Duflos, Emmanuel, and Vanheeghe, Philippe. Bayesian inference for linear dynamic models with Dirichlet process mixtures. *Signal Processing, IEEE Transactions on*, 56 (1):71–84, 2008.

Carroll, Raymond J. and Ruppert, David. Robust estimation in heteroscedastic linear models. *The annals of statistics*, pages 429–441, 1982.

Casella, George and Robert, Christian P. A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. Economics papers from University Paris Dauphine, Paris Dauphine University, 2011. URL http://ideas.repec.org/p/dau/papers/123456789-3549.html.

Chamberlain, Gary. Quantile regression, censoring and the structure of wages. In *Advances in Econometrics*. Elsevier, 1994.

Chanialidis, Charalampos; Evers, Ludger; Neocleous, Tereza, and Nobile, Agostino. Retrospective sampling in MCMC with an application to COM-Poisson regression. *Stat*, 3(1):273–290, 2014. ISSN 2049-1573. doi: 10.1002/sta4.61. URL http://dx.doi.org/10.1002/sta4.61.

Chernozhukov, Victor; Fernández-Val, Iván, and Galichon, Alfred. Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575, 2009.

Chernozhukov, Victor; Fernández-Val, Iván, and Galichon, Alfred. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010. ISSN 1468-0262. doi: 10.3982/ECTA7880. URL http://dx.doi.org/10.3982/ECTA7880.

Chib, Siddhartha and Greenberg, Edward. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, Nov 1995. URL http://www.jstor.org/stable/2684568.

Consul, P. C. and Famoye, Felix. Generalised Poisson regression model. *Communications in Statistics - Theory and Methods*, 21(1):89–109, 1992. doi: 10.1080/03610929208830766.

Conway, Richard W. and Maxwell, William L. A queuing model with state dependent service rate. *Journal of Industrial Engineering*, 12:132–136, 1962.

Del Castillo, Joan and Pérez-Casany, Marta. Weighted Poisson distributions for overdispersion and underdispersion situations. *Annals of the Institute of Statistical Mathematics*, 50(3):567–585, 1998. ISSN 0020-3157. doi: 10.1023/A:1003585714207.

Dempster, Arthur P.; Laird, Nan M., and Rubin, Donald B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, pages 1–38, 1977.

Dette, Holger and Volgushev, Stanislav. Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society: Series B*, 70(3):609–627, 2008. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2008.

00651.x. URL http://dx.doi.org/10.1111/j.1467-9868.2008.00651.x.

Dobson, Annette J. *An introduction to generalised linear models*. CRC press, 2001.

Dunson, David B. Empirical Bayes density regression. *Statistica Sinica*, 17: 481–504, 2007.

Dunson, David B. and Herring, Amy H. Semiparametric Bayesian latent trajectory models. 2006.

Dunson, David B. and Park, Ju-Hyun. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008. doi: 10.1093/biomet/asn012. URL http://biomet.oxfordjournals.org/content/95/2/307.abstract.

Dunson, David B. and Stanford, Joseph B. Bayesian inferences on predictors of conception probabilities. *Biometrics*, 61(1):126–133, 2005. ISSN 1541-0420. doi: 10.1111/j.0006-341X.2005.031231.x.

Dunson, David B.; Pillai, Natesh, and Park, Ju-Hyun. Bayesian density regression. *Journal of the Royal Statistical Society: Series B*, 69(2):163–183, 2007. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2007.00582.x.

Efron, Bradley. Double exponential families and their use in generalised linear regression. *Journal of the American Statistical Association*, 81(395): 709–721, 1986. doi: 10.1080/01621459.1986.10478327. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478327.

Escobar, Michael D. Estimating normal means with a Dirichlet process prior.

Journal of the American Statistical Association, 89(425):268–277, 1994. ISSN 01621459.

Escobar, Michael D. and West, Mike. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90: 577–588, 1994.

Fan, Jianqing; Yao, Qiwei, and Tong, Howell. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996. doi: 10.1093/biomet/83.1.189. URL http://biomet.oxfordjournals.org/content/83/1/189.abstract.

Ferguson, Thomas S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973. ISSN 0090-5364. doi: 10.1214/aos/1176342360.

Ferguson, Thomas S. Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, pages 287–302. Academic Press, 1983.

Frigyik, Bela A.; Kapila, Amol, and Gupta, Maya R. Introduction to the Dirichlet distribution and related processes. Technical Report 206, 2010.

Frühwirth-Schnatter, Sylvia. *Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes*. Springer, 2006.

Gagnon, David R.; Doron-LaMarca, Susan; Bell, Margret; O'Farrell, Timothy J, and Taft, Casey T. Poisson regression for modeling count and frequency outcomes in trauma research. *Journal of Traumatic Stress*, 21 (5):448–454, 2008.

Gelfand, Alan E. and Kottas, Athanasios. Nonparametric Bayesian modeling for stochastic order. *Annals of the Institute of Statistical Mathematics*, 53 (4):865–876, 2001.

Gelfand, Alan E.; Kottas, Athanasios, and MacEachern, Steven N. Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035, 2005.

Gelman, Andrew and Rubin, Donald B. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

Gelman, Andrew; Carlin, John B.; Stern, Hal S., and Rubin, Donald B. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

Gelman, Andrew; Carlin, John B; Stern, Hal S; Dunson, David B; Vehtari, Aki, and Rubin, Donald B. *Bayesian Data Analysis*. CRC Press, 3rd ed. edition, 2013.

Gelman, Andrew; Hwang, Jessica, and Vehtari, Aki. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

Ghosh, Jayanta K.; Tokdar, Surya T., and Zhu, Yu M. Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Analysis*, 5(2):319–344, 2010.

Griffin, Jim E. and Steel, Mark F.J. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006.

Guikema, Seth D. and Coffelt, Jeremy P. A flexible count data regression model for risk analysis. *Risk analysis : an official publication of the Society for Risk Analysis*, 28:213–223, 2008. ISSN 1539-6924. doi: 10.1111/j.1539-6924.2008.01014.x.

Hall, Byron. *LaplacesDemon Examples*, 2012. URL http://cran.r-project.org/web/packages/LaplacesDemon/index.html. R package version 12.05.07.

Hall, Peter; Wolff, Rodney C.L., and Yao, Qiwei. Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94(445):154–163, 1999.

Hall, Peter; Racine, Jeff, and Li, Qi. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468), 2004.

Hannah, Lauren A; Blei, David M., and Powell, Warren B. Dirichlet process mixtures of generalised linear models. *arXiv preprint arXiv:0909.5194*, 2009.

Hastie, Trevor J. and Tibshirani, Robert J. Generalised additive models. *Statistical science*, pages 297–310, 1986.

Hastie, Trevor J. and Tibshirani, Robert J. *Generalised additive models*, volume 43. CRC Press, 1990.

Hastings, W. Keith. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. doi: 10.1093/biomet/

57.1.97. URL http://biomet.oxfordjournals.org/content/57/1/97.abstract.

He, Xuming. Quantile curves without crossing. *The American Statistician*, 51(2):186–192, 1997.

Heller, Gillian Z; Mikis Stasinopoulos, D; Rigby, Robert A, and De Jong, Piet. Mean and dispersion modelling for policy claims costs. *Scandinavian Actuarial Journal*, 2007(4):281–292, 2007.

Hilbe, J. M. Cambridge University Press, 2007.

Holmes, Chris C; Held, Leonhard, and others, . Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1): 145–168, 2006.

Ishwaran, Hemant and James, Lancelot F. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453), 2001.

Ishwaran, Hemant and Rao, Sunil J. Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, pages 730–773, 2005.

Ismail, Noriszura and Jemain, Abdul Aziz. Handling overdispersion with negative binomial and generalised Poisson regression models. *Casualty Actuarial Society Forum*, pages 103–158, 2007.

Jara, Alejandro; Hanson, Timothy E.; Quintana, Fernando A.; Müller, Peter, and Rosner, Gary L. DPpackage: Bayesian non- and semi-parametric modelling in R. *Journal of statistical software*, 40(5):1–30, 2011. ISSN 1548-7660.

Johnson, Norman L; Kemp, Adrienne W, and Kotz, Samuel. *Univariate discrete distributions*, volume 444. John Wiley & Sons, 2005.

Kadane, Joseph B.; Shmueli, Galit; Minka, Thomas P.; Borle, Sharad, and Boatwright, Peter. Conjugate analysis of the Conway-Maxwell-Poisson distribution. 1(2):363–374, 2006.

Kalyanam, Kirthi; Borle, Sharad, and Boatwright, Peter. Deconstructing each items category contribution. *Marketing Science*, 26(3):327–341, 2007. doi: 10.1287/mksc.1070.0270. URL http://pubsonline.informs.org/doi/abs/10.1287/mksc.1070.0270.

Karlis, Dimitris and Xekalaki, Evdokia. Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3):577–590, 2003.

Kneib, Thomas; Konrath, Susanne, and Fahrmeir, Ludwig. High-dimensional structured additive regression models: Bayesian regularisation, smoothing and predictive performance, 2009. URL http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-9032-2.

Koenker, Roger. A note on l-estimates for linear models. *Statistics & Probability Letters*, 2(6):323 – 325, 1984. ISSN 0167-7152. doi: http://dx.doi.org/10.1016/0167-7152(84)90040-3. URL http://www.sciencedirect.com/science/article/pii/0167715284900403.

Koenker, Roger and Bassett, Gilbert. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Koenker, Roger and Hallock, Kevin F. Quantile regression. *The Journal*

of Economic Perspectives, 15(4):143–156, 2001. ISSN 08953309. doi: 10.2307/2696522.

Lee, Duncan. A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, 2(2):79–89, 2011. ISSN 1877-5853. URL http://www.biomedsearch.com/nih/comparison-conditional-autoregressive-models-used/22749587.html.

Lee, Duncan and Neocleous, Tereza. Bayesian quantile regression for count data with application to environmental epidemiology. *Journal of the Royal Statistical Society: Series C*, 59(5):905–920, 2010. ISSN 0035-9254 (print), 1467-9876 (electronic).

Leroux, Brian; Lei, Xingye, and Breslow, Norman. Estimation of disease rates in small areas: A new mixed model for spatial dependence. In Halloran, M.Elizabeth and Berry, Donald, editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, volume 116 of *The IMA Volumes in Mathematics and its Applications*, pages 179–191. Springer New York, 2000. ISBN 978-1-4612-7078-2. doi: 10.1007/978-1-4612-1284-3_4. URL http://dx.doi.org/10.1007/978-1-4612-1284-3_4.

Levin, Jesse. For whom the reductions count: A quantile regression analysis of class size and peer effects on scholastic achievement. *Empirical Economics*, 26(1):221–246, 2001.

Lindsay, Bruce G. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages 1–163. JSTOR, 1995.

Long, J. Scott. The origins of sex differences science. *Social Forces*, 68(4): 1297–1315, 1990.

Lord, Dominique; Guikema, Seth D, and Geedipally, Srinivas Reddy. Application of the Conway-Maxwell-Poisson generalised linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention*, 40(3): 1123–1134, 2008.

MacEachern, Steven N. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics - Simulation and Computation*, 23(3):727–741, 1994. doi: 10.1080/03610919408813196. URL http://dx.doi.org/10.1080/03610919408813196.

MacEachern, Steven N. Dependent nonparametric processes. 1999.

MacEachern, Steven N. Decision theoretic aspects of dependent nonparametric processes. *Bayesian methods with applications to science, policy and official statistics*, pages 551–560, 2001.

MacEachern, Steven N. and Müller, Peter. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2): 223–238, 1998. ISSN 10618600. doi: 10.2307/1390815.

Machado, José António Ferreira and Santos Silva, João M.C. Quantiles for counts. *Journal of the American Statistical Association*, 100:1226–1237, 2005.

McLachlan, Geoffrey and Peel, David. *Finite mixture models*. John Wiley & Sons, 2004.

Meligkotsidou, Loukia; Vrontos, Ioannis D., and Vrontos, Spyridon D. Quantile regression analysis of hedge fund strategies. *Journal of Empirical Finance*, 16(2):264–279, 2009.

Metropolis, Nicholas; Rosenbluth, Arianna W.; Rosenbluth, Marshall N.; Teller, Augusta H., and Teller, Edward. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1953.

Minka, Thomas P.; Shmueli, Galit; Kadane, Joseph B.; Borle, Sharad, and Boatwright, Peter. Computing with the COM-Poisson distribution. Technical report, CMU Statistics Department, 2003.

Mira, Antonietta. On Metropolis-Hastings algorithms with delayed rejection. 2001.

Miranda, Alfonso. Planned fertility and family background: a quantile regression for counts analysis. *Journal of Population Economics*, 21(1): 67–81, January 2008. URL http://ideas.repec.org/a/spr/jopoec/v21y2008i1p67-81.html.

Mitchell, Toby J. and Beauchamp, John J. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404): 1023–1032, 1988.

Møller, J.; Pettitt, A. N.; Reeves, R., and Berthelsen, K. K. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006. doi: 10.1093/biomet/93.2.451.

Moreira, Sara and Barros, Pedro Pita. Double health insurance coverage

and health care utilisation: evidence from quantile regression. *Health Economics*, 19(9):1075–1092, 2010. URL http://EconPapers.repec.org/RePEc:wly:hlthec:v:19:y:2010:i:9:p:1075-1092.

Müller, Peter and Quintana, Fernando A. Nonparametric Bayesian data analysis. *Statistical Science*, 19(1):95–110, 2004. ISSN 08834237. doi: 10.2307/4144375.

Müller, Peter; Erkanli, Alaattin, and West, Mike. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83(1):67–79, 1996. doi: 10.1093/biomet/83.1.67. URL http://dx.doi.org/10.1093/biomet/83.1.67.

Murray, Ian; Ghahramani, Zoubin, and MacKay, David J. C. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366. AUAI Press, 2006.

Neal, Radford M. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2): 249–265, 2000. ISSN 10618600. doi: 10.2307/1390653.

Nelder, J. A. and Wedderburn, R. W. M. Generalised linear models. *Journal of the Royal Statistical Society: Series A*, 135(3):370–384, 1972. URL http://www.jstor.org/stable/2344614.

Neocleous, Tereza and Portnoy, Stephen. On monotonicity of regression quantile functions. *Statistics & Probability Letters*, 78(10):1226–1229, 2008.

Orbanz, Peter and Teh, Yee W. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*. Springer, 2010.

Papaspiliopoulos, Omiros and Roberts, Gareth O. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95:169–186, 2008. ISSN 1464-3510. doi: 10.1093/biomet/asm086.

Park, Trevor and Casella, George. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. doi: 10.1198/016214508000000337. URL http://dx.doi.org/10.1198/016214508000000337.

Pilla, Ramani S. and Lindsay, Bruce G. Alternative EM methods for nonparametric finite mixture models. *Biometrika*, 88(2):535–550, 2001.

Plummer, Martyn; Best, Nicky; Cowles, Kate, and Vines, Karen. CODA: Convergence diagnostics and output analysis for MCMC. *R News*, 6(1): 7–11, 2006. URL http://CRAN.R-project.org/doc/Rnews/.

Powell, James L. Censored regression quantiles. *Journal of Econometrics*, 32(1):143–155, 1986. URL http://EconPapers.repec.org/RePEc:eee:econom:v:32:y:1986:i:1:p:143-155.

Powell, James L. Estimation of monotonic regression models under quantile restrictions. *Nonparametric and semiparametric methods in Econometrics,(Cambridge University Press, New York, NY)*, pages 357–384, 1991.

Qin, Xiao and Reyes, Perla E. Conditional quantile analysis for crash count data. *Journal of Transportation Engineering*, 137(9):601–607, 2011.

doi: 10.1061/(ASCE)TE.1943-5436.0000247. URL http://dx.doi.org/10.1061/(ASCE)TE.1943-5436.0000247.

R Core Team, . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org. ISBN 3-900051-07-0.

Ridout, Martin S. and Besbeas, Panagiotis. An empirical model for underdispersed count data. *Statistical Modelling*, 4(1):77–89, 2004. ISSN 1471-082X. doi: 10.1191/1471082X04st064oa.

Rigby, Robert A. and Stasinopoulos, Mikis D. The gamlss project: a flexible approach to statistical modelling. 2001.

Rigby, Robert A. and Stasinopoulos, Mikis D. Generalised additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C*, 54(3):507–554, 2005.

Rigby, Robert A.; Stasinopoulos, Mikis D., and Calliope, Akantziliotou. A framework for modelling overdispersed count data, including the Poisson-shifted generalised inverse Gaussian distribution. *Computational Statistics & Data Analysis*, 53(2):381 – 393, 2008. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/j.csda.2008.07.043. URL http://www.sciencedirect.com/science/article/pii/S0167947308003861.

Robert, Christian P. and Casella, George. *Introducing Monte Carlo Methods with R (Use R)*. Springer-Verlag, 1st edition, 2009.

Rodrigues, Josemar; de Castro Mário, ; G., Cancho Vicente, and N., Balakrishnan. COM-Poisson cure rate survival models and an application to a cu-

taneous melanoma data. *Journal of Statistical Planning and Inference*, 139 (10):3605 – 3611, 2009. ISSN 0378-3758. doi: 10.1016/j.jspi.2009.04.014.

Rogers, Simon and Girolami, Mark. *A First Course in Machine Learning*. Chapman & Hall/CRC, 1st edition, 2011. ISBN 1439824142, 9781439824146.

Romundstad, Päl; Andersen, Aage, and Haldorsen, Tor. Cancer incidence among workers in the Norwegian silicon carbide industry. *American journal of epidemiology*, 153(10):978–986, 2001.

Schulze, Niels. *Applied Quantile Regression: Microeconometric, Financial, and Environmental Analyses*. PhD thesis, 2004.

Scottish Government, . Scottish index of multiple deprivation, 2009.

Sellers, Kimberly F. and Shmueli, Galit. A flexible regression model for count data. *Annals of Applied Statistics*, 4(2):943–961, 2010. doi: 10.1214/09-aoas306.

Sellers, Kimberly F. and Shmueli, Galit. Data dispersion: Now you see it... now you don't. *Communications in Statistics - Theory and Methods*, 2013. doi: 10.1080/03610926.2011.621575.

Sellers, Kimberly F.; Borle, Sharad, and Shmueli, Galit. The COM-Poisson model for count data: a survey of methods and applications. *Appl. Stochastic Models Bus. Ind.*, 28(2):104–116, 2012. doi: 10.1002/asmb.918.

Sermaidis, Giorgos; Papaspiliopoulos, Omiros; Roberts, Gareth O.; Beskos, Alexandros, and Fearnhead, Paul. Markov chain Monte Carlo for exact

inference for diffusions. *Scandinavian Journal of Statistics*, 40(2):294–321, 2013. doi: 10.1111/j.1467-9469.2012.00812.x.

Sethuraman, Jayaram. A constructive definition of Dirichlet priors. Technical report, DTIC Document, 1991.

Shmueli, Galit; Minka, Thomas P.; Kadane, Joseph B.; Borle, Sharad, and Boatwright, Peter. A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society: Series C*, 54(1):127–142, January 2005. ISSN 0035-9254. doi: 10.1111/j.1467-9876.2005.00474.x.

Spiegelhalter, David J.; Best, Nicola G., and Carlin, Bradley P. Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical report, 1998.

Spiegelhalter, David J.; Best, Nicola G.; Carlin, Bradley P., and Van Der Linde, Angelika. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, 64(4):583–639, 2002. doi: 10.1111/1467-9868.00353.

Stasinopoulos, D Mikis and Rigby, Robert A. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46, 2007.

Teh, Yee W. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 2010.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.

Titterington, D. Michael; Smith, Adrian, and Makov, Udi E. *Statistical analysis of finite mixture distributions*, volume 7. Wiley New York, 1985.

Vehtari, Aki and Gelman, Andrew. Waic and cross-validation in stan. 2014.

Walker, Stephen and Damien, Paul. Sampling methods for Bayesian nonparametric inference involving stochastic processes. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 243–254. Springer, 1998.

Wang, Lianming and Dunson, David B. Bayesian isotonic density regression. *Biometrika*, 98(3):537–551, 2011.

Watanabe, Sumio. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, 11:3571–3594, 2010.

Wedderburn, Robert W.M. Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method. *Biometrika*, 61(3):439–447, 1974.

West, Mike. *Hierarchical priors and mixture models, with application in regression and density estimation*. 1994.

Winkelmann, Rainer. Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics*, 13(4):467–474, 1995.

Winkelmann, Rainer. Health care reform and the number of doctor visits: An econometric analysis. *Journal of Applied Econometrics*, 19(4):455–472, 2004.

Winkelmann, Rainer. Reforming health care: Evidence from quantile regressions for counts. *Journal of Health Economics*, 25(1):131–145, 2006.

Wood, Simon. *Generalised additive models: an introduction with R*. CRC press, 2006.

Wu, Hui; Gao, Liwei, and Zhang, Zhanmin. Analysis of crash data using quantile regression for counts. *Journal of Transportation Engineering*, 140 (4):04013025, 2014. doi: 10.1061/(ASCE)TE.1943-5436.0000650. URL http://dx.doi.org/10.1061/(ASCE)TE.1943-5436.0000650.

Wu, Yichao and Liu, Yufeng. Stepwise multiple quantile regression estimation using non-crossing constraints. *Statistics and Its Interface*, 2:299–310, 2009.

Yu, Keming; Lu, Zudi, and Stander, Julian. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D*, 52(3):331–350, 2003. ISSN 1467-9884. doi: 10.1111/1467-9884.00363. URL http://dx.doi.org/10.1111/1467-9884.00363.