The Robustness of Count Models in the Presence of Measurement Error and Process Error

Shu Zhen Tan

May 8, 2021

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Mathematics and Statistics

Tekijä — Författare — Author
Tan Shu Zhen

Työn nimi — Arbetets titel — Title

The Robustness of Count Models in the Presence of Measurement Error and Process Error

Oppiaine — Läroämne — Subject		
Statistics		
Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
MSc Thesis	May 8, 2021	57 pages

Tiivistelmä — Referat — Abstract

In practice, outlying observations are not uncommon in many study domains. Without knowing the underlying factors to the outliers, it is appealing to eliminate the outliers from the datasets. However, unless there are scientific justification, outlier elimination amounts to alteration of the datasets. Otherwise, heavy-tailed distributions should be adopted to model the larger-than-expected variabiltiy in an overdispersed dataset.

The Poisson distribution is the standard model to model the variation in count data. However, the empirical variability in observed datsets is often larger than the amount expected by the Poisson. This leads to unreliable inferences when estimating the true effect sizes of covariates in regression modelling. It follows that the Negative Binomial distribution is often adopted as an alternative to deal with the overdispersed datasets. Nevertheless, it has been proven that both Poisson and Negative Binomial observation distributions are not robust against the outliers, in a sense that the outliers have non-negligible influence on the estimation of the covariate effect size. On the other hand, the scale mixture of quasi-Poisson distributions (called the robust quasi-Poisson model), which is constructed similarly to the construction of the Student's t-distribution, is a heavy-tailed alternative to the Poisson. It is proven to be robust against outliers. The thesis shows the theoretical evidence on the robustness of the 3 aforementioned models in a Bayesian framework.

Lastly, the thesis considers 2 simulation experiments with different kinds of the outlier source – process error and covariate measurement error, to compare the robustness between the Poisson, Negative Binomial and robust quasi-Poisson regression models in the Bayesian framework. The model robustness was assessed, in terms of the model ability to infer correctly the covariate effect size, in different combination of error probability and error variability. It was proven that the robust quasi-Poisson regression model was more robust than its counterparts because its breakdown point was relatively higher than the others, in both experiments.

Avainsanat — Nyckelord — Keywords
Count response, Outliers, Process error, Covariate measurement error, Poisson, Negative Binomial, Robust quasi-Poisson, Robust statistics, Bayesian statistics

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

Chapter 1

Introduction

In practice, atypical observations are frequently observable in many study domains. Usually, such observations are eliminated from data sets before model fitting, with the argument that they are extremely large (or small) and do not follow the general pattern of the rest of the observations. Neyman and Scott [22] stated that the outlier elimination is allowed if it can be justified with domain-related knowledge.

The existence of outliers leads to an overdispersed data set. While there are numerous sources to overdispersion, the thesis concentrates on 2 particular sources that are common in many study fields – observation error and process error. Apparently, the observation error occurs during data collection. It can be due to human or instrumental mistakes. The process error occurs when some factors, that can explain a portion of data variability, are unidentified. It is obvious that the extra variation, caused by the aforementioned errors, fail to be explained by the set of covariates of interest, particularly when lumping all sources of error, including the random error, into one error term. The model assumptions considered by the standard models, such as Normal, Poisson and Binomial distributions always fail to characterize the unexplained variation adequately.

It is important to model the error variation appropriately, because in simple modeling with only one variable, the estimation of the location parameter (the mean) depends on the specified error distribution. Similarly, in regression modelling, the estimation on the effect sizes (parameters) of covariates depends on the distribution specified on the unexplained variation. The inadequacy of characterization of the variation may lead to biased and unreliable inferences on the parameters. The phenomenon can also be explained in terms of outliers and model robustness, that is, the model is robust if the estimations based on the model is free of the impact of the outliers (or if the extra-variability is modelled adequately).

A Poisson distribution is the standard model to model the variation in count data. It is not robust in the presence of outliers due to the fact that its mean and variance are

determined by one single parameter. Thus, the Poisson often cannot adequately describe the amount of variability in count data. The idea of allowing the mean to be random extends the Poisson to the Negative Binomial distribution if the means follow a Gamma distribution. It can be proven that the Negative Binomial distribution is not completely robust though. On the other hand, the scale mixture of quasi-Poisson distributions (called the robust quasi-Poisson likelihood) proposed by West [35] allows the Poisson variance to be random and follow some distribution, analogous to the Student's t-distribution which is a scale mixture of normal distributions. It is a heavy-tailed alternative to the Poisson and comparatively more robust that its counterpart.

In the thesis, the robustness of the Poisson, Negative Binomial and robust quasi-Poisson Bayesian regression models were studied in the presence of either observation or process error, in a Bayesian framework. The model robustness was investigated via their ability to estimate the effect size of a covariate correctly, in different combination of error probability and error variability.

1.1 Earlier Research

There are several research works dealing with overdispersion for count data. The most frequently used model is the Negative Binomial model [1, 21], which assumes quadratic mean-variance relation of the response. The quasi-Poisson likelihood model [33] assumes the Poisson-like mean-variance relation. Ver Hoef [32] compared the Negative Binomial and quasi-likelihood model with a real dataset and emphasize the model selection based on sound scientific reasoning rather than the goodness of fit to the dataset. Besides, there are research concerning the extension of the classical count model. Lindén and Mäntyniemi [18] extended the classical Negative Binomial distribution to a variaties of the Negative Binomial with different mean-variance structure. Hilbe [13] discussed a variaties of Negative Binomial distributions that deals with the overdispersed data due zero counts, data censoring and so on. Some works concerning the development of count distributions that can handle both underdispersion and overdispersion [9]. While some studies focus on the specification of the mean-variance relation of the data-generating process, Tsou [31] noticed that the inference validity of Poisson model depends heavily on the correct specification of the mean-variance relation. He proposed a robust Poisson regression that provides reliable inferences when the Poisson assumption fails or the underlying process is not Poisson.

1.2 Outline

Chapter 2 considers the description of exponential dispersion family to which the Poisson and Negative Binomial distributions belong, and its corresponding Generalized Linear Models (GLMs). Chapter 3 discusses Bayes' Theorem and Monte Carlo-based Bayesian parameter estimation. The chapter aims to distinguish Bayesian approaches from Frequentist approaches. It ends with the discussion of model comparison. Chapter 4 discusses robust inferences based on a robustified likelihood. Chapter 5 covers an introductory experiment and the description of the simulation experiments and the sampling setups for Poisson, Negative Binomial and robust quasi-Poisson Bayesian regression models. Chapter 6 presents the simulation results. Lastly, the thesis ends with a discussion section.

Chapter 2

Generalized Linear Models For Count Response

A statistical model is a simplication of a data-generating process. It is built on our assumptions. Considering more assumptions increases the complexity of a model. Assume that the variation in a sampled data set is due to random sampling, the variation can be modeled simply with a suitable distribution. This is the simplest model. Assume further that a set of factors influences the generating process, then the random model is extended by including a systematic component, which is a function of the influences. In this thesis, we consider additive influence, that is often expressed in a linear form. This results in a linear regression model.

This section covers the fundamental elements that constitute a generalized linear model (GLM). It begins with a general description of a distribution in the exponential dispersion family (Section 2.1). The focus is then narrowed down to exponential-family distributions - Poisson and Negative Binomial - that are usually adopted to model count data (Section 2.2). Generalized linear models are then presented (Section 2.3), following the consideration of factors that cause the variation in the data. As in Section 2.1, the focus is narrowed down to log-linear and Negative Binomial regression models.

2.1 Exponential Dispersion Family

Usually, we assume observations are independently and identically (iid) distributed according to some distribution. The assumption implies constant variance among observations. Natural exponential distributions, such as Normal, Poisson and Binomial distributions, have the assumption of constant variance. However, the assumption is unrealistic in practice, because observations are more likely to be heterogeneous. The exponential dis-

persion family is a generalization of natural exponential family by introducing a dispersion parameter, ϕ [16].

Suppose that $(y_1, y_2, \cdots, y_n)$ are independent random variables that have the same distributional form with an equal mean but unequal variances. If the distribution of y_i is a member of the exponential dispersion family (EDF), then its probability density function (pdf) or mass function (pmf) can be written in the form of [1, 20]

$$f(y_i; \theta, \phi) = \exp\left\{\frac{y_i\theta - b(\theta)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{2.1}$$

where θ and ϕ are the natural (or canonical) parameter and dispersion parameter of the distribution, respectively. $c(y, \theta)$ is a functional term, independent of θ. $a_i(\phi) = \phi/w_i$ is a dispersion function, where ϕ is constant and w_i is a weight that may vary by observation. $b(\theta)$ is known as the cumulant function that is used to define moments of the distribution. The first-order and second-order moments - mean and variance - can be derived in terms of $b(\theta)$ and $a(\phi)$ [7]:

$$E(y_i) = \mu = b'(\theta), \tag{2.2}$$

$$Var(y_i) = b''(\theta)a_i(\phi) = V(\mu)a_i(\phi), \tag{2.3}$$

where $\theta = b'^{-1}(\mu)$ and $V(\mu)$ is called the variance function, a function of the dispersion parameter, ϕ and expected value, μ. μ is a function of the canonical parameter, θ and thus determined by θ alone or vise versa. $V(\mu)$ determines the relationship between the mean and variance of y and thus the structure of the distribution of y, in the exponential dispersion family. In other words, every exponential dispersion distribution has its unique variance function. This is the uniqueness property in EDF [7].

Following Equation 2.3, the variance of an observation, y_i can vary among observations. Therefore, $a_i(\phi)$ adds flexibility to a model by allowing observations to have the same distributional form with same mean but unequal variances. Further, if $a_i(\phi) = 1$ and $c(y_i, \phi) = c(y_i)$ for all y_i, the distribution belongs to the natural exponential family [7].

2.2 Statistical Models on Count Data

Count data are naturally positive and discrete. They are usually modeled with the basic count model - Poisson distribution. Then, the models fitted to the data can be further extended to a relatively complex count model, such as Negative Binomial and quasi-Poisson models [1], if the assumptions of Poisson are not adequate to characterize the data set.

2.2.1 Poisson Distribution

Suppose y is a count random variable and assumed to have a Poisson distribution with mean, $E(y) = \lambda$. The time or space, in which each observation, y is measured, is constant. Then, the probability mass function (pmf) is

$$f(y;\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}. \tag{2.4}$$

and the function can be written in the form of EDF:

$$\begin{aligned} f(y;\theta) &= \exp\{y\log\lambda - \exp(\log\lambda) - \log(y!)\} \\ &= \exp\{y\theta - \exp(\theta) - \log(y!)\}, \end{aligned} \tag{2.5}$$

where $b(\theta) = \exp(\theta)$, $a(\phi) = 1$ and $c(y) = -\log(y!)$. The rewritten form proves that Poisson distribution is a member of the exponential dispersion family. Besides, it is important to notice that the natural parameter of a Poisson distribution, θ is $\log(\lambda)$. The log function has an important role as a link function in GLMs (Section 2.3). Further, $a(\phi) = 1$ is constant in the Poisson distribution and shows that the Poisson distribution belongs to the one-parameter exponential family.

From the expression above, we can prove that $E(y) = var(y) = \lambda$. It shows that Poisson is characterized by the equivalance of the mean-variance relation. Poisson, thereof, is a singe-parameter distribution that is entirely determined by λ. In practice, the equivalance of mean-variance relation does not always adequately characterize the data because the variability of the data can be greater than that expected by Poisson ($var(y) > E(y)$). This phenomenan is called overdispersion. One of the reasons is lumping all sources of error - process error, sampling error, observation error - into all-in-one uncertainty, which is then specified with some distribution, instead of expressing each of them with separate model. Or, it could possibly be that modellers do not realize the existence of other sources of error, apart from sampling process [13, 14, 18]. One of the solutions to deal with overdispersion is improving the flexibility of Poisson by adding extra parameter(s). This illustrates the tradeoff between the complexity and flexibility of a model. Considering certain assumptions on the overdispersion (the mean-variance relationship), this leads to the next topic – Negative Binomial Distribution.

2.2.2 Negative Binomial (NB) Distribution

Before introducing the probability density function (pdf) of NB distribution, it is worthwhile giving the idea of how the distribution is constructed based on multiple Poisson

distributions. Suppose that the time or space, in which each observation, y_i is measured, is constant. The mean of each count observation, y_i is unequal, implying heterogeneity among observation. Without knowing the source of heterogeneity, y_i can be assumed to be $Poisson(\lambda\epsilon_i)$, where λ is an underlying expected value common to all units and ϵ_i is a multiplicative error with $E(\epsilon) = 1$ so that the mean of $\lambda\epsilon_i$ over the observations is equal to λ [19].

$$E(\lambda\epsilon) = \lim_{n\to\infty} \frac{\sum \lambda\epsilon_i}{n} = \lambda. \tag{2.6}$$

If we further make the assumption that $\epsilon_i \sim Gamma(k, \frac{1}{k})$, where k is a dispersion parameter, its mean, $E(\epsilon_i)$, is 1 and variance, $var(\epsilon_i)$, is $\frac{1}{k}$. Further, let $\tilde{\lambda}_i = \lambda\epsilon_i$. Thus, $y_i \sim Poisson(\tilde{\lambda}_i)$ and $\tilde{\lambda}_i \sim Gamma(k, \frac{\lambda}{k})$ with $E(\tilde{\lambda}_i) = \lambda$ and $var(\tilde{\lambda}_i) = \frac{\lambda^2}{k}$. The overall Poisson distribution by averaging the individual distribution, $y_i \sim Poisson(\tilde{\lambda}_i)$ over $\tilde{\lambda}_i$ results in a wider tail than a basic Poisson with the same mean, because it includes the variability around the common mean, λ_i [17]. Mathematically, it is a result of marginalizing the Poisson distribution over $\tilde{\lambda}_i$. This leads to a Negative Binomial distribution. $y \sim NB(\lambda, k)$ and its probability mass function (pmf) is

$$f(y; \lambda, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\lambda}{\lambda+k}\right)^y \left(\frac{k}{\lambda+k}\right)^k, \tag{2.7}$$

where its mean, $E(y) = \lambda$ and variance, $var(y) = \lambda + \frac{\lambda^2}{k} = \lambda(1 + \lambda/k)$. This parameterization NB2 specifies that the variance is a quadratic function of mean. Besides, NB distribution is an exponential dispersion distribution of a discrete variable if k is known.

$$\begin{aligned} f(y;\theta) &= \exp\left\{y \log\frac{\lambda}{\lambda+k} + k\log\frac{k}{\lambda+k} + \log\Gamma(y+k) - \log\Gamma(y+1) - \log\Gamma(k)\right\} \\ &= \exp\left\{y\theta + k\log(1-\exp(\theta)) + [\log\Gamma(y+k) - \log\Gamma(y+1) - \log\Gamma(k)]\right\}, \end{aligned} \tag{2.8}$$

where k is known, $\theta = \log\frac{\lambda}{\lambda+k}$, $b(\theta) = -k\log(1-\exp(\theta))$, $a(\phi) = 1$ and $c(y) = \log\Gamma(y+k) - \log\Gamma(y+1) - \log\Gamma(k)$ [13]. From the formulation, we can see that if k is unknown, the NB does not belong to the EDF because the term, $-k\log(1-\exp(\theta))$ is a function of θ and k. and cannot be denoted by $b(\theta)$.

The aforementioned parameterization is only one of many parameterizations of NB distribution [1, 13, 18]. Different parameterization leads to different assumption on the mean-variance relation of count data and the pattern of overdispersion. Further, Lindén and Mäntyniemi [18] proposed a parameterization of the NB by introducing one more dispersion parameter in the mean-variance functional relationship, to deal with scenarios in which multiple uncertainties occur.

In addition, if ϵ_i is formulated as a log-Normal error term, with $E(\epsilon_i) = 1$ and $var(\epsilon_i) = \exp(v) - 1$, for any v, the resulting distribution is also a NB distribution with the mean, λ and variance, $var(y) = \lambda(1 + var(\epsilon_i))$ [18].

2.3 Generalized Linear Models (GLMs)

In the previous section (2.2), only the response, y is involved in the model. In practice, researchers are interested in factors (or covariates) that contribute to the variation in y and the relation between them. Suppose the response variable y has independent observations $(y_1, \cdots, y_n)$ from a parametric distribution. Each observation, y_i has a $1 \times (p+1)$ covariate vector, $\mathbf{x}_i^T = [1\ x_{i1} \cdots x_{ip}]$. The inclusion of covariates is the same as dividing observation into subgroups such that units in group, i, have common combination of covariate values, $\mathbf{x}_i$ and thus common $E(y|\mathbf{x}_i)$.

When y is not distributed normally, the assumptions of classical linear models do not fit the characterization of count data. Linear models ($y_i = \mathbf{x}_i^T\beta + \epsilon_i$) assume that the error term, ϵ_i is normally distributed and has constant variance across all observations. However, the assumptions are already violated when the data, y_i has Poisson distribution where the variance increases with the mean. Thus, a generalized linear model is required by relaxing some assumptions in the normal linear models.

Generalized linear models (GLMs) allow the response distribution to be an exponential dispersion distribution [21]. The GLMs comprise of 3 fundamental components:

Random component GLMs allow the response, y_i, to have an exponential dispersion distribution. Thus, the approach relaxes the normality assumption by allowing a broad family of distributions. Further, the covariates, $\mathbf{x}_i$, are introduced into the model through the expected value, $E(y_i|x_i)$ (or θ_i). The Equation 2.1 is rewritten as

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{2.9}$$

where θ_i is a function of $E(y_i|x_i)$ (a function of covariates). The moments of the pdf are assumed to depend on the covariates, via θ_i.

Linear predictor/ systematic component This component allows the interaction between the covariates and response. It is a linear combination of a $n \times (p+1)$ model

matrix, $X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$ and a $p \times 1$ parameter vector, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$. Note that n and p indicate the number of observations and number of independent variables, respectively. The linear predictor can be written as

$$\eta_i = \sum_{j=0}^{p} \mathrm{x}_{ij}\beta_j = \mathbf{x_i}^T\boldsymbol{\beta} \tag{2.10}$$

Link function It is a function that maps $E(y_i)$ to $\eta_i = \mathbf{X}\boldsymbol{\beta}$. In other words, it links the mean of the random component with the systematic component (or the covariates). Thus, it can be written as

$$g(E(y_i)) = \eta_i \tag{2.11}$$

Note that $g(.)$ is a monotonic and differentiable function. From here, we can see that an observation with a different set of covariate values leads to a same functional distribution with different mean.

Classical linear models are the special case of GLMs. It assumes that the distribution of y_i is normal and the link function is the identity link function, such that $g(E(y_i)) = E(y_i) = \eta_i$. GLMs allow a variety of link functions to be applied in a model.

Remember that the natural parameter, θ_i is a function of $E(y_i)$. Canonical link function is a link function that equates the linear predictor to the natural parameter.

$$g(E(y_i)) = \theta_i = \mathbf{x_i}^T\boldsymbol{\beta} \tag{2.12}$$

2.3.1 Poisson Log-Linear Regression Model

Suppose y is a count response and assumed to have Poisson distribution, $y_i \overset{ind}{\sim} Poisson(\lambda_i)$ with an individual λ_i. By using its canonical link, a Poisson regression model is written as

$$\log \lambda_i = \mathbf{x_i}^T\boldsymbol{\beta}. \tag{2.13}$$

This model is called the Poisson log-linear model. This is an additive model on log scale such that one-unit increase of a covariate j, x_{ij} increases $\log \lambda_i$ by β_j, given that the rest of the covariates are fixed. It still shows the linearity, as in the classical linear model, but on the logarithm scale of count data. This model can also be written on the original scale by exponentiating both sides of the equation.

$$\lambda_i = \exp(\mathbf{x_i}^T \boldsymbol{\beta}). \tag{2.14}$$

This equation shows the multiplicative model on original scale, such that one-unit increase in a covariate j, x_{ij} leads to multiply λ_i by a factor of $\exp^{\beta_j}$.

Another way to look at the log-transformation of λ_i is that the exponentiation ensures that λ_i remain positive for all values of linear predictor. This feature certainly cannot be achieved by applying classical linear model.

The pmf can then be written in terms of $\mathbf{x_i}$ as

$$f(y; \mathbf{x_i}, \boldsymbol{\beta}) = \frac{\exp^{-\exp(\mathbf{x_i}^T \boldsymbol{\beta})} \exp(\mathbf{x_i}^T \boldsymbol{\beta})^y}{y!}. \tag{2.15}$$

2.3.2 Negative Binomial Regression Model

Section 2.2.1 has explained how Poisson fails to fit count data with extra-variability and examples on how overdispersion could arise, without the consideration of covariates. I extend the illustration to when covariates are considered in the model, here.

Suppose there are two covariates that have truly influence on the count data. Only one of them are measured, such that $\mathrm{x_i} = (\mathrm{x_{obs_i}}, \mathrm{x_{not_i}})$. The Poisson model specification becomes $\log(E(y_i)) = \beta_0 + \mathrm{x_{obs_i}}\beta_1$. Further suppose that the units having the same value of x_{obs} form groups. Each group, i is indexed by a Poisson model, $y|x_{obs} \sim Poisson(\lambda_{x_{obs}} = \exp(\beta_0 + \mathrm{x_{obs}}\beta_1))$. However, the within-group heterogeneity is larger than the Poisson random variability, due to the unmeasured x_{not} [1]. The observations in a group have different means between each other. This leads to extra-variability in each group, i.e. $var(y|x_{obs_i}) > E(y|x_{obs})$.

The NB model expands the Poisson variability by assuming that each observation in the group has its mean that deviates from the group mean, $\lambda_{x_{obs}}$, by a random Gamma error. It is expressed as $y_j|x_{obs_i} \sim Poisson(\lambda_{x_{obs}} \cdot \epsilon_j)$, where $j = \{1, \cdots, n_{xobs_i}\}$ and $\epsilon_j \sim Gamma(k, 1/k)$. The NB formulates the extra-variability due to the unknown factor, x_{obs} in terms of a Gamma uncertainty. Marginalizing the Poisson group model over the Gamma ϵ_j results in a NB group model, $y|x_{obs} \sim NB(\lambda_{x_{obs}},\ k)$ with $var(y|x_{obs}) = k(1 + k\lambda_{x_{obs}})$ (See also Section 2.2.2).

In a NB2 GLM model, the link function is the log link function, $g(\lambda_i) = \log(\lambda_i) = \mathbf{x_i}\boldsymbol{\beta}$, instead of its canonical link function, $g(\lambda_i) = \log \frac{\lambda_i}{k+\lambda_i} = \mathbf{x_i}\boldsymbol{\beta}$. (See Section 2.2.2). Hence, the pmf in terms of $\mathbf{x_i}$ is expressed as

$$f(y_i; \mathbf{x_i}, \boldsymbol{\beta}, k) = \frac{\Gamma(y_i + k)}{\Gamma(k)\Gamma(y_i + 1)} \left(\frac{\exp(\mathbf{x_i}^T \boldsymbol{\beta})}{\exp(\mathbf{x_i}^T \boldsymbol{\beta}) + k} \right)^{y_i} \left(\frac{k}{\exp(\mathbf{x_i}^T \boldsymbol{\beta})} \right)^{k}. \tag{2.16}$$

Chapter 3

Bayesian Inference

This chapter firstly presents Bayes' theorem (Section 3.1) which Bayesian inferences are based on. The theorem is then presented in terms of data and model parameters - likelihood function and prior distributions. It is followed by the discussion of the Monte Carlo Markov Chain (MCMC)- based posterior inference and the asymptotic properties of Bayesian inference (Section 3.2). It ends with the desciption of the WAIC-based model comparision (Section 3.3)

3.1 Bayes Theorem

There are 2 schools of statistical inference: frequentist and Bayesian inferences. Both approaches have different notions on the nature of uncertainty, which is expressed in probability. The difference results in different interpretations on the nature of parameters by these 2 approaches. The frequentist allows uncertainty due to random variability in the data-generating mechanism and the Bayesian allows uncertainty due to imperfect knowledge on the quantities of interest - parameters, apart from the random variability. The parameters are considered random variables in the Bayesian framework. They are random, not because of randomness, but imperfect knowledge on them. Although the unknown parameters are fixed, the uncertainty due to lack of knowledge on the parameters can be expressed in probabilities over a set of plausible values. In contrast, the Frequentist statistics describes a probability as a relative frequency over infinite trials. Following this notion, the parameters, which are fixed and certainly not subject to random variability, should not be treated as random variables and expressed in a probabilty distribution [24, 30].

Both frequentist and Bayesian approaches require model specification (likelihood specification). The frequentist approach - Maximum Likelihood Estimation (MLE) - requires a

likelihood function and find the estimate value that maximizes the likelihood, conditional on the observed data. The Bayesian approach requires a likelihood function and a prior distribution for the parameters. The frequentist approach results in a point estimate, while the Bayesian results in a set of plausible values for the parameters, with each value assigned an updated probability.

Bayesian statistics uses Bayes' Theorem to combine the uncertainties due to random variability in population and lack of knowledge on parameters. Bayes' theorem is

$$P(A|B) = P(A \cap B) = \frac{P(B|A)P(A)}{P(B)}, \tag{3.1}$$

where $P(A)$ and $P(B)$ are the probability of event A and event B. $P(A|B)$ describes the conditional probability that event A occurs, when event B has occurred. If A and B are dependent, knowing that B has occurred gives us new information about A and results in an updated probability, $P(A|B)$. In the Bayesian framework, A is replaced by θ, representing some parameter value and B by the observations, $\boldsymbol{y}$. Thus, it becomes

$$P(\theta|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\theta)P(\theta)}{P(\boldsymbol{y})} \propto \frac{L(\theta;\boldsymbol{y})P(\theta)}{P(\boldsymbol{y})} \propto L(\theta;\boldsymbol{y})P(\theta). \tag{3.2}$$

$P(\boldsymbol{y}|\theta)$ is the probability distribution of $\boldsymbol{y}$, when the parameter value is θ. In Bayes' theorem, it is treated as a function of θ, given $\boldsymbol{y}$. It is called the likelihood function, denoted by $L(\theta;\boldsymbol{y})$ [5, 24, 30]. Instead of using the complete expression of $P(\boldsymbol{y}|\theta)$, the likelihood function can be redefined by scaling $P(\boldsymbol{y}|\theta)$ by some constant multiplier, that is free of θ. Thus, it becomes $P(\boldsymbol{y}|\theta) \propto L(\theta;\boldsymbol{y})$. $P(\theta)$ is a prior distribution for the parameter of interest, θ. It can also be described as a personal probability (or probability density) that the parameter takes some value, θ. $P(\boldsymbol{y})$ is the probability of observing the observed data set. It does not depend on the parameter and acts as a normalizing (or proportionality) constant. This leads to the rightmost proportionality in the expression. Finally, $P(\theta|\boldsymbol{y})$ is the posterior distribution of θ, after observing the data. It can be described as an updated knowledge on the parameter values, after observing the data.

Equation 3.2 shows that data influences the posterior via the likelihood and the personal belief via the prior. The former is rather objective, whereas the latter can be either objective or subjective. Bayes' theorem combines these 2 sources of information via multiplication. That is, the likelihood is multiplied by the prior, at every value of θ.

3.1.1 Likelihood Function

The likelihood function, $L(\theta|\boldsymbol{y})$, can be interpreted as the strength of support, provided by the data, over a range of possible values for the parameter, θ. Besides, since $P(\boldsymbol{y}|\theta)$ is

viewed as a function of θ, it no longer follows the law of probability, in which a probability function must integrate to 1.

The expression, $L(\theta; \boldsymbol{y}) \propto P(\boldsymbol{y}|\theta)$, shows that the likelihood specification corresponds to the model specification. The underlying data-generating process is usually unknown and the model specification, thus, is subjective. However, researchers tend to select models that have nice properties (e.g. mathematical tractability, easy interpretability) out of habit. The specification is often left unjustified for its compatibility to the true data-generating process. The influence of the likelihood function on the posterior can be assessed through a sensitivity analysis of the likelihood [30].

Similar to prior specification, the analysis can begin with a simple and conventional model and if the model checking shows that the model does not fit to the data, we can either switch to alternative models or extend the model to a relatively flexible one. For count data, we may start with Poisson modeling. If the goodness-of-fit of the Poisson to the data is unsatisfactory, we may extend the model to Negative Binomial distribution.

The likelihood function plays an important role in both frequentist and Bayesian frameworks. Observed data sets impact the inference via the likelihood. In Bayesian framework, the data set updates the prior (or the uncertainty) for the parameter through the likelihood. When the sample size increases, the weight Bayes' theorem assigns to the likelihood increases and to the prior decreases. See Section 3.2.2 for the asymptotic properties of Bayesian inference.

3.1.2 Prior Distribution

The Bayesian statistics assumes that uncertainty due to random variability or imperfect knowledge can be expressed in a probability distribution. A parameter θ is considered random due to lack of knowledge and regarded as a random variable in the Bayesian framework. The uncertainty (or personal belief) about the parameter is then expressed in a probability distribution (pmf or pdf), whose support is a set of possible values for θ. (The support of a distribution includes values that are not assigned to zero.) The distribution is called the prior distribution that distinguishes the Bayesian statistics from the frequentist (or classical) statistics. Note that the prior is defined before observing the data.

As in the likelihood function [5], the informativeness of the prior at some value is evaluated via the curvature around that value. The informativeness of the likelihood depends solely on the data, whereas the informativeness of the prior depends on the modelers and the subject-matter experts. A prior can be categorized into diffuse, weakly informative or informative priors, with respect to its degree of informativeness.

A diffuse prior is used when the researchers would like to have the Bayesian inference to be objective such that the prior is independent of their beliefs or have no knowledge

of the parameter. They would like to let the data speak for itself by assigning equal prior density value (or mass probability) over the range of possible parameter values. Thus, the prior appears to be flat. It does not favor any particular value such that none of the value receives relatively higher density (or mass probability) than the rest. Therefore, in combination with the data via the likelihood, the likelihood for each parameter value is multiplied by a fixed value. The posterior turns out to be proportional to the likelihood only. This shows that the posterior is determined solely by the likelihood. The likelihood dominates the Bayes' theorem. Thus, the mode (or peak) of the posterior is at the same suggested parameter value as the likelihood. The mode is aligned with the maximum likelihood estimate. However, only Bayesian inference allows to make probability statement about the true parameter value (probability that θ lies in an interval), with the posterior. The frequentist approach, in contrast, does not own such privilege. A hypothesis, such as the true parameter is some value (e.g. the estimate value, $\hat{\theta}$) or lies in some interval, is the matter of true or false, in the frequentist framework. The hypotheses do not own probabilities. Since the true parameter is unknown, the approach can only answer how likely or unlikely the data would be observed, by assuming that the hypothesis - $\theta = \hat{\theta}$ - is true. The frequentist, however, does not account for the reasonability of the hypothesis.

A diffuse prior does not mean that the prior is non-informative for all times. Suppose a parameter is given a diffuse prior. If some transformation of the parameter is of interest as well, its prior can be derived from the prior of the parameter, via variable transformation. The prior is an the induced prior. However, the induced prior might be far from being diffuse. This prompts unintended influence on the posterior of the transformed functions. The influence is pronounced if the sample size is small. Therefore, there are arguments on naming any diffuse prior as a non-informative prior, because the diffuse prior contains implicit information concerning other parameters, instead of the parameters of interest.

However, the diffuse prior can lead to some drawbacks that modelers want to avoid by using a weakly informative prior or informative prior. The drawbacks are mainly due to the support of the prior is not restricted. One of them is that an improper diffuse prior might lead to a improper posterior. Following the probability rules, a probability function is improper if its integral over the support is not finite. The instance of an improper diffuse prior is when the parameter space assumed by the modeler is unlimited (e.g. from $-\infty$ to ∞). Another drawback is that the posterior distribution might contain a range of improbable parameter values in the support. This happens especially when the sample size is small.

Weakly informative priors sit between diffuse and informative priors, in terms of the informativeness. The main usage of weakly informative priors is to restrict the parameter space to a range of plausible values, to avoid the drawbacks of diffuse priors. However, a weakly informative prior does not contain specific information like an informative prior. Thus, it has relatively smaller impact in Bayes' theorem than an informative one. The

likelihood is, hence, weighted more in the theorem when the former is used.

Last but not least, an informative prior contains a high degree of certainty about the parameters. It contains much specific information about the parameters than the other two kinds of priors. To construct an informative prior requires a careful prior elicitation. Otherwise, the resulting posterior might not reflect the true nature of the parameters. The effect is pronounced when the sample size is small.

The influence of a prior on the posterior decreases when more data arrive. (Consider the weight of a source is inversely related to the weight of the other source.) When the sample size is sufficiently large, the influence from a prior becomes negligible. The posterior is then dominated by the data (the likelihood) and the inference approaches to be an objective one. This also implies that the prior specification is not important anymore when the sample size is large, because different priors converge to the same posterior results, assuming that the support of each prior is similar.

3.2 Posterior Inference

After the likelihood and prior specification, the parameter estimation is computed by fitting the model to the observed data. The parameter estimation, in the Bayesian framework, involves estimation of the posterior distribution of the parameters. Instead of presenting the entire posterior distribution, the posterior distribution is usually summarized with its summary statistics. The instances of summary statistics are mean, variance and moments of a distribution. Remember that the posterior is proportional to the product of the likelihood and prior. The normalizing constant is an integral (or summation) of the product, over the parameter space. It is usually difficult to evaluate the normalizing constant, because the integral is generally not in closed form. Thus, the posterior and its associated summary statistics cannot be evaluated explicitly. Monte Carlo Markov Chain-based algorithms, which is presented in the next section, can be adopted for calculating numerical approximation of integrals.

3.2.1 Monte Carlo Markov Chain (MCMC)-based Posterior Inference

Integration of a function over a space becomes difficult, when the expression is analytically intractable or the integral is taken over a high dimensional space. In Bayesian framework, integration is required for evaluation of the normalizing constant in the posterior, its summary statistics and moments as well as credible interval, and the marginal posterior distribution of a parameter. The mathematical intractability in the equations makes direct computation hardly possible. Indirect computational methods (e.g. MCMC-based

algorithms) have been widely used in Bayesian framework.

Markov chain algorithm is adopted to randomly sample from the target distribution, whereas Monte Carlo is defined as the empirical approximation of the integrals over the target distribution, using the sampled values from the Markov chain. Concisely, the idea of MCMC is to approximate the integrals with a tractable summation over a sample set from the posterior distribution [2]. The Monte Carlo estimates can be proved to be unbiased and consistent to its corresponding statistic, based on the strong law of large numbers, which states that when $n \to \infty$, the average of iid observations, $\boldsymbol{y}$ (or $f(y)$) converges almost surely to the expected value of y, $E(y)$ or $(E(f(y)))$. Further, random sampling from the target distribution allows the target distribution to be known up to a constant.

The unbiasedness of the Monte Carlo estimates [2] is based on the fact that the observations are drawn independently from the same distribution. However, each state (or value) in a Markov chain is sampled, conditional on previous state. The mechanism violates the assumption of independence [30]. The degree of the autocorrelation in the chain depends on the choice of MCMC algorithm and its setup. Autocorrelation might result in poor exploration on the distribution and require to sample large amount of values to approximate the distribution. Although autocorrelation in the chain can be resolved by thinning the chain, Thinning leads to loss of the draws and makes the sampling algorithm less efficient. Besides, convergence of the Markov chain to the target distribution is a must to achieve unbiasedness in the estimation. The convergence means that the sampler has explored the entire target distribution, such that distribution (or the histogram) of the draws is almost similar to the target distribution and remains unchanged, as the number of draws reaches or exceeds some (large) value. Autocorrelation highly affects the Markov chain to converge efficiently. An ideal Monte Carlo estimate is computed with the sample that is collected after the convergence.

Independence in the realizations and chain convergence highly depend on the transition kernel in a MCMC-based sampler. The transition kernel determines how the next draw is proposed, conditional on the current draw and whether the stationary distribution is the target distribution. It is generally composed of a proposal distribution and an acceptance probability. The proposal distribution proposes a new parameter value or a set of values, if multiple parameters (or quantities) are of interest. The new value or the set is either accepted or rejected with the acceptance probability. If it is accepted, the Markov chain proceeds to the value, or else the chain stays at the same value at the next iteration. The well-known instances of the MCMC transition kernel are Metropolis-Hastings method [2], Hamiltonian Monte Carlo (HMC) method [4] and Gibbs sampler [2].

In this thesis, no-U-Turn sampler (NUTS)[15] was adopted via Stan [28]. The No-U-Turn Sampler is an extension form the HMC sampler and is the default sampler in Stan. The motivation of NUTS algorithm is to avoid random-walk behavior (that arises in

Metropolis-Hasting and Gibbs samplers), and prevent autocorrelation in a Markov chain. The algorithm aims to improve the efficiency of the exploration on the target distribution.

The efficiency [8] of a sampler on the target distribution should be always checked after model fitting, because it influences the reliability of the Monte Carlo estimates. Common quantitative indicators of the sampling efficiency are $\hat{R}$ statistic and effective sample size. $\hat{R}$ statistic is evaluated for convergence diagnostic. $\hat{R}$ is computed using multiple Markov chains with different initial values. It is recommended that only using the sample for Monte Carlo integration, if $\hat{R}$ is less than 1.01 [29]. If it is larger than 1, the chains explored different regions of the distribution. Each chain did not manage to explore the entire distribution. Note that the draws prior to convergence (in the warm-up phase) should be removed.

On the other hand, the effective sample size (ESS) is used to measure the autocorrelation of the sample. ESS is the number of independent sampled values that contains same amount of information or estimation power as the entire set of sampled values, which are autocorrelated.

3.2.2 Large-Sample Properties of Bayesian Approach

As in the frequentist inference, Bayesian inference enjoys large-sample properties - asymptotic normality and consistency. Bayesian inference is inherently consistent, in the sense the posterior probability that θ falls in a neighbourhood of the true parameter value, θ_0 tends to be 1 asymptotically, provided that the model $p(y|\theta)$ is the true data-generating process, $f(y)$ and the prior probability that θ equals the true parameter value, θ_0 or falls in a neighbourhood of θ_0 is not zero. The posterior distribution becomes concentrated in the neighbourhood of θ_0 with probability of 1.

Suppose that θ_0 is a parameter value that uniquely minimizes Kullback-Leibler divergence measure of some assumed data distribution, $p(y|\theta)$ relative to the true distribution, $f(y)$,

$$KL(f \to p) = E_f\left[\log \frac{f(y)}{p(y|\theta)}\right] = \int f(y) \log \frac{f(y)}{p(y|\theta)} dy. \tag{3.3}$$

If the assumed model is correct, that is, $f(y) = p(y|\theta)$, then θ_0 is the true parameter value. Otherwise, the minimizer, θ_0 is the parameter value that makes $p(y|\theta)$ closest to the true distribution.

Suppose $y_1, y_2, \cdots, y_n$ are identically and independently distributed random variables and the parameter space, Θ is discrete. The convergence of a posterior distribution to a value can be proved in terms of log posterior odds,

$$\log \frac{p(\theta_0|y)}{p(\theta|y)} = \log \frac{p(\theta_0)}{p(\theta)} + \sum_{i=1}^{n} \log \frac{p(y_i|\theta_0)}{p(y_i|\theta)} \tag{3.4}$$

For fixed $\theta \neq \theta_0$, the second term on the right of the equation is the sum of n independently and identically distributed random variables. By the strong law of large numbers, the summation tends to $nE_f\left(\log \frac{p(y|\theta_0)}{p(y|\theta)}\right)$. The expectation is equal to $KL(f \to p(y|\theta)) - KL(f \to p(y|\theta_0))$ and is 0 when $\theta = \theta_0$, and larger than 0, otherwise.

Provided that the prior probability, $p(\theta_0) \neq 0$, $E_f\left(\log \frac{p(y|\theta_0)}{p(y|\theta)}\right) \to \infty$, as $n \to \infty$. If the log prior odds in Equation 3.4 is finite, then $\log \frac{p(\theta_0|y)}{p(\theta|y)} \to \infty$, so does $\frac{p(\theta_0|y)}{p(\theta|y)}$. This shows that $p(\theta|y)$ is infinitely higher at $\theta = \theta_0$ than anywhere else. Therefore, the probability mass becomes concentrated at θ_0, that is, $p(\theta_0) \to 1$ as $n \to \infty$. This shows the convergence to θ_0, considering the discrete parameter space. The proof can be extended to continuous parameter space by considering that each value, θ, in the parameter space, has a small neighborhood, A_θ with θ as the center. The convergence to θ_0 is achieved if $p(A_{theta_0}) \to 1$ and $p(A_\theta) \to 0$, for $\theta \neq \theta_0$.

The property of consistency is achieved when the assumed data model is specified correct. It means that the posterior distribution converges asymptotically to the true parameter value.

On the other hand, the limiting posterior distribution is a normal distribution with mean, θ_0. The asymptotically normality can be proved via taking Taylor expansion on the logarithm of the posterior density, around the posterior mode, θ_0.

(3.5)

$$
\begin{aligned}
\log p(\theta|y) &= \log p(\theta_0) - (\theta - \theta_0) \log \left[\frac{d}{d\theta} \log p(\theta|y)\right]_{\theta=\theta_0} + \frac{(\theta-\theta_0)^2}{2} \log \left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right]_{\theta=\theta_0} + R \\
&= \log p(\theta_0) + \frac{(\theta-\theta_0)^2}{2} \log \left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right]_{\theta=\theta_0} + R \\
&= \log p(\theta_0) + \frac{(\theta-\theta_0)^2}{2} \log \left[\frac{d^2}{d\theta^2} \log p(\theta)\right]_{\theta=\theta_0} + \frac{(\theta-\theta_0)^2}{2} \sum_{i=1}^{n} \log \left[\frac{d^2}{d\theta^2} \log p(y_i|\theta)\right]_{\theta=\theta_0} + R \\
&\to constant + \frac{(\theta-\theta_0)^2}{2} nE \left[\frac{d^2}{d\theta^2} \log p(y|\theta)\right]_{\theta=\theta_0} + R \\
&= constant - \frac{(\theta-\theta_0)^2}{2} \frac{1}{\frac{-1}{nE\left[\frac{d^2}{d\theta^2} \log p(y|\theta)\right]_{\theta=\theta_0}}} + R \\
&= constant - \frac{(\theta-\theta_0)^2}{2v},
\end{aligned}
$$

where R contains higher order terms of the expansion. The linear term in the first line of Equation 3.5 is 0 because $\log p(\theta|y)$ peaks at the posterior mode, θ_0. By taking

exponentiation on the last line in Equation 3.5, it is equal to a normal density, with mean, θ_0 and variance, v. Following the proof for the convergence of the posterior distribution, as $n \to \infty$, the support of the posterior distribution for θ becomes very small. The absolute deviation, $|\theta - \theta_0|$, for $\theta \neq \theta_0$ is small and higher orders of the deviation in R tends to 0 as $n \to \infty$. This results that the normality approximation becomes accurate. Thus,

$$\theta \sim_{asym} N\left(\theta_0, v = \frac{-1}{nE\left[\frac{d^2}{d\theta^2}\theta=\theta_0\right]}\right). \tag{3.6}$$

3.3 Model Comparison

A model is robust if it is able to capture the general pattern of a data-generating process in the presence of other sources of error, apart from the sampling error induced by sampling from the population. This implies its capability of generalizing its estimation to any data set generated by the same data-generating process. Predictive accuracy of a model is an indicator of this capability. The rationale of using the predictive accuracy is to study the predictive power of a model on new data sets. Mean squared error (MSE) and log predictive density are the choices of the predictive accuracy measure [8]. MSE is the average of squared difference of true data value and its predicted value, across data points. However, the measure performs badly if the true data process is not normally distributed. Thus, log predictive density is used as the measure of predictive density in the thesis. Furthermore, the robustness of models, that are fitted to the same data set can be compared relatively, using predictive accuracy.

3.3.1 Log Predictive Density

The reason that log predictive density is used in statistical model comparison is due to its connection with Kullback-Leibler (KL) divergence. The KL divergence measures the difference in expected loss, $E_y(-\log \text{likelihood})$, between 2 models [25]. The equation below is written in Bayesian context.

$$\begin{aligned} KL(f \to p) &= \int f(y) \log \frac{f(y)}{p_{post}(y)} dy \\ &= E_y\left(\log \frac{f(y)}{p_{post}(y)}\right) \\ &= E_y[\log f(y)] - E_y[\log p_{post}(y)], \end{aligned} \tag{3.7}$$

where $f(.)$ is the true data-generating process, $p(.)$ is an assumed statistical model for the data and $p_{post}(.) = \int p(y|\theta)g(\theta|y)dy$ is the expected posterior predictive density for a

data point, y, over the posterior distribution for θ, $g(\theta|y)$.

$f(\cdot)$ is usually unknown but fixed. Hence, the first term, $E_y[\log f(y)]$, in Equation 3.7 is also fixed. However, the second term, $E_y[\log p_{post}(y)] = \int \log p_{post}(y) f(y)$, is the expected log posterior predictive density (elpd) for a data point, y and cannot be computed directly because of the unknown $f(\cdot)$. It can be estimated based on the observations and $p(\cdot)$. The estimate is denoted by $\widehat{elpd}$. Besides, since $p(\cdot)$ is the assumed model on the data and is up to the choice of modellers, minimizing the KL divergence between $f(y)$ and $p_{post}(y)$ is the same as maximizing $elpd$. This also implies that the model with the highest $\widehat{elpd}$ wins the comparison [17].

3.3.2 Widely-Applicable Information Criterion (WAIC)

There are several ways to measure $elpd$ using either within-sample or out-of sample. However, there are limitations on using either kind of sample. The out-of sample is hardly available, whereas using within-sample to estimate $elpd$ leads to upward biased estimate due to overfitting [8].

The ideal estimation to $elpd$ is Leave-One-Out Cross Validation (LOO-CV) [8, 17]. However, it is computationally intensive. Thus, in the thesis, the estimation is settled for Widely Available Information Criterion (WAIC)[33], which is one of the measures strongly recommended by Gelman et al [8]. There are several reasons to opt for this measure. Its computational time is much shorter than that of LOO-CV and its approximation quality is much better than Deviance Information Criterion (DIC) and Akaike Information Criterion (AIC). The reasons can be further explained through the derivation of WAIC, as stated below.

$$\begin{aligned} \hat{elppd} &= \sum_{i=1}^{n} lppd_i - p_{WAIC} \\ &= \sum_{i=1}^{n} \log[E_{post} p(y_i|\theta)] - \sum_{i=1}^{n} var_{post}(\log p(y_i|\theta)), \end{aligned} \tag{3.8}$$

where $elppd = \sum_{i=1}^{n} E_f(\log p_{post}(y_i))$ is the expected log pointwise predictive density for a data set. The idea of $elppd$ is, by considering partitioning the data set into n individual data points and taken one at a time, then the expected log posterior predictive density for a data set becomes the sum of individual lppd. The idea is same as LOO-CV, which also divides data set into n points, but it saves large amount of computational time [8].

The first term, $\sum_{i=1}^{n} lppd_i$ in Equation 3.8, which represents the estimate of predictive accuracy, take into account the posterior uncertainty about the parameter, θ, $p(\theta|\boldsymbol{y})$, in the calculation. This is the feature that distinguish WAIC from AIC and DIC, which both use log predictive density given that θ is MLE and the mean of the posterior distribution

for θ, respectively. The second term, p_{WAIC} is the bias correction for the first term in order to offset the issue of overfitting due to using the same data. It is the variance of $\log p(y_i|\theta)$ over the posterior uncertainty about θ. This, thus, indicates that the higher the posterior uncertainty about the parameter value, the heavier the penalty applied on the predictive accuracy estimate. The inclusion of posterior uncertainty in both terms makes $\hat{elppd}$ a fully Bayesian estimate and improves its approximation quality for the out-of-sample predictive accuracy of a model [8].

To be comparable with other information criteria. the scale of $\hat{elppd}$ has to be adjusted to the deviance scale by multiplying $\hat{elppd}$ by a factor of -2.

$$WAIC = -2 \times \hat{elppd} \tag{3.9}$$

Chapter 4

Robust Inference

This chapter begins with the motivation to robust inferences (Section 4.1). It explains the role of likelihoods in a robust model (Section 4.2). It ends with the development of a robustified likelihood (Section 4.3).

4.1 Motivation

A statistical model simplifies a real-world problem (data-generating process). Since the underlying process, including the observation process, to a dataset is unknown, we inevitably encounter the inadequacy of model descriptions with available factors and inconsistency of the model assumptions with the dataset. Usually, the issues can be diagnosed by investigating the residual variation after model fitting. Extra variability in the data set is one of the symptoms showing model inadequacy. (Section Experiment illustrates 2 particular sources of extra variability.) Particularly, observations lie far away from a bulk of observations. Such observations are called outliers. If the outliers follow the general pattern of the rest, they help extrapolation in analyses. Otherwise, they blur the true data pattern and bias the inferences. There are multiple ways to view the latter case.

It can be thought as a conflict within the dataset, such that there exist 2 distributions – a main distribution and an outlier distribution. The main distribution generates the majority of the observations in a dataset and the outlier distribution generates atypical observations. This perspective makes outlier removal appealing. However, there is an argument that, without reasonable domain-related justification, outliers should not be simply removed before analysis [22]. Neyman and Scott [22] also stated that outliers are not uncommon in some domains of study and outlier elimination amounts to mutilation of the datasets.

Instead, the dataset can be thought of being generated from a relatively heavy-tailed

distribution (or an error-prone distribution) instead of a standard distribution. The data variation is assumed to be much dispersed. The standard models we usually utilize, such as normal, poisson and binomial distributions, are convenient but less flexible when dealing with extra-variability. They provide optimal estimation only when their assumptions are consistent with the datasets. Otherwise, the inference may be misleading and biased. This leads to the topic of model robustness. Both frequentist and Bayesian approaches have the same definition on a robust statistical model. A model is robust if it can provide reliable inferences with or without outliers. Put differently, it can detect outliers and downweight their influence in inferences. There is also a different definition on model robustness. A model is robust if it is less sensitive to the change of model assumptions [8]. The inferences do not vary too much when components in the model vary. The former definition is the interest of the thesis.

4.2 The role of the Likelihood in a Robust Model

A statistical model is composed of a systematic part, that explains a portion of variability in the data set in terms of factors, and a random part, that describes the unexplained variation with a distribution. It is obvious to see that these 2 parts are complementary, in the sense that one part 'takes over' the left-over variation that cannot be 'handled' by the other. The combination of these 2 components is expected to explain adequately the data variability. However, if one part fails to explains the portion that it is expected to do it well, the portion will be taken over by the other part and this results in an unreliable inference. For example, the random variability is much larger than expected by a specified distribution(random component). The systematic part takes over and try to explain this left-over random variability with covariates. This leads to biased estimates on the effect size of the covariates. Thus, it is important to adequately describe the variability in a data set, in order to capture the true effect size of the factors on the response. Usually, an overdispersed alternative to the standard distributions can be adopted, when dealing with data overdispersion. For example, a Student's t-distribution is an alternative to the normal distribution, whereas the Negative Binomial distribution to the Poisson distribution. This section shows that the Negative Binomial distribution is not a completely robust distribution or outlier-prone distribution [23, 26] and does not lead to a robust Bayesian model.

In the Bayesian framework, West [35, 36] and Ramsay and Novick [27] explained how the choice of the likelihood affects the posterior distribution. By differentiating the log posterior with respect to μ, the posterior score is

$$\frac{d}{d\mu}\log p(\mu|\boldsymbol{y}) = \frac{d}{d\mu}\log p(\mu) + \sum_{i=1}^{n}\frac{d}{d\mu}\log p(y_i|\mu), \tag{4.1}$$

where $\frac{d}{d\mu}\log p(y_i|\mu)$ is called the likelihood score with the respect to the parameter, μ. We can see how the terrain/steepness (or mathematically, score) of the log posterior is determined by the terrains of the log prior and n individual and identical log-likelihoods, via addition. Any individual information source whose score is extremely positively high or negatively low, at a parameter value, μ, can dominate the terrain of the posterior over the rest of sources whose score functions are finite. This illustrates that any individual log support, whose slope is extremely steeper than the rest, at μ, is more influential in determining the behavior of the posterior distribution at that parameter value. Hereafter, the score is also called the influence of an information source on the posterior.

It is desirable if the dominating information sources are reliable. Then, the terrain of the posterior distribution across the parameter space are built on them. However, suppose that there exists an outlier in the data set, we may not want the posterior distribution to be thoroughly dominated by it. Preferably, the influence of the outlier on the posterior terrain is limited or goes to 0 in parameter regions whose μ are far from the outlier value. The score that is limited across the parameter space, is called a bounded score, whereas the score is redescending if it goes to 0 when the parameter values, μ are far from the outlier value.

By expressing $y_i = \mu + \epsilon_i$ in terms of its error terms, such that $p(y_i|\mu) = p(\epsilon_i|\mu)$, the last term on the right side of Equation 4.1, $\frac{d}{d\mu}\log p(y_i|\mu)$ can be re-expressed as

$$\begin{aligned}\frac{d}{d\mu}\log p(y_i|\mu) &= \frac{d}{d\mu}\log p(\epsilon_i|\mu)\\ &= \frac{d}{d\epsilon_i}\log p(\epsilon_i|\mu)\cdot\frac{d}{d\mu}\epsilon_i\\ &= \frac{d}{d\epsilon_i}\log p(\epsilon_i|\mu)\cdot\frac{d}{d\mu}(y_i-\mu)\\ &= -\frac{d}{d\epsilon_i}\log p(\epsilon_i|\mu).\end{aligned} \tag{4.2}$$

Therefore, Equation 4.1 can be rewritten as

$$\frac{d}{d\mu}\log p(\mu|\boldsymbol{y}) = \frac{d}{d\mu}\log p(\mu) - \sum_{i=1}^{n}\frac{d}{d\epsilon_i}\log p(\epsilon_i|\mu). \tag{4.3}$$

Since $p(y_i|\mu) = p(\epsilon_i|\mu)$, the derivation in Equation 4.2 shows that the likelihood score, $\frac{d}{d\mu}\log p(y_i|\mu)$, corresponds to the negative gradient of $\log p(y_i|\mu)$ with respect to y_i, $-\frac{d}{dy_i}\log p(y_i|\mu)$. This implies that the choice of distribution, $p(y_i|\mu)$, plays an important role in determining the behavior of the posterior distribution [23, 26].

I consider unimodal distributions hereafter. The distribution whose $\log p(y|\mu)$ is getting steeper as y moves farther from μ, is defined as a light-tailed distribution/outlier-resistant distribution. It has an unbounded score across the parameter space (Equation

4.2). For example, exponential dispersion distributions. In contrast, for a heavy-tailed distribution, there exists a threshold distance of y from μ, denoted by d_{thres} such that the distribution starts to become less steep as y hits and go beyond d_{thres}. Thus, the heavy-tailed distribution/outlier-prone distribution has bounded and redescending score function (Equation 4.2) because the steepness stops increasing and starts decreasing after some point. This also means that an observation does not influence the posterior distribution across the entire parameter space. Such distributions are useful when dealing with suspiciously unrealiable information sources, so as to limit their influential area of the parameter space when constructing the posterior distribution.

The likelihood score function, $\frac{d}{d\mu}\log p(y_i|\mu)$, can be written in the form [36]

$$\frac{d}{d\mu}\log p(y|\mu) = \frac{d}{d\mu}\log p(\epsilon|\mu) = \omega(y-\mu)\cdot(y-\mu), \tag{4.4}$$

where $\omega(y-\mu)$ is the weight function of $y-\mu$. In order to have a bounded and redescending score function, it requires $\omega(y-\mu)$ to decay faster than $\frac{1}{|y-\mu|}$ as $|y-\mu|$ increases [36]. Otherwise, the score function is unbounded as $|y-\mu|$ increases. Hence, the weight function plays a role of adjusting the influence of y on the posterior distribution of μ.

West [35] showed that EDF distributions are outlier-resistant, using the definitions in O'Hagan's paper [23, 26]. An intuitive way to show that if a distribution is robust is by checking visually

(a) if its $\omega(y-\mu)$, decays faster than $\frac{1}{|y-\mu|}$ as $|y-\mu|$ increases, for $|y-\mu| \geq d_{thres}$;

(b) if the likelihood score function decreases as $|y-\mu|$ increases, for $|y-\mu| \geq d_{thres}$.

The log-likelihood of an EDF distribution is

$$\log L(\theta, \phi; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi), \tag{4.5}$$

where θ and ϕ are the canonical and scale parameters and $a(.), b(.)$ and $c(.)$ are some functions corresponding to the distribution (See Section 2.1). Its first-order derivative with respect to θ is then

$$\frac{d\log L(\theta, \phi; y)}{d\theta} = \frac{y - b'(\theta)}{a(\phi)} = \frac{y-\mu}{a(\phi)}, \tag{4.6}$$

where $\mu = E(y)$. Since the canonical parameter is not always of the interest, as compared

to μ, we can use the chain rule to derive the score function with respect to μ,

$$\begin{aligned} \frac{d \log L(\theta, \phi; y)}{d\mu} &= \frac{dL(\theta, \phi; y)}{d\theta} \cdot \frac{d\theta}{d\mu} \\ &= \frac{y - \mu}{a(\phi)} \cdot \frac{a(\phi)}{var(y)} \\ &= \frac{y - \mu}{var(y)}, \end{aligned} \tag{4.7}$$

where μ can be expressed in terms of θ, via the inverse link function and the weight function, $\omega(y - \mu)$ is the inverse of $var(y)$. Consider that the score function is a function of y, given a fixed parameter value, μ. Since the weight, $1/var(y)$, is always the same over y, the absolute score of y increases as $|y - \mu|$ increases. Observations that are far from μ will have larger influence on the posterior at μ than observation closer to μ. Thus, the EDF likelihood is not robust.

For the Poisson likelihood, whose $a(\phi) = 1$ and $var(y) = \mu$, its score function of μ is

$$\frac{d \log L(\theta, \phi; y)}{d\mu} = \frac{1}{\mu} \cdot (y - \mu), \tag{4.8}$$

where the weight function, $\omega = \frac{1}{\mu}$. For the Negative Binomial likelihood, whose $a(\phi) = 1$ and $\mu = \mu(1 + \mu/k)$, where k is the dispersion parameter,

$$\frac{d \log L(\theta, \phi; y)}{d\mu} = \frac{k}{\mu(\mu + k)} \cdot (y - \mu), \tag{4.9}$$

where $\omega = \frac{k}{\mu(\mu+k)}$. We can see that all y values are given equal weight at any μ, regardless of how atypical the values are, given μ, for both likelihood models. (See Appendix 1 for the weight and score plots of the Poisson and Negative Binomial likelihoods.)

In order to achieve outlier rejection, the EDF likelihoods should either be abandoned or modified to accommodate the outliers. In case of a Normal observation model, which is also an EDF distribution, it is robustified by assigning an inverse Chi-square distribution to its variance and integrating it with respect to the variance. The resultant distribution is a scale-mixture of Normal distributions, known as the Student-t distribution, whose score is bounded and redescending [36, 35]. However, for the Poisson model, there are no robust alternatives that correspond to a well-defined sampling distribution for y. Then, West [35] proposed the generalization of the Normal robustification approach to any EDF distribution. The proposal involves the quasi-likelihood approach [34], in which an approximate likelihood is constructed based on the specification of the mean-variance relationship of y. This leads us to the next section – robust quasi-likelihood.

4.3 Robust Quasi-Likelihood

This section introduces robust quasi-likelihoods, proposed by West[35]. The idea is to derive an approximate sampling distribution, based on a scaled EDF score function, whose scale parameter is a random quantity, and marginalize the quasi-likelihood with respect to the scale parameter.

4.3.1 Quasi-Likelihood

To generalize the Normal robustification to any EDF distribution, it requires the variance of the distribution to be a random quantity, while keeping the mean-variance relationship of the distribution, $V(\mu)$ that characterizes the distribution. To allow for this, a scale parameter, $\Phi > 0$ is introduced to scale the variance of the EDF distribution, $var(y)$. The scaled variance is expressed as

$$\begin{aligned} var^*(y) &= \Phi\ V(\mu) \\ &= \Phi\ a(\phi)\ V(\mu), \end{aligned} \tag{4.10}$$

where $a(\phi) = 1$ for the Poisson distribution. Following the derivation in Equation 4.10, $var^*(y)$ can be treated as a random quantity without altering the EDF mean-variance relationship, when Φ is random.

Concisely, the main idea is, thus, to construct a sampling distribution with μ and $\Phi var(y)$, that shares the similar mean-variance relationship with the EDF distribution. This idea extends to the topic of quasi-likelihood [34] in which an approximate likelihood of μ is constructed based on the mean-variance relationship.

A quasi-likelihood is usually adopted when modellers apply minimal assumptions on the underlying data-generating process [34]. It does not require the specification of the data distribution. It only requires the specification of the mean, μ and variance, $var^*(y)$ of the observations, where $var(y)$ is expressed in terms of the mean-variance relationship, $V(\mu)$.

Suppose the observation, y has a distribution whose mean, μ and variance, $var^*(y)$. A quasi-likelihood of μ, given y can be constructed based on μ and $var^*(y)$ and is expressed as

$$\int_y^\mu \frac{y-\mu}{var^*(y)} d\mu. \tag{4.11}$$

Its first-order derivation with respect to μ, known as the quasi-score is

$$\frac{y-\mu}{var^*(y)}. \tag{4.12}$$

Suppose $var^*(y) = \Phi\ var(y)$, where $var(y)$ is a function of $V(\mu)$ of an EDF distribution. The quasi-likelihood for μ becomes

$$\int_y^\mu \frac{y-\mu}{\Phi\ var(y)} d\mu \tag{4.13}$$

and the quasi-score with respect to μ is

$$\frac{y-\mu}{\Phi\ var(y)}. \tag{4.14}$$

It differs from the GLM score (Equation 4.7) with the addition of Φ. Its weight function, $\omega(y-\mu)$ is $\frac{1}{\Phi var(y)}$. Apparently, the weight does not decrease to zero as $|y-\mu|$ increases, for any fixed value of Φ.

The quasi-likelihood is in an integral form and does not have an explicit probability distribution form. The approach was invented in the frequentist framework in which the first derivative of a quasi-EDF likelihood, $\frac{y-\mu}{\Phi var(y)}$ is required to estimate μ. This makes the quasi-likelihood approach rather unattractive in the Bayesian framework, which requires a proper expression of the likelihood of μ. Nevertheless, West [35] adopted the quasi-likelihood approach in the Bayesian framework by developing an approximate sampling model based on the quasi-score (Equation 4.14), following his previous work [36], in which he proposed a robustified Normal likelihood.

Suppose that $y \sim EDF(\mu, \Phi var(y))$ is an EDF distribution with the mean, μ and the scaled variance, $var(y)^* = \Phi var(y))$. However, not all EDF distribution can be expressed in terms of Φ, such as Binomial and Poisson models, because the variance is a function of mean (Equation 2.3). Thus, the distribution of an EDF random variable with a scaled variance cannot be derived easily. In contrast, the Normal distribution has its variance independent of the mean, which makes the introduction of the extra scale parameter, Φ easier. The distribution of a Normal random variable with a scaled variance can be expressed in terms of Φ properly, without difficulty.

The density of $y \sim N(\mu, 1)$ is proportional to

$$\exp\left(-\frac{(y-\mu)^2}{2}\right) = \exp\left(-\frac{D(y;\mu)}{2}\right), \tag{4.15}$$

where $D(y;\mu) = (y-\mu)^2$ is the deviance of the normal distribution, while the density of $y \sim N(\mu, \Phi var(y))$ is proportional to

$$\begin{aligned} &\frac{1}{\sqrt{\Phi var(y)}} \exp\left(-\frac{(y-\mu)^2}{2\Phi var(y)}\right) \\ &= \frac{1}{\sqrt{\Phi var(y)}} \exp\left(-\frac{D(y;\mu)}{2\Phi var(y)}\right). \end{aligned} \tag{4.16}$$

West generalized Equation 4.15 and Equation 4.16 to EDF distributions by writing the distribution of the EDF random variable, $y \sim EDF(\mu, 1)$, in terms of deviance, like Equation 4.15 and deriving the approximate density of random variable for $y \sim EDF(\mu, \Phi var(y))$, using Equation 4.16. Following the expression of the deviance of a probability distribution,

$$D(y;\mu) = -2\log\frac{L(\mu;y)}{L(y;y)}, \tag{4.17}$$

a probability distribution can be expressed as

$$\begin{aligned} p(y;\mu) &= L(y;y)\cdot\exp\left(-\frac{D(y;\mu)}{2}\right) \\ &\propto \exp\left(-\frac{D(y;\mu)}{2}\right) \end{aligned} \tag{4.18}$$

which is in the same functional form as Equation 4.15. Then the approximate sampling distribution of $y \sim EDF(\mu, \Phi var(y))$ or the quasi-likelihood of μ can be derived, using Equation 4.16, as

$$m(y;\mu,\Phi) = \frac{1}{\sqrt{\Phi var(y)}}\exp\left(-\frac{D(y;\mu)}{2\Phi var(y)}\right). \tag{4.19}$$

4.3.2 Scale Mixture of Quasi-Likelihoods

Remember that quasi-likelihoods are not robust for a fixed value of Φ. West [35] suggested to robustify the approximate scaled likelihood, $m(y;\theta,\Phi)$ (Equation 4.19) by setting Φ to be a random quantity and integrating $m(y;\mu,\Phi)$ over Φ with some prior. The resultant likelihood is a scale-mixture of approximate scaled likelihoods. He derived an analytical expression for the scale mixture of the likelihoods (namely, the robust quasi-Poisson likelihood hereafter), by having $k \cdot var^*(y) \sim \chi^2_k$,

$$\begin{aligned} m(y;\mu) &= \int_0^\infty m(y;\mu,\Phi)\cdot p(\Phi)\ d\Phi \\ &\propto [k + D(y;\mu)]^{-\frac{k+1}{2}}, \end{aligned} \tag{4.20}$$

where k is the degree of freedom of the Chi-square distribution. Further, its score function, with respect to μ is

$$\frac{d}{d\mu}\log m(y;\mu) = -\frac{k+1}{2(k+D(y;\mu))}\frac{d}{d\mu}(k+D(y;\mu)) = \frac{k+1}{var(y)\cdot(k+D(y;\mu))}(y-\mu). \tag{4.21}$$

Following Equation 4.4, the weight function, $\omega(y-\mu)$ is $\frac{k+1}{var(y)\cdot(k+D(y;\mu))}$. It approaches to 0 when $|y-\mu|$ increases because the deviance, $D(y;\mu) \geq 0$, increases when μ moves away from y. Thus, Equation 4.21 is bounded and redescending. For the Poisson distribution, the deviance, $D(y;\mu) = 2(y\log\frac{y}{\mu} - y + \mu)$ and its robust quasi-Poisson likelihood is proportional to

$$[k + 2(y\log\frac{y}{\mu} - y + \mu)]^{-\frac{k+1}{2}}. \tag{4.22}$$

(See Appendix 1 for the weight and score plots of the robust quasi-Poisson likelihoods.)

Conclusively, the Poisson distribution is outlier resistant and not robust against outliers. The Negative Binomial distribution, which has been suggested as an overdispersed alternative for the Poisson, is not robust either against the outliers. In contrast, the robust quasi-Poisson likelihood is an approach, which fulfills the formal requirements of robust likelihood even though it does not correspond to any known distribution.

Chapter 5

Experiments

This chapter begins with a single-observation-based experiment on the model robustness between 3 Bayesian models with different likelihoods (Section 5.1). It is followed by the motivations to the simulation experiments (Section 5.2). The setups of 2 simulation experiments - measurement/observation error and process error and the Monte Carlo computations for the posterior summary statistics are presented in Section 5.3.

5.1 An Introductory Experiment

This section illustrates the robustness of 3 Bayesian models against an outlier. The experiment illustrates the extreme case in which the prior is at the end of the parameter range and very close to zero and the outlier is very far from the prior. The Bayesian models share the same prior but have different likelihoods for the mean – Poisson, Negative Binomial and robust quasi-Poisson likelihoods.

Suppose that the prior is reliably informative

$$\lambda \sim N_+(3, 1) \tag{5.1}$$

and a single observation is an outlier, $y = 20$, that is in conflict with the prior. The expectation is that the posterior rejects the data source and converges to the reliable

prior. The likelihoods that were compared are

$$
\begin{aligned}
&\textbf{Poisson:} \\
&L(\lambda; y = 20) \propto \exp(-\lambda) \cdot \lambda^{20} \\
&\textbf{Negative Binomial:} \\
&L(\lambda; y = 20, k_{nb}) \propto \frac{\lambda}{\lambda + k_{nb}}^{y} \cdot \frac{k_{nb}}{\lambda + k_{nb}} \\
&\textbf{Robust Quasi-Poisson:} \\
&L(\lambda; y = 20, k_q) \propto [k_q + D(y = 20; \lambda)]^{-\frac{k_q+1}{2}},
\end{aligned}
\tag{5.2}
$$

where k_{nb} is the dispersion parameter of the Negative Binomial distribution and k_q is the degrees of freedom of the χ^2_k distribution. The Negative Binomial and robust quasi-Poisson likelihood were tested with 2 different values of $k_{nb} = k_q = \{1, 5\}$. The weight and score functions of the aforementioned likelihoods are shown in the Appendix 1.

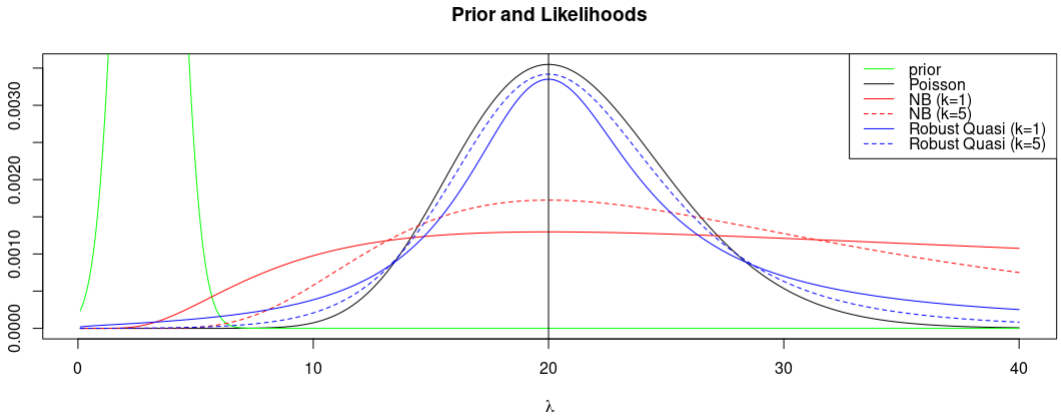

Figure 5.1: A normal prior, $N_+(\lambda|3, 1)$ and different likelihoods, given $y = 20$.

From Figure 5.1, we can see that the Poisson likelihood given the outlier covers the smallest range of λ, as compared to the rest. The Negative Binomial likelihood ($k = 1$) supports a wide range of values but does not support values close to 0, while ($k = 5$) covers relatively small range of λ. The Negative Binomial distribution has a heavy tail on one side only. It assumes that λ close to 0 are less likely to be true parameter. On the other hand, the quasi-likelihood ($k = 1$) supports λ that are either extremely less than or greater than the outlier. It does not avoid the possibility that the true λ lies from the observation on either side.

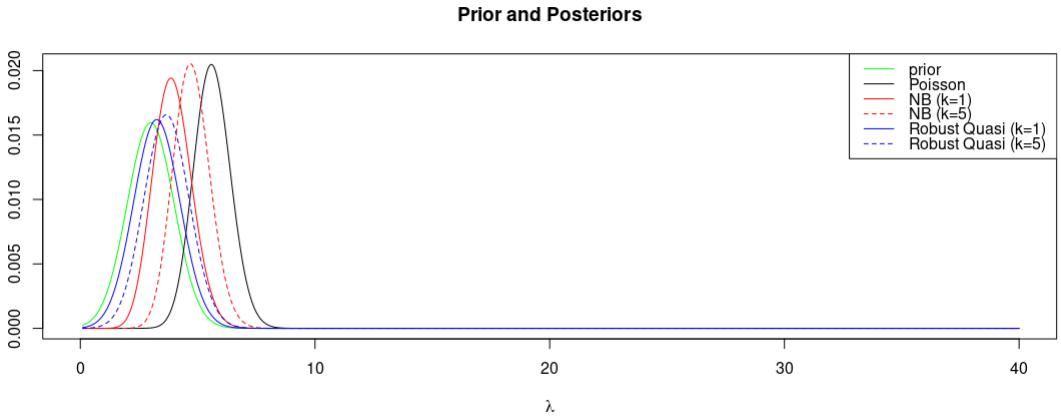

Figure 5.2: A normal prior, $N_+(\lambda|3,1)$ and posteriors with different likelihoods, $y = 20$.

We can see that from Figure 5.2, the posterior distribution with the robust quasi-Poisson likelihood ($k = 1$) lies closest to the prior, while the posteriors with the Poisson likelihood lies further from the rest. The comparison illustrates that the Bayesian model based on the robust quasi-Poisson likelihood is a relatively robust model as compared to the rest.

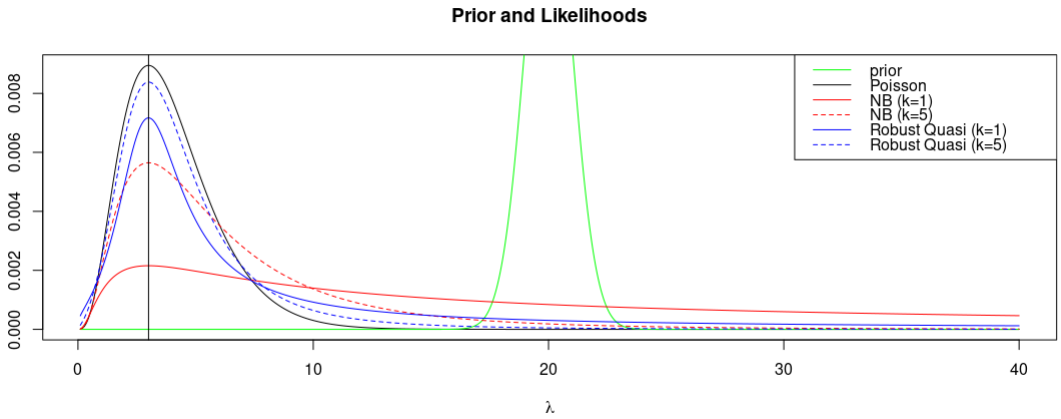

Figure 5.3: A normal prior, $N_+(\lambda|20,1)$ and different likelihoods, given $y = 3$.

In contrast, if y is 3 and the prior is on the right side of y, $N_+(\lambda|20,1)$ (See Figure 5.3). All other posteriors converge to the normal prior, except the Poisson-based posterior (Figure 5.4). In overall performance, the robust quasi-Poisson does better than the rest.

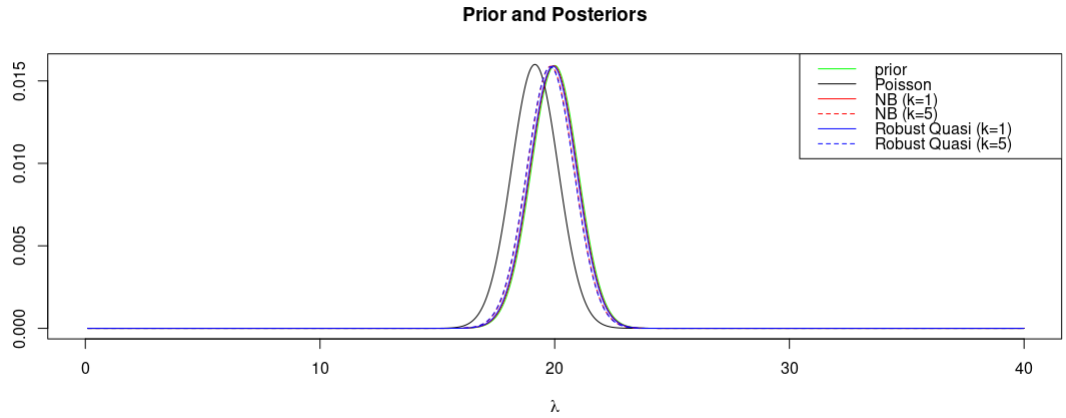

Figure 5.4: A normal prior, $N_+(\lambda|20, 1)$ and posteriors with different likelihoods, given $y = 3$.

The takeaway of this section is that a Bayesian model can adopt a heavy-tailed likelihood to detect unreliable sources, reduce their influences and move to reliable sources. The unreliable sources can be outliers or priors. The next section considers the overdispersed data points as unreliable sources and the rest of the data points are reliable sources. The section also illustrates that the percentage of unreliable sources can also determine the model robustness.

5.2 Motivation

We usually assume that there exists a data-generating process that generates the variable of interest, namely the response. There are a set of factors that are 'operating' within the process such that the response value is determined by the combinations of these factors' values and the relationship between the factors and the response. These set of factors are unknown and researchers have to discover or hypothesize the factors and the relationship between the factors and the response. Usually, only a subset of the factors are identified. Sometimes, researchers select deliberately factors of interest only, rather than considering the whole set. Suppose that the factors and response of each unit are measured correctly. The researchers build a model with this subset of factors. The residual variance/error, thus, is due to the process error from the excluded/unknown factors and sampling error. The variation due to unknown influences are termed the process error. Ignoring the presence of unknown influences might lead the researchers to build a model with smaller variance. This results in overdispersion in the dataset. The presence of process error is

the theme of Experiment 2. The topic has obtained much attention in many domains. There are numerous literature works regarding to it [9, 12, 13, 18]. In the experiment, random process error was considered rather than an excluded fixed factor and the model capability was tested in different combinations of process stability (the presence of the process error) and process error variance.

Furthermore, overdispersion worsens when the sample is measured or observed imperfectly during the observation process. This results in observation error. The true data set becomes latent. If a model is build based on the subset of true factors and sampling error, the residual error is a combination of the observation error and process error. The true amount of variation in the population is larger than the the variance expected by the model. The presence of observation error is the theme of Experiment 1. In the experiment, only the covariate was subject to the measurement error. The true covariate was a latent variable. The model capability was tested in different combinations of measurement reliability and measurement error variability [6, 10, 11].

Each response type has its corresponding basic model to start with due its simplicity. It is Poisson for count response. However, basic models, in practice, are relatively unrealistic when considering multiple error sources are possible in a dataset. The goal of the thesis is to compare a basic count model, Poisson with its extended models, Negative Binomial and robust quasi-Poisson model, in the presence of process or observation error.

5.3 Simulation Studies

Frequentist-type simulation setup was used in the thesis. Frequentist setup means that data sets are sampled repeatedly from a fixed data-generating process [3].

5.3.1 Generating Simulated Data Sets

CASE A - Covariate measurement error

In practice, an observed value is a function of its true value and some measurement error. If the measurement process is perfect, the observed value is same as its true value. Otherwise, the measurement is error-prone. In the case of a data set, the observed data set is not a true reflection of the true status of the sample, and of the target population, even if the units are sampled randomly from the population. While measurement errors could arise in either response or covariates, the experiment in the thesis focuses on measurement error in a single covariate. Measurement error in a covariate results in biased and inconsistent estimates, because the estimations are generally based on the assumption that the covariates are measured without error [6].

In this experiment, consider n independent observations (x_i, y_i), for $i = 1, \cdots, n$ that follow the distribution, $y_i \sim Pois(\exp(a + bx_i))$, where a and b are the intercept and slope parameters. Assume that the occurrence of measurement error is probabilistic, in a way that units are "picked" randomly and independently, to be measured erroneously. There are 3 pre-defined probabilities of the covariate of an observation being measured erroneously and 3 degrees of variability due to measurement error. The variability due to measurement error indicates the average deviation of an observed value from its true value. Overall, there are 9 settings in which the robustness of the models are tested.

50 simulations were performed in each setting. In each simulation, a dataset of size $n = 500$ were generated. x_i were drawn from a uniform distribution with the minimum as 1 and the maximum as 2. y_i were drawn from a Poisson distribution, $Pois(\exp(2 + 2x_i))$, where both the slope and intercept parameters equal to 2. Each observation, i, had an equal probability, $p \in \{0.05, 0.2, 0.6, 1\}$ that x_i was measured with error. If the unit was "picked", its true value, x_i was replaced with

$$x_{obs_i} = x_i + \epsilon_i, \tag{5.3}$$

where the measurement error, $\epsilon_i \sim N(0, \sigma^2_{err})$ and $\sigma^2_{err} \in \{0.1^2, 0.3^2, 0.5^2\}$. Otherwise, the observed value was the same as its true value, $x_{obs_i} = x_i$.

CASE B - Process error

The data-generating process of a data set is much sophisticated than the assumed model with a limited set of covariates. A subset of truly influential factors is considered in the model, whereas its complement is either unidentified or not of interest for the researchers and thus not included in the model. The variation that can be explained by the omitted variables is named the process error variance. The process error contributes to a portion of the random variability that is modelled with a probabilistic distribution. Neglecting the process error may lead to specifying a distribution that expects smaller random variation.

In this experiment, I assumed that the occurrence of process error for an observation (or the presence of unknown factor(s) in an observation) is probabilistic and the effect size is random. There are 3 predefined probabilities of error occurrence and 3 degrees of variability of the process error (effect size). Thus, there are 9 different scenarios in which the robustness of the models are tested.

50 simulations were performed in each setting. A data set of size $n = 500$ was generated per each simulation. x_i, same as in Case A, were generated from a uniform distribution with the minimum as 1 and the maximum as 2. y_i were generated either with the absence,

$$Pois(\exp(a + bx_i)) = Pois(\exp(\lambda_i)), \tag{5.4}$$

or presence of the random effect

$$Pois(\exp(a + bx_i + \epsilon_i)) = Pois(\exp(\lambda_i)\exp(\epsilon_i)), \tag{5.5}$$

with an error probability, $p \in \{0.05, 0.2, 0.6, 1\}$. The process errors, ϵ_i were generated from $N(0, \sigma^2_{err})$ where $\sigma^2_{err} \in \{0.1^2, 0.3^2, 0.5^2\}$ and were assumed to have an additive effect on the the logarithm of the mean. On the original count scale, the unknown factor(s) affected the mean multiplicatively, by a factor of $\exp(\epsilon_i)$. The multiplicative errors can be written as $\log[\exp(\epsilon_i)] \sim N(0, \sigma^2_{err})$. The perturbed data-generating process corresponds to a log-linear mixed-effects Poisson model.

5.3.2 Alternative Models

Poisson, Negative Binomial and robust Quasi-Poisson Bayesian models were fitted to the simulated data sets from both cases. The focus of the thesis emphasizes less on prior specification and more on the likelihood specification. Therefore, the priors specified here are mostly vague priors and bring limited information and less influence to the posterior inference.

Poisson Bayesian Regression Model

The Poisson regression model assumes that the count response, y has independent observations from Poisson distributions, $Pois(\exp(a+bx_i))$, with different means. The likelihood function of the parameters, a and b, given a Poisson observation, (x_i, y_i) is

$$L(a, b|y_i) = \frac{\exp\{-\exp(a + bx_i)\}\{\exp(a + bx_i)\}^{y_i}}{y_i!}. \tag{5.6}$$

The intercept, a and slope, b are the parameters of interest. I assumed independent wide normal priors for both parameters.

$$\begin{aligned} a &\sim N(0, 10^6) \\ b &\sim N(0, 10^6) \end{aligned} \tag{5.7}$$

Negative Binomial Bayesian Regression Model

The Negative Binomial regression model assumes that y has independent observations from Negative Binomial distributions, $NB2(\exp(a + bx_i), k)$, with some fixed dispersion parameter, k but different means. Notice that the link function is $\log \lambda = a + bx_i$. The Negative Binomial likelihood of a, b and k, given an observation, is

$$L(a, b, k|y_i) = \frac{\Gamma(y_i + k)}{\Gamma(k) + \Gamma(y_i + 1)} \left[\frac{\exp(a + bx_i)}{\exp(a + bx_i) + k}\right]^y \left[\frac{k}{\exp(a + bx_i) + k}\right]^k, \tag{5.8}$$

I assumed independence between the parameters,a, b and k. The same normal priors were adopted for a and b. The dispersion parameter, k should be positive. Thus, a Gamma distribution was assigned to define the uncertainty over the parameter space of k.

$$\begin{aligned} a &\sim N(0, 10^6) \\ b &\sim N(0, 10^6) \\ r &\sim Gamma(2, 0.2) \end{aligned} \tag{5.9}$$

Robust Quasi-Poisson Bayesian Regression Model

The robust quasi-likelihood of a and b, given y_i is

$$L(a, b|y_i) = k + 2\left[y_i + \log\left(\frac{y_i}{\exp(a + bx_i)}\right) - y_i + \exp(a + bx_i)\right] \tag{5.10}$$

where k is the number of degrees of freedom. Still, I considered the log link function of λ_i. The observations are assumed to be independent, such that the log quasi-likelihood of a data set is the sum of n individual robust quasi-likelihood.

I assumed independence between a and b, and assigned diffuse normal priors to them.

$$\begin{aligned} a &\sim N(0, 10^6) \\ b &\sim N(0, 10^6) \end{aligned} \tag{5.11}$$

k was set as 5 after tuning. If k was given much smaller value, it took long sampling time and did not converge within a predefined length of chains.

5.3.3 MCMC

The MCMC setup was the same for both models, such that 4 parallel chains were run, with 1000 burn-in samples, 2000 main samples per chain. Further, the step size, denoted by adapt_delta was set 0.99 and the maximum number of steps per each iteration was 20, to ensure no post-warmup divergence during sampling from the posterior distributions and avoid the sampler hitting the maximum treedepth. The sampler algorithm was No-U-Turn sampler [15, 28].

5.3.4 Bayes Estimate and its Sampling Distribution

The posterior distribution for a parameter was summarized with its mean, which is motivated by the asymptotic properties of Bayesian inference (Section 3.2.2). The posterior

mean was regarded as the point estimate of the parameter. The posterior mean will hereafter be denoted by "Bayes estimate". Following the discussion in Section 3.3.1, the Bayes estimate was computed empirically using

$$mean = \frac{\sum_{s=1}^{8000} \theta^s}{8000}, \tag{5.12}$$

where θ^s represents a sampled parameter value.

In each setting, the data set in every simulation was generated from a fixed data-generating mechanism and contributed a Bayes estimate. Following a frequentist point of view, the collection of Bayes estimates for a parameter forms a sampling distribution of the Bayes estimate. The sampling variability of the estimate was summarized with its mean and standard error. The sampling mean determines if a Bayesian model is capable of drawing an unbiased estimate for the parameter. The sampling variability, measured by the standard error, indicates how far, on average, a Bayes estimate deviates from the sampling mean. The chance of a sample estimate equal or close to the sampling mean is lower in a wider sampling distribution. Therefore, if the sampling mean approaches the true parameter value but the standard error is large, the probability of the sample estimate hitting the neighborhood of true parameter value is low. In contrast, a small standard error indicates that a sample estimate is, in expectation, close to the sampling mean.

5.3.5 Model Comparison

The predictive accuracy of the models were evaluated using WAIC. The model with smaller WAIC has better fit to the data-generating process. WAIC of each model was computed empirically following the expressions

$$\begin{aligned} lppd &= \sum_{n=1}^{500} \log \left(\frac{1}{8000} \sum_{s=1}^{8000} p(y_i|\theta^s) \right) \\ p_{waic} &= \sum_{n=1}^{500} var_i(\log[p(y_i|\theta^s)]) \\ WAIC &= -2lppd - p_{WAIC}. \end{aligned} \tag{5.13}$$

However, the expression of WAIC holds for the Poisson and Negative Binomial regression models, but not for the robust quasi-Poisson regression model. The reason is that the robust quasi-Poisson method does not assume any probabilistic distribution, $p(y|\theta)$, for the underlying data-generating process.

Chapter 6

Results

This chapter presents the empirical results for the 2 simulation studies. A subjective discussion on the result is discussed in Chapter Discussion.

6.1 CASE A: Measurement Error in Covariate

Figure 6.1 shows the empirical means and variances of the sampling distributions of the Bayes estimates (a and b) and WAIC, over 9 scenarios and 3 Bayesian regression methods, in the presence of random measurement error in a single covariate.

Interpretations All methods had a systematic increase in their empirical sampling means of $\hat{a}_{Bayes}$, denoted by ave($\hat{a}_{Bayes}$), and a systematic decrease in in their empirical sampling means of $\hat{b}_{Bayes}$, denoted byave($\hat{b}_{Bayes}$), when the measurement error probability increased and/or error variance increased. This shows that the sampling means of $\hat{a}_{Bayes}$ and $\hat{b}_{Bayes}$ moved further from their true values(2) as the contamination level worsens.

The robust quasi-Poisson and Negative Binomial models performed better than Poisson model in all settings, in terms of the bias level. Nevertheless, the robust quasi-Poisson performed the best among the models in all settings, In particular, when the error probability was below 0.2, its sampling means were very close to the true parameter value, regardless how large sd was. This manifests that the robust quasi-Poisson was robust against the error-prone until a breakdown point in $p = (0.2, 0.6)$. The other models showed non-robustness already at $p = 0.05$ with $sd > 0.1$.

The sampling variances of $\hat{a}_{Bayes}$ and $\hat{a}_{Bayes}$ were below 0.07, in all settings. The coefficients of variation, computed by $\sqrt{var(\hat{a}_{Bayes})}/\text{ave}(\hat{a}_{Bayes})$ were, thus, considerably small. This showed that a sample estimate did not deviate much from its sampling mean.

In terms of predictive accuracy, the Negativa Binomial model, on average, had smaller

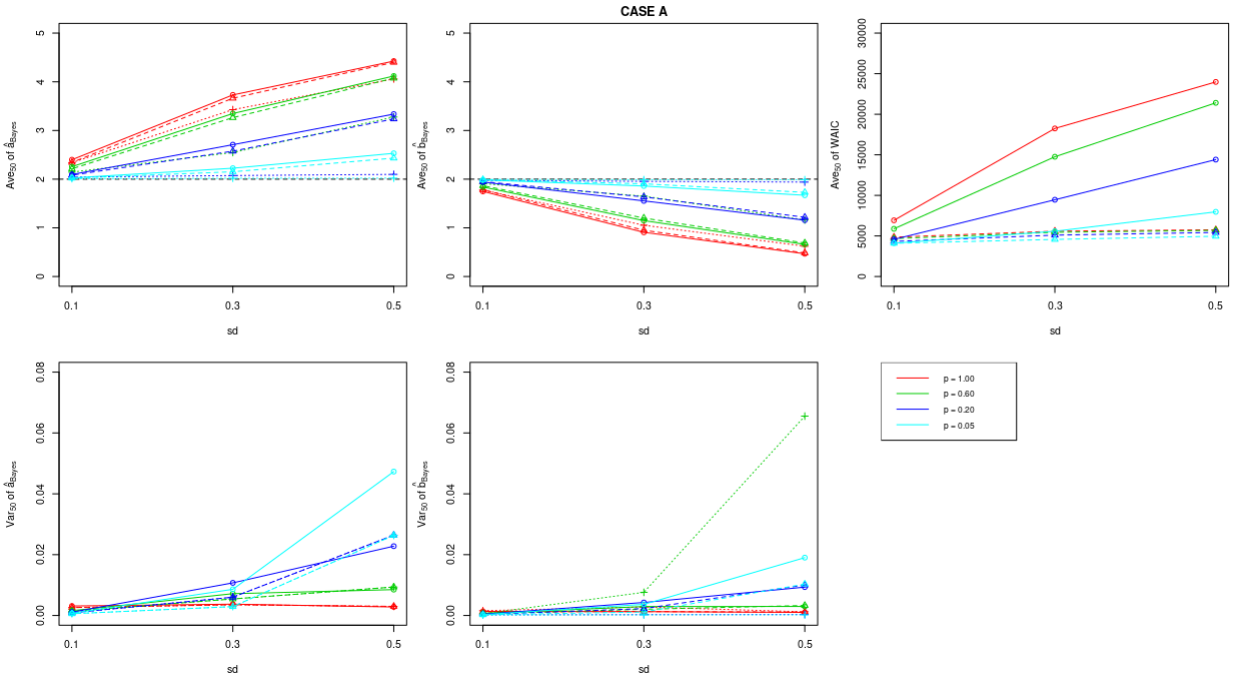

Figure 6.1: CASE A. The plots in the left column shows the mean and variance of the sampling distribution of the Bayes intercept estimate, $\hat{a}_{Bayes}$, for each scenario. The plots in the middle column consists of the mean and variance of the sampling distribution of the Bayes slope estimate, $\hat{b}_{Bayes}$. The plot in the right column illustrates WAICs of the Poisson and NB models, for each scenario. Poisson, NB and robust quasi-Poisson models are denoted by solid, dashed and dotted lines, respectively. The measurement reliability and measurement error variance are denoted by p and sd.

WAIC than the Poisson, in most of the settings. The Negative Binomial model had a better capability of generalizing its inference to new data sets. The performance gap between the Poisson and Negative Binomial became wider when either p or σ^2_{err} increased.

In conclusion, when dealing with an additive error-in-covariates, robust quasi-Poisson performed better than the rest.

6.2 CASE B: Process Error

Figure 6.2 shows the empirical means and variances of the sampling distributions of the Bayes estimates (a and b) and WAIC, over 9 scenarios and 3 Bayesian regression methods, in the presence of random process error.

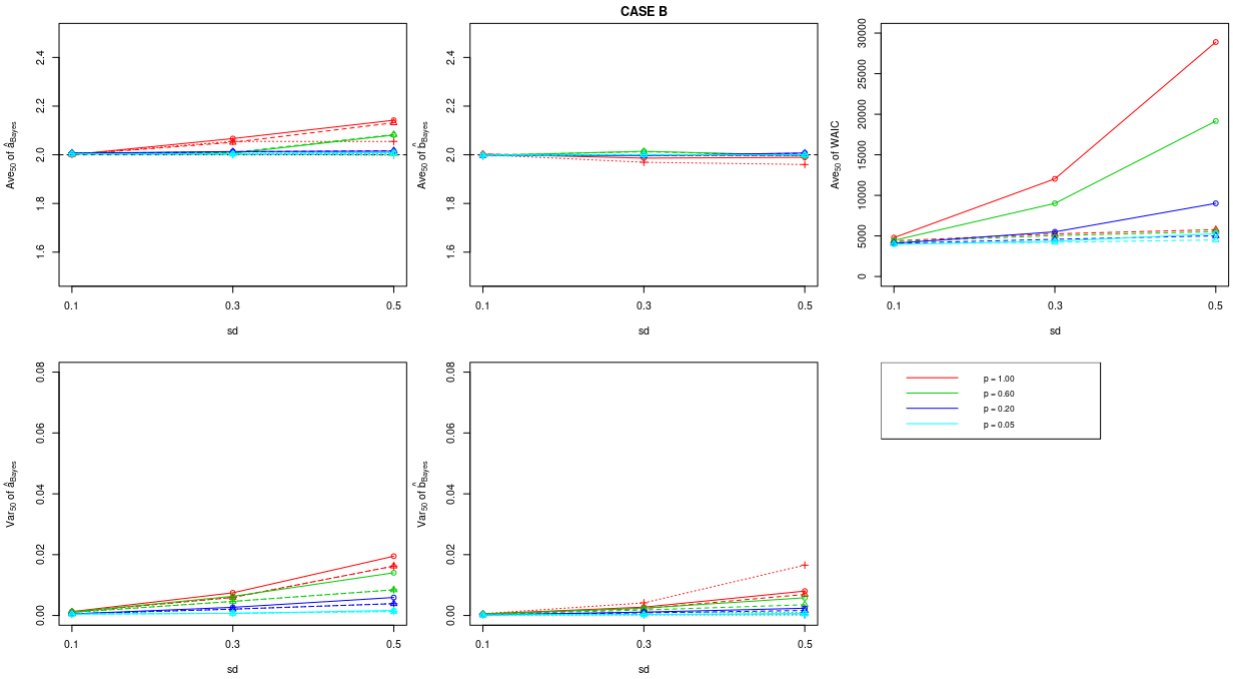

Figure 6.2: CASE B.The plots in the left column shows the mean and variance of the sampling distribution of the Bayes intercept estimate, $\hat{a}_{Bayes}$, for each scenario. The plots in the middle column consists of the mean and variance of the sampling distribution of the Bayes slope estimate, $\hat{b}_{Bayes}$. The plot in the right column illustrates WAICs of the Poisson and NB models, for each scenario. Poisson, NB and robust quasi-Poisson models are denoted by solid, dashed and dotted lines, respectively. The process error probability and process error variance are denoted by p and sd.

Interpretation In the presence of process error, the bias levels of the models were relatively smaller than in Case A (error-in-covariate). All the models were able to infer correctly a and b when p was 0.05 or 0.2. Considering the sampling distributions of both $\hat{a}_{Bayes}$ and $\hat{b}_{Bayes}$, the Negative Binomial and Poisson models broke down in settings whose $p \geq 0.6$ and $sd \geq 0.3$, while the robust quasi-Poisson broke down at $(p, sd) = (1.0, 0.5)$.

The coefficent of variations of 2 estimates, in all settings were considerably small. This shows the estimates of either parameters were close to the the corresponding sampling mean. In terms of predictive accuracy, the Negative Binomial performed better than the Poisson model in most settings.

In conclusion, although the models were not tested in the setting of $0.2 < p < 0.6$, as long as p stayed lower than 0.2, these 3 models were robust against the dispersion due to the process error and able to infer correctly. Nevertheless, it is certain that the breakdown

point of the robust quasi-Poisson model was higher than the other 2 models.

Chapter 7

Discussion

Simulation study A shows that the robust quasi-Poisson model performs better than the Poisson and Negative Binomial distribution, in the presence of measurement error in the covariate. The robust quasi-Poisson model is able to infer approximately correct inference on the effect size of the covariate, under a tolerable error probability and error variance. It broke down much slower than its counterparts. The performance gap between the robust quasi-Poisson model and its counterpart is larger than the gap between the Negative Binomial and Poisson models, although the Negative Binomial model has better predictive accuracy than the Poisson model.

Simulation study B shows that the robust quasi-Poisson model performs empirically better than the Poisson and Negative Binomial models overall, in the presence of process error. Although all the models are able to infer correctly the effect size of the covariates from highly 'contaminated' data sets, the Negative Binomial and Poisson models do not infer the intercept correctly when more than half of the data points are 'contaminated'. The study also shows that the robust quasi-Poisson likelihood models has a relatively higher breakdown point.

The scale mixture of quasi-likelihood discussed here downweights equally the influence of aberrant values on both sides of the mean. However, there are some domains which allow one-side dispersion. Thus, it is important to notice the choice of model is based on scientific reasoning and there is no model that can perform the best generally and is compatible with any data-generating process.

There are some points in the thesis that are worthwhile to have further investigation:

(a) The settings chosen in the thesis are random that are not intended to study the breakdown point of the models. Future investigation can be conducted concerning the breakdown point of models.

(b) In Case A, the models assume that the response is a function of the surrogate (observed) variable, instead of the covariate of interest. The models do not contain the

assumption of measurement error in the covariate. The extension of the models by taking the measurement error into account can be a follow-up to the current work. There are several books and articles concerning different form of measurement error in nonlinear models [6, 10, 11].

(c) The simulation experiments consider the simple form of process error or measurement error, which is still quite unrealistic to real data sets. In practice, the observed data sets could be a combination of measurement error and process error. The measurement error can occur in both response and covariates. A further work can be based on this idea and the model feasibility of real-word data.

(d) The robust quasi likelihood discussed here is derived by having the scale parameter Chi-square distributed. Alternative priors for the scale parameter lead to different likelihood models [35] and the resultant likelihood models can be topics of future studies.

Appendix 1: Weight and Score Functions

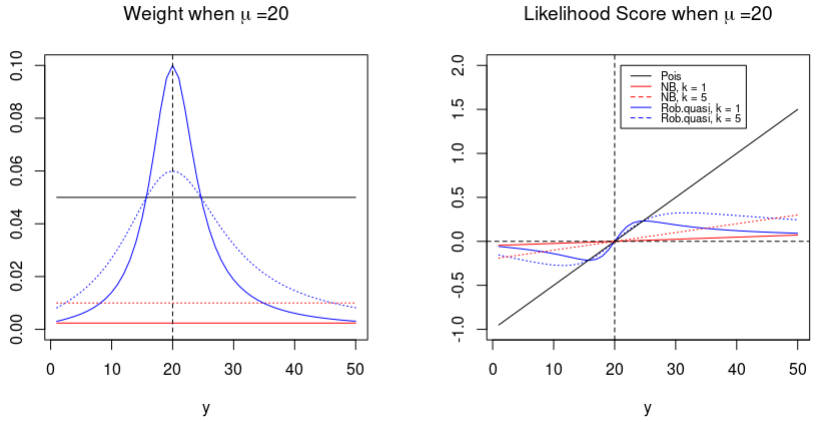

Figure 7.1: Weight (left) and score (right) functions of Poisson, Negative Binomial and robust quasi-poisson likelihoods, given $\mu = 20$ (shown by the dashed vertical line). **Left plot**: y receives the equal weight, independent of $|y - 20|$, for the Poisson and Negative Binomial ($k = 1, 5$) likelihoods, whereas the weight assigned to y decreases when $|y - 20|$ increases for the robust quasi-poisson likelihoods ($k = 1, 5$). **Right plot**: As $|y - 20|$ increases, the absolute score of the Poisson and Negative Binomial ($k = 1, 5$) likelihoods increases, whereas for the robust quasi-likelihoods ($k = 1, 5$), the scores increases up to a point before redescending.

Appendix 2: Stan's Sampling Code

Poisson Model

```
data{
int<lower = 1> n;
int<lower = 0> y[n];
vector[n] x;
int<lower = 1> np;
int<lower = 0> yp[np];
vector[np] xp;
}
parameters{
real a;
real b;
}
transformed parameters{
vector[np] log_lik;

for(i in 1:np){
log_lik[i] = poisson_lpmf(yp[i] | exp(a + b*xp[i]));
}
}
model{
a ~ normal(0, sqrt(10^6));
b ~ normal(0, sqrt(10^6));
y ~ poisson(exp(a + b*x));
}
```

Negative Binomial Model

```
data{
int<lower = 1> n;
int<lower = 0> y[n];
vector[n] x;
int<lower = 1> np;
int<lower = 0> yp[n];
vector[n] xp;
}
parameters{
real a;
real b;
real<lower = 0> r;
}
transformed parameters{
vector[np] log_lik;

for (i in 1:np) {
log_lik[i] = neg_binomial_2_lpmf(yp[i] |  exp(a + b*xp[i]), r);
}
}
model{
a ~ normal(0, sqrt(10^6));
b ~ normal(0, sqrt(10^6));
r ~ gamma(1,1);

for (i in 1:n) {
y[i] ~ neg_binomial_2(exp(a + b*x[i]), r);
}
}
```

Robust Quasi-Poisson Likelihood Model

```
functions{
real robust_lpmf(int Y, real lambda, real k){
return log(k + 2 * (Y * log(Y/lambda) - Y + lambda ) )^(-(k+1)/2);
}
```

```
}
data{
int<lower = 1> n;
int<lower = 0> y[n];
vector[n] x;
real<lower = 0> k;
}
parameters{
real a;
real b;
}
model{
a ~ normal(0, sqrt(10^6));
b ~ normal(0, sqrt(10^6));

for(i in 1:n){
y[i] ~ robust(exp(a + b*x[i]), k);
}
}
```

Appendix 3: R-code

```
library("matrixStats")
library(rstan)
options(mc.cores = parallel::detectCores())

setwd("/home/tanshuz/[Masters] Code")

pois_model <- stan_model(file = "pois_model.stan")
NB_model2 <- stan_model(file = "NB_model2.stan")
robust_pois <- stan_model(file = "Robust_pois.stan")

set.seed(25)
n = 500
a = 2
b = 2
P = c(1, 0.6, 0.2, 0.05)# contamination level
sds = c(0.5, 0.3, 0.1) # measurement error variance
sigmas = c(0.5, 0.3, 0.1) # process error variance
R = 50 # number of realizations

colnames = c("a_mu", "b_mu", "WAIC")
mtds = c('pois', 'nb', 'robust') # methods

# Create list [ests x method x P x sds/sigmas]
temp = matrix(NA, nrow = R, ncol = length(colnames),
dimnames = list(NULL, colnames))
temp = sapply(mtds, function(x) temp, USE.NAMES = TRUE, simplify = "array")
temp = sapply(as.character(P), function(x) temp, USE.NAMES = TRUE,
simplify = "array")
summ_A = sapply(as.character(sds), function(x) temp, USE.NAMES = TRUE,
```

```
simplify = "array")
summ_B = sapply(as.character(sigmas), function(x) temp, USE.NAMES = TRUE,
simplify = "array")

# Storing means for a, b and WAIC
temp = matrix(NA, nrow = 1, ncol = length(colnames),
dimnames = list(NULL, colnames))
temp = sapply(mtds, function(x) temp, USE.NAMES = TRUE,
simplify = "array")
temp = sapply(as.character(P), function(x) temp, USE.NAMES = TRUE,
simplify = "array")
mean_A = var_A = sapply(as.character(sds), function(x) temp, USE.NAMES = TRUE,
simplify = "array")
mean_B = var_B = sapply(as.character(sigmas), function(x) temp, USE.NAMES = TRUE,
simplify = "array")
```

Case A

```
for(s in 1:length(sds)){
for(p in 1:length(P)){
for(r in 1:R){
idx = rbinom(n = n, size = 1, prob = P[p]) + 1
x.true = runif(n, min = 1, max = 2)
x.obs = x.true + rnorm(n, mean = 0, sd = c(0, sds[s])[idx])
y.obs = rpois(n = n, lambda = exp(a + b*x.true))

dataset = list("n" = n, "y" = y.obs, "x" = x.obs,
"np" = n, "yp" = y.obs, "xp" = x.obs)

# POISSON
post = sampling(pois_model, data = dataset, warmup = 1000,
iter = 3000, chains = 4, thin = 1,
control = list(adapt_delta = 0.99, max_treedepth = 20))
M = extract(post)
lppd = sum(apply(M$log_lik, 2, logSumExp) - log(dim(M$log_lik)[1]))
p_waic = sum(colMeans((M$log_lik)^2) - (colMeans(M$log_lik))^2)
WAIC_pois = -2*(lppd - p_waic)
```

```
summ_A[r, , 'pois', p, s] = c(mean(M$a), mean(M$b), WAIC_pois)

# NEGATIVE BINOMIAL
post_nb = sampling(NB_model2, data = dataset, warmup = 1000,
iter = 3000, chains = 4, thin = 1,
control = list(adapt_delta = 0.99, max_treedepth = 20))
M_nb = extract(post_nb)
lppd = sum(apply(M_nb$log_lik, 2, logSumExp) - log(dim(M_nb$log_lik)[1]))
p_waic = sum(colMeans((M_nb$log_lik)^2) - (colMeans(M_nb$log_lik))^2)
WAIC_nb = -2*(lppd - p_waic)
summ_A[r, , 'nb', p, s] = c(mean(M_nb$a), mean(M_nb$b), WAIC_nb)

# ROBUST QUASI-LIKELIHOOD
dataset = list("n" = n, "y" = y.obs, "x" = x.obs,
"k" = 5)
post_rob = sampling(robust_pois, data = dataset, warmup = 1000,
iter = 3000, chains = 4, thin = 1,
control = list(adapt_delta = 0.99, max_treedepth = 20))
M_rob = extract(post_rob)
summ_A[r, , 'robust', p, s] = c(mean(M_rob$a), mean(M_rob$b), NA)
}

for(m in mtds){
mean_A[ , , m, p, s] = apply(summ_A[, , m, p, s], 2, mean)
var_A[ , , m, p, s] = apply(summ_A[, , m, p, s], 2, var)
}
}
}
```

Case B

```
for(s in 1:length(sigmas)){
for(p in 1:length(P)){
for(r in 1:R){
idx = rbinom(n = n, size = 1, prob = P[p]) + 1
x.true = runif(n, min = 1, max = 2)
x.obs = x.true
eps = rnorm(n, mean = 0, sd = c(0, sigmas[s])[idx])
```

```
y.obs = rpois(n = n, lambda = exp(a + b*x.true + eps))

# POISSON
post = sampling(pois_model, data = dataset, warmup = 1000,
iter = 3000, chains = 4, thin = 1,
control = list(adapt_delta = 0.99, max_treedepth = 20))
M = extract(post)
lppd = sum(apply(M$log_lik, 2, logSumExp) - log(dim(M$log_lik)[1]))
p_waic = sum(colMeans((M$log_lik)^2) - (colMeans(M$log_lik))^2)
WAIC_pois = -2*(lppd - p_waic)
summ_B[r, , 'pois', p, s] = c(mean(M$a), mean(M$b), WAIC_pois)

# NEGATIVE BINOMIAL
post_nb = sampling(NB_model2, data = dataset, warmup = 1000,
iter = 3000, chains = 4, thin = 1,
control = list(adapt_delta = 0.99, max_treedepth = 20))
M_nb = extract(post_nb)
lppd = sum(apply(M_nb$log_lik, 2, logSumExp) - log(dim(M_nb$log_lik)[1]))
p_waic = sum(colMeans((M_nb$log_lik)^2) - (colMeans(M_nb$log_lik))^2)
WAIC_nb = -2*(lppd - p_waic)
summ_B[r, , 'nb', p, s] = c(mean(M_nb$a), mean(M_nb$b), WAIC_nb)

# ROBUST QUASI-LIKELIHOOD
dataset = list("n" = n, "y" = y.obs, "x" = x.obs,
"k" = 5)
post_rob = sampling(robust_pois, data = dataset, warmup = 1000,
iter = 3000, chains = 4, thin = 1,
control = list(adapt_delta = 0.99, max_treedepth = 20))
M_rob = extract(post_rob)
summ_B[r, , 'robust', p, s] = c(mean(M_rob$a), mean(M_rob$b), NA)
}

for(m in mtds){
mean_B[ , , m, p, s] = apply(summ_B[, , m, p, s], 2, mean)
var_B[ , , m, p, s] = apply(summ_B[, , m, p, s], 2, var)
}
}
```

Bibliography

[1] Agresti, A. (2015). Foundations of Linear and Generalized Linear Models. In John Wiley & Sons, INC.

[2] Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. Machine Learning, 50(1-2), 5-43. https://doi.org/10.1023/A:1020281327116

[3] Bartlett, J. W., & Keogh, R. H. (2018). Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration. Statistical methods in medical research, 27(6), 1695-1708. https://doi.org/10.1177/0962280216667764

[4] Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. arXiv: Methodology.

[5] Casella, G., & Berger, R. (2001). Statistical inference (2nd ed.). Thomson Learning.

[6] Carroll, R.J., Ruppert, D., Stefanski, L.A., & Crainiceanu, C.M. (2006). Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420010138

[7] Clark, D.R., & Thayer, C. (2004). A Primer on the Exponential Family of Distributions.

[8] Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2014). Bayesian data analysis (Third edition.). CRC Press.

[9] Guikema, S. D., & Goffelt, J. P. (2008). A flexible count data regression model for risk analysis. Risk Analysis, 28(1), 213-223. https://doi.org/10.1111/j.1539-6924.2008.01014.x

[10] Gustafson, P. (2003). Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780203502761

[11] Hardin, J., & Carroll, R. (2003). Measurement Error, GLMs, and Notational Conventions. The Stata Journal, 3, 329-341.

[12] Hayat, M. J., & Higgins, M. (2014). Understanding poisson regression. Journal of Nursing Education, 53(4), 207-215. https://doi.org/10.3928/01484834-20140325-04

[13] Hilbe, J. (2011). Negative binomial regression (2nd ed.). Cambridge University Press.

[14] Hobbs, N. T. & Hooten, M. B. (2015). Bayesian models: A statistical primer for ecologists. Princeton University Press.

[15] Hoffman, M. D. & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research. 15.

[16] Jϕrgensen, B. (1992). Exponential Dispersion Models and Extensions: A Review. International Statistical Review / Revue Internationale De Statistique, 60(1), 5-20. doi:10.2307/1403498

[17] Lambert, B. (2018). A student's guide to Bayesian statistics. SAGE, Los Angeles.

[18] Lindén, A., & Mäntyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. Ecology, 92(7), 1414-1421. https://doi.org/10.1890/10-1831.1

[19] Liu, J., & Vanhatalo, J. (2020). Bayesian model based spatiotemporal survey designs and partially observed log Gaussian Cox process. Spatial statistics, 35, [100392]. https://doi.org/10.1016/j.spasta.2019.100392

[20] McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models. Chapman and Hall, London.

[21] Nelder, J., & Wedderburn, R. (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), 135(3), 370-384. doi:10.2307/2344614

[22] Neyman, J. $ Scott, E. L.(1971). Outlier proneness of phenomena and of related distributions. Optimizing Methods in Statistics. Academic Press, New York.

[23] O'Hagan, A. (1979) On outlier rejection phenomena in Bayes inference. J. R. Statist. Soc. B, 41, 358-367.

[24] O'Hagan, A. (2004). Bayesian statistics: principles and benefits. 3.

[25] O'Hagan, A., Forster, J., & Kendall, M. (2004). Kendall's advanced theory of statistics. Volume 2B, Bayesian inference (Second edition.). Arnold.

[26] O'Hagan, A., & Pericchi, L. (2012). Bayesian heavy-tailed models and conflict resolution: A review. Brazilian Journal of Probability and Statistics, 26(4), 372-401. Retrieved March 31, 2021, from http://www.jstor.org/stable/43601225

[27] Ramsay, J. O. & Novick, M. R. (1980). PLU robust Bayesian decision theory: point estimation. J. Amer. Statist. Ass., 75,901-907.

[28] Stan Development Team. (2019). Stan Modeling Language Users Guide and Reference Manual, Version 2.26. https://mc-stan.org

[29] Stan Development Team. (2020, July 26). Brief Guide to Stan's Warnings. https://mc-stan.org/misc/warnings.html

[30] van de Schoot, R., Depaoli, S., King, R. et al. (2021). Bayesian statistics and modelling. Nat Rev Methods Primers 1, 1. https://doi.org/10.1038/s43586-020-00001-2

[31] Tsou, T. (2006). Robust Poisson regression. Journal of Statistical Planning and Inference, 136, 3173-3186.

[32] Ver Hoef, J. M., and P. L. Boveng. (2007). Quasi-Poisson vs. negative binomial regression: How should we model over- dispersed count data? Ecology 88:2766-2772.

[33] Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. Journal of Machine Learning Research, 11, p. 3571–3594.

[34] Wedderburn, R. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. Biometrika, 61(3), 439-447. doi:10.2307/2334725

[35] West, M (1985). Generalized linear models: Scale parameters, outlier accommodation and prior distributions. In Bernardo, JM DeGroot, MH Lindley, DV Smith, AFM eds. Bayesian Statistics 2. North-Holland: Elsevier, 531-58.

[36] West, M. (1984), Outlier Models and Prior Distributions in Bayesian Linear Regression. Journal of the Royal Statistical Society: Series B (Methodological), 46: 431-439. https://doi.org/10.1111/j.2517-6161.1984.tb01317.x