Université de Montréal

Robust Gamma Generalized Linear Models with Applications in Actuarial Science

par

Yuxi Wang

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en statistique

September 12, 2022

Université de Montréal

Faculté des études supérieures et postdoctorales

Ce mémoire intitulé

Robust Gamma Generalized Linear Models with Applications in Actuarial Science

présenté par

Yuxi Wang

a été évalué par un jury composé des personnes suivantes :

Maciej Augustyniak
(président-rapporteur)

Philippe Gagnon
(directeur de recherche)

Louis Doray
(membre du jury)

Mémoire accepté le :

_____

Abstract

Generalized linear models (GLMs) form one of the most popular classes of models in statistics. This class contains a large variety of commonly used regression models, such as normal linear regression, logistic regression and gamma GLMs. In GLMs, the response variable distribution defines an exponential family. A drawback of these models is that they are non-robust against outliers. For models like the normal linear regression and gamma GLMs, the non-robustness is a consequence of the exponential tails of the densities. The difference in trends in the bulk of the data and the outliers yields skewed inference and prediction.

To our knowledge, there is no Bayesian robust approach specifically for GLMs. The most popular method is frequentist; it is that of Cantoni and Ronchetti (2001). Their approach is to adapt the robust M-estimators for linear regression to the context of GLMs. However, their estimator is derived from a modification of the derivative of the log-likelihood, instead of from a modification of the likelihood (as with robust M-estimators for linear regression). As a consequence, it is not possible to establish a clear correspondence between the modified function to optimize and a model. Having a robust model has two advantages. First, it allows for an understanding and an interpretation of the modelling. Second, it allows for both frequentist and Bayesian analysis. The method we propose is based on ideas from Bayesian robust linear regression. We adapt the approach proposed by Gagnon et al. (2020), which consists of using a modified normal distribution with heavier tails for the error term. In our context, the distribution of the response variable is a modified version where the central part of the density is kept as is, while the extremities are replaced by log-Pareto tails, behaving like $(1/|x|)(1/\log|x|)^{\lambda}$. The focus of this thesis is on gamma GLMs. The performance is measured both theoretically and empirically, with an analysis of hospital costs data.

Keywords: Bayesian statistics; heavy-tailed distributions; outlier detection; outliers; Pearson residuals.

Résumé

Les modèles linéaires généralisés (GLMs) constituent l'une des classes de modèles les plus populaires en statistique. Cette classe contient une grande variété de modèles de régression fréquemment utilisés, tels que la régression linéaire normale, la régression logistique et les gamma GLMs. Dans les GLMs, la distribution de la variable de réponse définit une famille exponentielle. Un désavantage de ces modèles est qu'ils ne sont pas robustes par rapport aux valeurs aberrantes. Pour les modèles comme la régression linéaire normale et les gamma GLMs, la non-robustesse est une conséquence des ailes exponentielles des densités. La différence entre les tendances de l'ensemble des données et celles des valeurs aberrantes donne lieu à des inférences et des prédictions biaisées.

A notre connaissance, il n'existe pas d'approche bayésienne robuste spécifique pour les GLMs. La méthode la plus populaire est fréquentiste ; c'est celle de Cantoni and Ronchetti (2001). Leur approche consiste à adapter les M-estimateurs robustes pour la régression linéaire au contexte des GLMs. Cependant, leur estimateur est dérivé d'une modification de la dérivée de la log-vraisemblance, au lieu d'une modification de la vraisemblance (comme avec les M-estimateurs robustes pour la régression linéaire). Par conséquent, il n'est pas possible d'établir une correspondance claire entre la fonction modifiée à optimiser et un modèle. Le fait de proposer un modèle robuste présente deux avantages. Premièrement, il permet de comprendre et d'interpréter la modélisation. Deuxièmement, il permet l'analyse fréquentiste et bayésienne. La méthode que nous proposons s'inspire des idées de la régression linéaire robuste bayésienne. Nous adaptons l'approche proposée par Gagnon et al. (2020), qui consiste à utiliser une distribution normale modifiée avec des ailes plus relevées pour le terme d'erreur. Dans notre contexte, la distribution de la variable de réponse est une version modifiée où la partie centrale de la densité est conservée telle quelle, tandis que les extrémités sont remplacées par des ailes log-Pareto, se comportant comme $(1/|x|)(1/\log|x|)^{\lambda}$. Ce mémoire se concentre sur les gamma GLMs. La performance est mesurée à la fois théoriquement et empiriquement, avec une analyse des données sur les coûts hospitaliers.

Mots-clés : Statistiques bayésiennes ; distributions à ailes relevées ; détection des valeurs aberrantes ; valeurs aberrantes ; résidus de Pearson.

Contents

List of Tables

List of Figures

List of Acronyms

GLM: generalized linear model ix

OLS: ordinary least squares xi

PDF: probability density function xi

LPTN: log-Pareto-tailed normal 5

PMF: probability mass function 8

MLE: maximum likelihood estimator 10

MAP: maximum a posteriori estimator 18

MCMC: Markov chain Monte Carlo 18

CDF: cumulative distribution function 36

Acknowledgment

Before all, I would like to thank my supervisor, whose financial support for my research helped to reduce enormously the financial burden. He was very patient and responsible in helping me revise my scholarship applications, which led to different precious awards. Here, I would like to thank all the industry, associations, and institution that provided me with such a great support. They are LifeWorks, Fin-ML, Mitacs, the faculty of Graduate and Postdoctoral Studies of the Université de Montréal, and of course, the department of Mathematics and Statistics.

I first direct my gratitude to Professor Philippe Gagnon, who is indelibly credited with my project. I must mention that without his advice at the beginning of my program, I would not have even chosen the thesis-based track. From the first day of this research project, we have been meeting regularly once a week, which had a very positive supervisory effect. During the whole experience, I have rarely felt lost or uncertain, because Philippe always pointed me in a clear and specific direction. I have enjoyed countless pleasant academic discussions with him, from general ideas of a statistical model to a detail in a mathematical proof. During the writing of this thesis, Philippe constantly proposed pertinent comments, and thanks to them, I did see my progress step by step. His generous and amiable personality also made my research experience a real pleasurable journey.

I would like to thank all the teachers who taught me during my undergraduate and graduate studies. In particular, Professor Christian Léger opened my door to the world of statistics with his course *concepts and methods in statistics*. Professor Mylène Bédard not only gave attractive lectures, but also offered genuine concern and advice. I am grateful for the generous availability of Professor Robert Owens and the enthusiast communication of a math problem with him. The weekly lecture club, organized by Professor Florian Maire, allowed me to exchange ideas with other intelligent minds having an appetite for statistics. Finally, I would like to thank Professors Pierre Duchesne and Marlène Frigon for their recognition and trust in me to teach the course *sampling theory*.

Thank you to all the people contributing to my university life, who impressed me by their integrity, helped me with their kindness, and motivated me by their passion for mathematics.

My special thanks go to An, Dorchelle, Émilyne, Gabriel, Guillaume, Ismael, Siying, Xurui and Yan.

In the end, I would like to thank my parents for their endless tolerance and support. Their pride in me is my greatest motivation to move forward. Also, thanks to the precious friends in my life who share the joy and the sufferings. Without them to balance my work and life, I would not have been able to make all the achievements. My special thanks go to Florent, Hanwen, Kira, Siying, Xin, Xinran, Yiyao, Yuli and Ziqi.

Introduction

Generalized linear models (GLMs) are regression models introduced by Nelder and Wedderburn (1972). They generalize normal linear regression, i.e. linear regression with normally distributed errors, in the following way: in normal linear regression, the dependent variable is modelled by using a normal distribution with parameters that are functions of explanatory variables; with GLMs, the distribution of the dependent variable defines an exponential family, which is the case for the normal distribution, and parameters of the assumed distribution depend on the explanatory variables as with normal linear regression. As a result, GLMs can handle both discrete and continuous responses, with distribution shapes that offer flexibility regarding in particular the skewness. A specificity of GLMs is that the expectation of the response variable is linear in the explanatory variables, up to a transformation. GLMs are among the most widely used classes of statistical models, with applications ranging from actuarial science (Goldburd et al., 2019) to medicine (Casals et al., 2014). GLMs indeed cover many popular statistical models such as, as mentioned, the classical linear regression for normally distributed responses, logistic regression for binary ones, Poisson regression for count data, gamma regression for right-skewed positive data, plus many other statistical models obtained through its general model formulation. For an excellent reference about GLMs, their applications and features, we refer readers to the book of Dobson and Barnett (2018).

The use of GLMs in actuarial fields can be traced back to the early 1980s. Indeed, McCullagh and Nelder (1983) give many examples of the fitting of GLMs to insurance data, such as average claim costs data from a motor insurance portfolio. In the insurance industry, levels of interest and rates of adoption for this class of models have increased to the point where it now seems as though GLMs are near-ubiquitous (Goldburd et al., 2019). Among all the GLMs, the gamma and inverse Gaussian GLMs, meaning that the dependent variable has a gamma or inverse Gaussian distribution, are commonly used for modelling insurance claim severity due to the similar characteristics of the PDFs with those of the observed data. The PDFs of these two distributions are both right-skewed, have a sharp peak with an exponential right tail, and are supported on the positive real numbers. Compared to the gamma, inverse Gaussian has a sharper peak at a positive value and a wider tail, and

is therefore appropriate for situations where the skewness of the severity curve is expected to be somewhat extreme (Goldburd et al., 2019). Gamma distribution can have a peak at a positive real value or at 0, whereas the mode is always positive in inverse Gaussian distribution. Thus, the shape parameter of the gamma distribution allows the model to fit different scenarios. Our understanding is that gamma GLMs are preferred in typical situations in practice. It can for instance be used by an insurance company to determine the factors that contribute the most to the claim size and how they influence the latter, and to predict claims based on a given set of explanatory variables, ultimately leading to the pricing of insurance products.

Insurance companies, like any companies exploiting data for commercial use on a daily basis, are however not shielded from issues such as data quality. Also, extreme claims are often present in their data bases. Both issues have a negative impact on the conclusions drawn and predictions made from statistical analyses. This is due to the non-robustness of the regression models typically employed, such as gamma GLMs, against outliers, and therefore against data with gross errors and extreme claims. More specifically, the non-robustness is a consequence of the exponential tails of those models, combined with the difference in the trends in the bulk of the data and the outliers. When the likelihood function is evaluated at parameter values reflecting the trends in the bulk of the data, the exponential tails penalize heavily those values for the outliers, diminishing significantly the likelihood function value. The analogous phenomenon arises when the likelihood function is evaluated at parameter values reflecting the trends in the outliers: those values are heavily penalized for the bulk of the data. All this makes values in between those mentioned more likely, representing an undesirable compromise. The resulting maximum likelihood estimates are thus consistent with neither the bulk of the data nor the outliers. Because the model adjusts itself for the outliers, another undesirable consequence is that identifying outliers using standard measures such as Pearson residuals (which will be defined in detail in Chapter 1) may be ineffective. This is called *the masking effect* (Hadi and Simonoff, 1993), as outliers may mask one another due to an adjustment of the model. Moreover, univariate analyses of extreme values may not allow to deal with the problem, because outliers here are considered as outliers with respect to the model employed, i.e. data points that are unlikely under that model when using parameter values reflecting the trends in the bulk of the data. A data point can thus be an outlier with respect to the model without having any extreme values in the explanatory or response variables.

All that motivates the use of robust GLMs in situations where one wants protection in case the data set to be analysed contains outliers. The non-robustness properties of classical maximum likelihood estimators in the context of GLMs have been studied by several authors. Pregibon (1982) proposed a resistant fitting method for logistic regression, by applying different loss functions at the step of estimation. Stefanski et al. (1986) and

Kunsch et al. (1989) studied optimally bounded score functions for estimating parameters in GLMs. However, all the aforementioned methods focus particularly on the logistic regression. Cantoni and Ronchetti (2001) studied robust estimators for GLMs based on the notion of quasi-likelihood, and their approach is to adapt the robust M-estimators for linear regression of Huber (1973). Their approach is valid for any GLM, and is also the most commonly used approach for robust GLMs. These are all, of course, frequentist approaches. On the Bayesian side, we did not find any robust approach specifically for GLMs. A general approach is that of Bissiri et al. (2013) which introduces another statistical paradigm based on the premise that the model assumed is incorrect, but we consider it as another type of approach and will not focus on such approaches in the current document. Bayesian robust approaches typically consist in adapting the original model to the presence of outliers by replacing the distribution by one that is similar, but with heavier tails. A famous example is a robust Bayesian linear regression where the normal distribution of the errors is replaced by a Student distribution (West, 1984).

Frequentist and Bayesian robust methods are often seen as being fundamentally different. In former methods, the loss function to be minimized or the likelihood function to be maximized at the step of estimation is modified for the purpose of diminishing the impact of outliers, whereas in the latter, the original PDF is directly replaced by another density which, while being as similar as possible to the original one, has heavier tails. Interestingly, these two approaches are connected. Indeed, the modified loss function in linear regression is often quadratic below a certain threshold, but then grows more slowly beyond that threshold. As shown in detail in Chapter 4, this can be seen as using a modified normal PDF with tails that have been replaced by heavier ones. With the approach of Cantoni and Ronchetti (2001) applicable for distributions which have tails for GLMs, it is not possible to establish a clear correspondence between the modified function to optimize for estimation and a model. This is also explained in Chapter 4.

The approach of using a modified normal PDF with tails that have been replaced by heavier ones has been proposed by Desgagné (2015) in a context of Bayesian location-scale models. The PDF used is called log-Pareto-tailed normal (LPTN) distribution because the central part of this continuous density is that of the standard normal and the tails are log-Pareto, meaning that they behave like $(1/|x|)(1/\log|x|)^{\lambda}$, hence is name. This distribution belongs to the class of log-regularly varying distributions introduced in Desgagné (2015). This approach was subsequently adapted to the context of Bayesian linear regression by Gagnon et al. (2020). The authors assumed that the distribution of the error was LPTN instead of normal. Estimation of parameters based on this model was shown to be more robust than the Student one; the latter is the most popular Bayesian solution for robust linear regression.

With this project, we take one step further by adapting that approach to GLMs, which means that the distribution of the dependent variable is a modified version where the central is kept as is, while the extremities are replaced by log-Pareto tails. We focus on gamma GLMs throughout the document, but our approach remains valid for any GLM based on a distribution with tails, whether it is continuous or discrete, such as inverse Gaussian or Poisson GLMs. Our approach has two advantages over the most popular approach to robust GLMs of Cantoni and Ronchetti (2001):

- firstly, we have a precise characterization of the model, which is important from a modelling point of view;
- secondly, given that the approach is a direct modification of the distribution of the dependent variable, it can be used for both frequentist and Bayesian analyses.

We now present how the document is organized. We start this document by introducing, in Chapter 1, the general ideas of GLMs, gamma GLMs in particular, as well as estimation and inference for GLMs. In Chapter 2, we present a data set of health care expenditure at a hospital in Lausanne and an analysis based on (non-robust) gamma GLMs. Such analysis can be conducted by actuaries in insurance companies to get insights regarding the key factors influencing the costs. This data set contains several outliers, and with this example, issues mentioned earlier regarding non-robust GLMs become apparent. In particular, there is an evident masking effect which has a significant negative impact on the outlier detection. In Chapter 3, we present in detail the approach of Cantoni and Ronchetti (2001) and apply it to analyse the same data set as in Chapter 2. Presenting their approach is useful to compare it with ours. It helps to highlight in Chapter 4 that their modified function cannot be clearly connected to a model, because the function that they modified to gain in robustness is the derivative of the log-likelihood, instead of the log-likelihood, as with robust linear regression. In Chapter 4, we also explain how some frequentist robust linear regression approaches can be connected to the use of heavy-tailed distributions. This motivates the introduction of our method in Chapter 5. In Chapter 5 we also explain how to perform inference using our method, which is applied to analyse the data set mentioned earlier. Theoretical results are also presented: we present sufficient conditions under which the posterior distribution for a Bayesian analysis is proper, and results that allow to characterize the asymptotic behaviour of the likelihood function and posterior distribution as outliers move away from the bulk of the data. We also present simulation results that support the latter results.

Chapter 1

Generalized Linear Models

We start this chapter in Section 1.1 with a general definition of GLMs. We present three components characterizing this class of models:

- random component: a response variable that is assumed to have a distribution defining an exponential family;
- linear predictor: a set of explanatory variables and regression coefficients that together produce a linear predictor;
- link function: a function which links previous components together.

We then have a closer look in Section 1.2 at the special case of gamma GLMs, representing, as mentioned in the introduction, the focus of our work. In Section 1.3, inference and estimation methods for GLMs from both frequentist and Bayesian perspectives are explained.

1.1. General Definition

Proposed by Nelder and Wedderburn (1972), GLMs were created out of a desire to bring under one umbrella a wide variety of regression models that span a spectrum from normal linear regression to logistic regression. The response variable follows no longer strictly a Gaussian distribution, such as with normal linear regression, but any distribution defining an exponential family, such as a Poisson distribution, a Binomial distribution, a gamma distribution, etc. In addition to the diversity of the distribution of the response, the relation between the response and explanatory variables is not necessarily linear in GLMs; it is determined through a link function $l(\cdot)$ chosen by users. This function often relates non-linearly a linear combination of the explanatory variables to the mean of the distribution. Additionally, appropriate link functions have inverse functions which allow to map such a linear combination taking values on the real line to an interval corresponding to the support of the distribution of the response, so that predictions belong to the support. For instance, the exponential function (the inverse of the logarithmic link function) is used in gamma GLMs to map the linear combination to a positive number. Besides, certain GLMs, such as gamma and inverse Gaussian GLMs, take heteroscedasticity into account through a dispersion parameter,

allowing the variance of the response to vary with the explanatory variables. In brief, GLMs greatly improve the modelling flexibility over normal linear regression. Therefore, there are considerably more types of data that can be dealt with by using GLMs than normal linear regression.

Let us now present a formal definition of GLMs. Let $y \in \mathbb{R}$ be a random variable representing the random variable that we are interested in modelling, called the *response* in our context. A GLM is such that the distribution of y defines an exponential family, i.e. its PDF or probability mass function (PMF) is such that

$$f_{\theta,\phi}(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{1.1.1}$$

where $\theta \in \mathbb{R}$ is the canonical parameter, $\phi > 0$ is a dispersion parameter, and $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are some specific functions that define the distribution of y and are seen to satisfy regularity conditions. Typically, $a(\phi) = \phi$ or $a(\phi) = \phi/w$, where the weight $w > 0$ is usually known. For example, with binomial distributions, the weight w is the number of independent experiments. For an exponential family, the expected value is $\mathbb{E}[y] = b'(\theta) =: \mu$ and the variance is $\mathbb{V}ar[y] = a(\phi)b''(\theta)$, where b' and b'' denote the first and second derivatives of b, respectively. The term $b''(\theta) = b''[(b')^{-1}(\mu)] = v(\mu)$ represents what is called the *mean-variance relationship*, where $(b')^{-1}$ denotes the inverse function of b'. The function v determines the effect of the mean on the variance.

In a GLM, the information carried by the explanatory variables is incorporated by setting $l(\mu) =: \eta := \boldsymbol{x}^T\boldsymbol{\beta}$, where η is called the *linear predictor* in our context, $\boldsymbol{x} := (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ is a vector of p explanatory variables, and $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is the vector of regression coefficients. Regarding the explanatory variables, $x_1 = 1$ to introduce an intercept in the model; other explanatory variables can be quantitative, qualitative, or mixed, in the case of an interaction term that is the product of a quantitative variable and a qualitative factor. As mentioned, the function l is called the *link function* in our context. It satisfies the following condition: strictly monotone and differentiable. For example, the link is often chosen to be the identity when the distribution of the response is normal, and to be the logarithm when the distribution of the response is Poisson or gamma. It is named *canonical link* when $l(\mu) = (b')^{-1}(\mu)$, thus the resulting GLM assumes that $\mu = l^{-1}(\theta)$ when this link function is applied. We can obtain benefits of convenient mathematical and algorithmic properties from using the canonical link in modelling, as will be explained in the Section 1.3.1. Even if convenient, this link can be replaced with a different one for the purpose of practicality or enhanced interpretation. For instance, in gamma GLMs, the logarithmic link $\log(\mu)$ is preferable to the canonical link $-1/\mu$. The predicted value of a response using the former link is

always positive, which corresponds to the domain of gamma observations. Indeed, the predicted value is $\hat{\mu} = \exp(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})$, where $\hat{\mu}$ and $\hat{\boldsymbol{\beta}}$ are mean and regression coefficient estimates, respectively. However, if the canonical link is applied, the predicted value is $\hat{\mu} = -1/(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})$, which can be negative. In terms of interpretation, the log link yields a multiplicative effect of the explanatory variables (other than the intercept) with respect to what can be seen as a base rate, $\exp(\beta_1)$, which is the value when all explanatory variables (except the intercept) are null. This interpretation makes sense when all explanatory variables are categorical (0–1) variables or standardized continuous variables, given that a departure from 0, which can be seen as a base value, results in a multiplicative adjustment of the base rate. For example, let x_2 be the sex of a person, and y be the cost of stay at a certain hospital. Considering that $x_2 = 0$ signifies that the person is a female, the expected cost of stay for a man is the base rate $\exp(\beta_1)$ that is adjusted by multiplying by a factor of $\exp(\beta_2)$, if all other explanatory variables are null. This interpretation is practical in actuarial science, and in particular in insurance pricing.

Notice that with GLMs, we cannot in general write y as with the traditional linear regression, i.e. $y = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$, where ϵ is a zero mean error. We make instead assumptions on the distribution of y whose parameters depend on $\boldsymbol{x}^T\boldsymbol{\beta}$, allowing us to break away from the assumption of additive, zero mean errors model. Nonetheless, we can still recover the normal linear regression model by letting $y \sim \mathcal{N}(\mu, \sigma^2)$, with an identity link $l(\mu) = \mu = \boldsymbol{x}^T\boldsymbol{\beta}$.

In summary, in any GLM, the distribution of the response variable y defines an exponential family with θ that depends on $\boldsymbol{x}$ and $\boldsymbol{\beta}$ and that controls the mean of the response, and with ϕ that controls the variance. The explanatory variables $\boldsymbol{x}$ are considered known and fixed; the dispersion parameter ϕ is known in certain distributions such as Bernoulli and Poisson, but unknown and to be estimated in other distributions such as Gaussian and gamma; regression coefficients $\boldsymbol{\beta}$ are parameters to be estimated. Although GLMs are flexible, they still have limitations:

- the distribution of the response must define an exponential family, meaning that distributions not satisfying this are excluded, such as the Student distribution, the hypergeometric distribution and the log-normal distribution;
- the predicted value is linear in the explanatory variables, up to a transformation, i.e. $l(\hat{\mu}) = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$;
- the estimation is sensitive to outliers if no modification to the model or the estimation procedure is applied.

The latter issue is the focus of our document.

1.2. Gamma GLMs

In this document, we are particularly interested in gamma GLMs, as this is a model commonly used in actuarial science, especially in insurance pricing. We will show through a

data analysis in the next chapter that it is non-robust to outliers. We will propose a slight modification to the model that will allow to gain significantly in robustness. The approach will be seen to be valid for other GLMs with distributions having tails, such as inverse Gaussian and Poisson GLMs.

Recall the PDF of a gamma distribution:

$$f_{\alpha,\beta}(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y), \quad y > 0,$$

where $\alpha > 0$ and $\beta > 0$ are shape and rate parameters, respectively, and Γ is the gamma function. We now rewrite the gamma PDF using the reparametrization $\mu = \alpha/\beta$ and $\nu = \alpha$:

$$f_{\mu,\nu}(y) = \exp\left\{\frac{-y/\mu - \log\mu}{1/\nu} + (\nu-1)\log y + \nu\log\nu - \log(\Gamma(\nu))\right\}. \tag{1.2.1}$$

We thus have that the gamma distribution defines an exponential family with:

$$\theta = \frac{-1}{\mu}, \quad b(\theta) = \log\left(-\frac{1}{\theta}\right), \quad \phi = a(\phi) = \frac{1}{\nu},$$

$$c(y,\phi) = \left(\frac{1}{\phi} - 1\right)\log y + \frac{1}{\phi}\log\left(\frac{1}{\phi}\right) - \log\left(\Gamma\left(\frac{1}{\phi}\right)\right).$$

With this distribution, $\mathbb{E}[y] = b'(\theta) = \mu$ and $\mathbb{V}ar[y] = a(\phi)b''(\theta) = \mu^2/\nu$. We notice that the standard deviation of a random variable with a gamma distribution is proportional to its mean (equal, up to a factor $1/\sqrt{\nu}$).

Information carried by the explanatory variables are incorporated in the model by setting $l(\mu) = \eta = \boldsymbol{x}^T\boldsymbol{\beta}$. As mentioned previously, the logarithmic link $l(\mu) = \log(\mu)$ is more appropriate than the canonical link for gamma GLMs. The expected value with this link function is $\mu = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$. The gamma PDFs in form of (1.2.1) and the logarithmic link will be used throughout this document.

1.3. Estimation and Inference

In this section, we explain how to estimate GLMs from data and how to perform inference. We first present in Section 1.3.1 a statistical framework for GLMs, and then derive the likelihood function. In Section 1.3.2, we start with the likelihood estimating equations for the maximum likelihood estimators (MLEs), and proceed with frequentist estimation and inference. In particular, we present two numerical methods for finding the MLEs, and a type of residual which is essential for the robust approaches that will be used in the following chapters. Next, we present in Section 1.3.3 a Bayesian perspective for parameter estimation and inference. Finally, in Section 1.3.4, we discuss estimation for a special case of gamma GLMs in particular.

1.3.1. Statistical Framework and Likelihood Function

We consider that we have access to a data set of the form $(\boldsymbol{x}_i, y_i)_{i=1}^n$, where $y_1, ..., y_n$ are n independent realizations of the response, each associated with a given set of explanatory variable data points $\boldsymbol{x}_i$. With some abuse of notation, we denote by $y_1, \ldots, y_n$ the random variables as well. We want to analyse this data set by using a GLM (which can be any GLM, not necessarily a gamma GLM for now). The PDF or PMF associated to the GLM evaluated at y_i is thus such that:

$$f_{\boldsymbol{\beta},\phi}(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{1.3.1}$$

where $\theta_i = (b')^{-1} \circ l^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})$ (recall that for an exponential family, $\mu = b'(\theta)$ and $l(\mu) = \boldsymbol{x}^T\boldsymbol{\beta}$, using the notation of Section 1.1), and $a_i(\phi) = \phi$ or $a_i(\phi) = \phi/\omega_i$ where w_i is considered to be a known weight. Note that in the first case, $a_i(\phi)$ can be viewed as being equal to ϕ/ω_i but with $\omega_i = 1$. We define the i-th linear predictor as $\eta_i := \boldsymbol{x}_i^T\boldsymbol{\beta}$, and the expected value as $\mu_i = l^{-1}(\eta_i)$.

The likelihood and log-likelihood functions are central to both frequentist and Bayesian estimation. We thus present these functions here and then present frequentist and Bayesian estimation in Section 1.3.2 and Section 1.3.3, respectively. The likelihood function is defined as

$$L(\boldsymbol{\beta}, \phi) = \prod_{i=1}^n f_{\boldsymbol{\beta},\phi}(y_i) = \prod_{i=1}^n \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}.$$

In practice, it is often more convenient to work with the log-likelihood function instead of directly with the likelihood function. Indeed, by taking the log-likelihood, we end up with a sum of terms which allows to calculate derivatives more easily than a product of terms. Note that maximizing the likelihood is equivalent to maximizing the log-likelihood, given that the log function is monotonically increasing. Another reason why we work with the log-likelihood instead of the likelihood is that with many observations, the likelihood can become extremely small (or large) such that we will run out of the floating point precision very quickly, yielding easily an underflow (or overflow). A problem of underflow or overflow occurs when a number becomes too small or too large to be processed or stored in allocated space correctly in the computer, meaning that significant rounding errors are introduced. This problem is alleviated by working on the log scale. The log-likelihood function is defined as

$$\ell(\boldsymbol{\beta}, \phi) = \log\left(L(\boldsymbol{\beta}, \phi)\right) = \sum_{i=1}^n \left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\} = \sum_{i=1}^n \ell_i(\boldsymbol{\beta}, \phi), \tag{1.3.2}$$

where $\ell_i(\boldsymbol{\beta}, \phi)$ is the contribution of one data point to the likelihood, i.e.

$$\ell_i(\boldsymbol{\beta}, \phi) := \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi).$$

If the link function is canonical, an advantage is that a sufficient statistic for $\boldsymbol{\beta}$ exists. Indeed, in this case, $\theta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, and when $a(\phi)$ is a fixed constant, the part of the log-likelihood involving both the data and the model parameters is

$$\sum_{i=1}^{n} y_i \left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) = \sum_{i=1}^{n} y_i \left(\sum_{j=1}^{p} \beta_j x_{ij}\right) = \sum_{j=1}^{p} \beta_j \left(\sum_{i=1}^{n} y_i x_{ij}\right).$$

The sufficient statistic for $\{\beta_1, \ldots, \beta_j\}$ is thus $\{\sum_{i=1}^n y_i x_{ij}, j = 1, \ldots, p\}$. For the purpose of estimating $\boldsymbol{\beta}$, the sufficient statistic contains all relevant information.

1.3.2. Frequentist Estimation

Frequentist estimation is commonly performed using the maximum likelihood method. In order to maximize the likelihood, or equivalently, the log-likelihood, we can obtain *likelihood estimating equations* by taking the derivative of the log-likelihood with respect to $\boldsymbol{\beta}$ and ϕ if the log-likelihood function is differentiable in $\boldsymbol{\beta}$ and ϕ, and by setting these partial derivatives to 0. Typically with GLMs, the log-likelihood is strictly concave, implying that the identified root of each likelihood estimating equation yields indeed a global maximum. These equations regarding $\boldsymbol{\beta}$ are defined as

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = 0, \quad j = 1, \ldots, p. \tag{1.3.3}$$

To differentiate ℓ_i, we use the chain rule,

$$\frac{\partial \ell_i(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \frac{\partial \ell_i(\boldsymbol{\beta}, \phi)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

Since we have

$$\begin{aligned}
\frac{\partial \ell_i(\boldsymbol{\beta}, \phi)}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)}, \\
\frac{\partial \theta_i}{\partial \mu_i} &= \frac{1}{b''(\theta_i)} = \frac{a_i(\phi)}{[y_i]}, \\
\frac{\partial \mu_i}{\partial \eta_i} &= \frac{1}{l'(\mu_i)}, \\
\frac{\partial \eta_i}{\partial \beta_j} &= x_{ij},
\end{aligned}$$

the equation (1.3.3) becomes

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\mathbb{V}ar[y_i]l'(\mu_i)} = 0, \quad j = 1, \ldots, p. \tag{1.3.4}$$

Recall that $\mathbb{V}ar[y_i] = a_i(\phi)v(\mu_i)$. As mentioned, $a_i(\phi)$ is typically ϕ/w_i and we will consider that it is the case here to simplify the explanation. As ϕ does not depend on i, the dispersion parameter can be cancelled out in these estimating likelihood equations, thus will not influence the estimation of $\boldsymbol{\beta}$. More precisely, regardless of the value of ϕ, the likelihood is maximized at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ which is the solution of (1.3.4). Therefore, to jointly maximize $\ell(\boldsymbol{\beta}, \phi)$ with respect to $\boldsymbol{\beta}$ and ϕ, we can first find $\hat{\boldsymbol{\beta}}$ and then maximize $\ell(\hat{\boldsymbol{\beta}}, \phi)$ with respect to ϕ.

The solution of $\boldsymbol{\beta}$ has an analytic form in very particular cases, such as a Gaussian distribution with an identity link, which corresponds to the OLS solution for classical linear regression. For the rest not having an analytic solution, numerical methods are commonly used to solve the equation (1.3.4) and provide estimation of model parameters. There are two numerical methods with iterative process that are classical in frequentist estimation for GLMs. They are Newton–Raphson method and Fisher scoring method. The latter is the most commonly used for GLMs estimation. It is, for instance, used in R (R Core Team, 2021), command glm. It can be seen as a modification of Newton–Raphson method. We now briefly describe Newton–Raphson method and then explain how Fisher scoring differs from it.

The *Newton–Raphson method* is an iterative method for solving nonlinear equations. One of its applications is to find the maximum of a function, as in our case of finding MLE for $\boldsymbol{\beta}$ in GLMs. It begins with an initial approximation $\boldsymbol{\beta}^{(0)}$ for the solution, then it obtains a quadratic approximation by approximating the function in a neighbourhood of the initial approximation by a Taylor polynomial of second-degree. Next, it finds the location of that polynomial's maximum value, which becomes the initial point $\boldsymbol{\beta}^{(1)}$ for the next iteration. This step is repeated until the sequence of approximations converges.

In the context of GLMs, the log-likelihood approximated by the quadratic polynomial at the t-th iteration is given by

$$\ell\left(\boldsymbol{\beta}^{(t+1)}\right) \approx \ell\left(\boldsymbol{\beta}^{(t)}\right) + \boldsymbol{U}\left(\boldsymbol{\beta}^{(t)}\right)\left(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\right) + \frac{1}{2}\left(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\right)^T \boldsymbol{H}\left(\boldsymbol{\beta}^{(t)}\right)\left(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\right), \tag{1.3.5}$$

where $\boldsymbol{U}(\boldsymbol{\beta}^{(t)})$ is the *score vector* evaluated at $\boldsymbol{\beta}^{(t)}$, and $\boldsymbol{H}(\boldsymbol{\beta}^{(t)})$ is the *Hessian matrix* evaluated at $\boldsymbol{\beta}^{(t)}$. We wrote $\ell(\boldsymbol{\beta})$, instead of $\ell(\boldsymbol{\beta}, \phi)$ to simplify the notation, and also because here we consider that ϕ is fixed to an arbitrary value that does not influence the outcome of the optimization process. The score vector is the gradient of the log-likelihood function with respect to the parameter vector. In our context, it is defined as

$$\boldsymbol{U}(\boldsymbol{\beta}) = \left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1}, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_2}, \ldots, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_p}\right)^T.$$

The Hessian matrix is a square matrix of second-order partial derivatives with respect to the parameter vector. In our context, it is defined as

$$\boldsymbol{H}(\boldsymbol{\beta}) = \begin{bmatrix} \dfrac{\partial \ell^2(\boldsymbol{\beta})}{\partial \beta_1^2} & \dfrac{\partial \ell^2(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_2} & \cdots & \dfrac{\partial \ell^2(\boldsymbol{\beta})}{\partial \beta_1 \beta_p} \\ \dfrac{\partial \ell^2(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_1} & \dfrac{\partial \ell^2(\boldsymbol{\beta})}{\partial \beta_2^2} & \cdots & \dfrac{\partial \ell^2(\boldsymbol{\beta})}{\partial \beta_2 \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial \ell^2(\boldsymbol{\beta})}{\partial \beta_p \beta_1} & \dfrac{\partial \ell^2(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_2} & \cdots & \dfrac{\partial \ell^2(\boldsymbol{\beta})}{\partial \beta_p^2} \end{bmatrix}.$$

To maximize the equation (1.3.5), we differentiate with respect to $\boldsymbol{\beta}^{(t+1)}$ and set the derivative equal to 0. We obtain $\boldsymbol{U}\left(\boldsymbol{\beta}^{(t)}\right) + \boldsymbol{H}\left(\boldsymbol{\beta}^{(t)}\right)\left(\hat{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{\beta}^{(t)}\right) = 0$. If $\boldsymbol{H}\left(\boldsymbol{\beta}^{(t)}\right)$ is invertible, the solution for $\boldsymbol{\beta}^{(t+1)}$ is thus

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\boldsymbol{H}\left(\boldsymbol{\beta}^{(t)}\right)\right]^{-1} \boldsymbol{U}\left(\boldsymbol{\beta}^{(t)}\right).$$

The expression above is the iterative formula for Newton–Raphson method. We consider that the algorithm has converged when $||\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}|| < \delta$ or $\left|\left|\boldsymbol{U}\left(\boldsymbol{\beta}^{(t)}\right)\right|\right| < \delta$, where δ is a stopping criterion chosen by users. If the Hessian matrix is invertible, which is typically the case with GLMs, the estimates given by Newton–Raphson method converge to the MLE.

Another method for solving likelihood equations is called the *Fisher scoring method*. Sometimes, the calculation of the Hessian matrix can be quite complicated in Newton–Raphson method. The idea of the Fisher scoring method is to use the so-called *Fisher information* (or *expected information*) matrix to replace the Hessian matrix, when the former is easier to calculate. It is easier to calculate the Fisher information in the context of GLMs. Let us define the Fisher information for $\boldsymbol{\beta}$, denoted by $\boldsymbol{I}(\boldsymbol{\beta})$, with entries

$$\begin{aligned} I_{jk}(\boldsymbol{\beta}) &= \mathbb{E}\left[\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j}\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_k}\right] = \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j}\right)\left(\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_k}\right)\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{y_i - \mu_i}{\mathbb{V}ar[y_i]}\right)^2 \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 x_{ij}x_{ik}\right] = \sum_{i=1}^{n}\left\{x_{ij}x_{ik}\left(\frac{1}{\mathbb{V}ar[y_i]}\right)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right\}. \end{aligned} \tag{1.3.6}$$

With GLMs, the Fisher information is equal to the expected value of the *observed information*, which is the negative of the Hessian matrix. This is in fact the case for statistical models that are considered to be regular enough (i.e. that satisfy some regularity conditions), which is the case for GLMs.

As shown in (1.3.6), $\boldsymbol{I}(\boldsymbol{\beta})$ can be written as $\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{X}$, where $\boldsymbol{X}$ is the design matrix, and $\boldsymbol{W}(\boldsymbol{\beta})$ is a $n \times n$ diagonal matrix with diagonal elements $W_{ii}(\boldsymbol{\beta}) =$

$\left(\frac{1}{\mathbb{V}ar[y_i]}\right)\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$. The iterative formula for Fisher scoring method is

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\boldsymbol{I}(\boldsymbol{\beta}^{(t)})\right]^{-1} \boldsymbol{U}\left(\boldsymbol{\beta}^{(t)}\right). \tag{1.3.7}$$

The Fisher information, defined as $\boldsymbol{I}(\boldsymbol{\beta}) = \mathbb{E}\left[\left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T\right]$ is positive semi-definite by design. If the design matrix is of full rank, and $\partial \mu_i / \partial \eta_i$ is non-null for all i, which are typically the case with GLMs, the Fisher information is invertible, so that the update in (1.3.7) is valid. If the matrix is invertible, the estimates given by Fisher scoring method converge to the MLE.

Moreover, if the canonical link is used, the observed information is equal to the expected information, which means Newton–Raphson method is exactly the same as Fisher Scoring method. Indeed, when this link is used, $\theta_i = \eta_i$, then

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) = \frac{\mathbb{V}ar[y_i]}{a(\phi)}.$$

Hence, $W_{ii}(\boldsymbol{\beta}) = \frac{\mathbb{V}ar[y_i]}{a(\phi)^2}$. Regarding the observed information, we observe that

$$\frac{\partial^2 \ell_i(\boldsymbol{\beta}, \phi)}{\partial \beta_j \partial \beta_k} = \frac{\partial}{\partial \beta_k}\left(\frac{y_i - \mu_i}{a(\phi)} x_{ij}\right) = \frac{x_{ij}}{a(\phi)}\left(-\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_k}\right) = -\frac{\mathbb{V}ar[y_i]}{a(\phi)^2} x_{ij} x_{ik} = -I_{jk}(\boldsymbol{\beta}).$$

Now that we have ways of computing a point estimate of $\boldsymbol{\beta}$, let us discuss interval estimation. The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ has an asymptotic normal distribution. Assuming that the GLM employed is well specified with a true coefficient vector $\boldsymbol{\beta}^*$ and that n is large enough, the distribution of $\hat{\boldsymbol{\beta}}$ is approximately a normal with a mean of $\boldsymbol{\beta}^*$ and a covariance of $\boldsymbol{I}(\hat{\boldsymbol{\beta}})^{-1}$. We are then able to construct an approximate confidence interval, also called the *Wald confidence interval*, with a confidence level α for β_j^*, $j = 1, \ldots, p$, by

$$\beta_j^* \in \left[\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\left[\boldsymbol{I}(\hat{\boldsymbol{\beta}})^{-1}\right]_{jj}}\right],$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

Let us now talk about the estimation of the dispersion parameter ϕ. Since the likelihood estimating equation with respect to ϕ is different for every distribution defining an exponential family, we are going to take gamma GLMs as an example. The log-likelihood function for gamma GLMs is defined as

$$\sum_{i=1}^{n}\left\{\frac{-y_i/\mu_i - \log \mu_i}{1/\nu} + (\nu - 1)\log y_i + \nu \log \nu - \log(\Gamma(\nu))\right\}. \tag{1.3.8}$$

We derive the log-likelihood function with respect to ν, which is the inverse of the dispersion parameter ϕ, and set the partial derivative to 0. We thus write the log-likelihood as a

function of ν instead of a function of ϕ here. Let us consider that we already found $\hat{\boldsymbol{\beta}}$ which maximizes the log-likelihood (regardless of the value of ν). The estimating likelihood equation with $\hat{\mu}_i = \exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$ is thus given by

$$\frac{\partial \ell(\hat{\boldsymbol{\beta}}, \nu)}{\partial \nu} = \sum_{i=1}^{n} \left(\frac{\hat{\mu}_i - y_i}{\hat{\mu}_i} - \log(\hat{\mu}_i) + \log(y_i) + \log(\nu) - \frac{\Gamma(\nu)'}{\Gamma(\nu)} \right) = 0,$$

which is equivalent to

$$2n \left(\log(\nu) - \frac{\Gamma(\nu)'}{\Gamma(\nu)} \right) = 2 \sum_{i=1}^{n} \left(\log\left(\frac{\hat{\mu}_i}{y_i} \right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right) = D(y, \hat{\mu}), \tag{1.3.9}$$

where $D(y, \hat{\mu})$ is called the *deviance*. Deviance is a measure of goodness of fit of data to the model; the greater the deviance, the poorer the fit. The value of ν which maximizes the likelihood is also the solution of the equation (1.3.9) concerned with the deviance.

The principal problem with maximum likelihood estimation of ϕ in gamma GLMs is that it is sensitive to rounding errors of small observation values due to divisions by y_i (as seen in (1.3.9) above). As a result, we prefer to use another method for estimating ϕ, called *Pearson method*, which is the method applied in R, command glm.

For this method, we use the *Pearson statistic*, defined as

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)\phi^*/w_i}, \tag{1.3.10}$$

where ϕ^* is the true value of the dispersion parameter, assuming that the model is well specified. In this case, the Pearson statistic has a distribution which is asymptotically equivalent to a chi-squared distribution with $n - p$ degrees of freedom, in the limit $n \to \infty$. The Pearson method for estimating ϕ is based on the Pearson statistic. This estimator is consistent and given by

$$\hat{\phi} = \sum_{i=1}^{n} \frac{w_i (y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)(n - p)}, \tag{1.3.11}$$

which can be derived by noticing that a chi-squared random variable with $n - p$ degrees of freedom, divided by $n - p$, converges in probability to 1.

Let us now present an important type of residuals, called *Pearson residual*, which is one of the most commonly used class of residuals for GLMs. It will also be involved in different robust approaches for GLMs in the next chapters. The Pearson residual is defined as

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\mathbb{V}ar[y_i]}}}. \tag{1.3.12}$$

To avoid ambiguity, it is necessary to mention that the Pearson residual in R is defined as

$$\tilde{r}_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)/w_i}},$$

which is similar to r_i, the difference being that it is not divided by $\sqrt{\hat{\phi}}$. We can consider $\tilde{r}_i$ as a non-standardized version of the Pearson residual. In this document, we use all along the standardized version, i.e. r_i, as the definition of Pearson residual.

1.3.3. Bayesian Estimation

From a Bayesian perspective, parameters $\boldsymbol{\beta}$ and ϕ are random variables. Let π be the *prior distribution*, i.e. the distribution that is assumed on the parameters before having collected the data. The response is assumed to have the same distribution as in (1.1.1), but it is here considered as a conditional distribution given $\boldsymbol{\beta}$ and ϕ. The random variables $y_1, \ldots, y_n$ are assumed to be independent as before, but conditionally on the parameters. The explanatory variables are assumed to be fixed and known as before.

Bayesian estimation and inference rest upon the *posterior distribution*, which is the conditional distribution of the parameters given the observed data set:

$$\pi(\boldsymbol{\beta}, \phi \mid \mathbf{y}) = \frac{1}{m(\mathbf{y})} \pi(\boldsymbol{\beta}, \phi) \prod_{i=1}^{n} f_{\boldsymbol{\beta},\phi}(y_i), \tag{1.3.13}$$

where $\mathbf{y} := (y_1, \ldots, y_n)^T$, and $m(\mathbf{y})$ is the marginal density evaluated at $\mathbf{y}$:

$$m(\mathbf{y}) = \iint \pi(\boldsymbol{\beta}, \phi) \prod_{i=1}^{n} f_{\boldsymbol{\beta},\phi}(y_i) \, \mathrm{d}\phi \, \mathrm{d}\boldsymbol{\beta}.$$

Note that the product term in (1.3.13) corresponds to the likelihood function, but here it is considered, as mentioned, as the joint conditional density of $(y_1, \ldots, y_n)^T = \mathbf{y}$, given parameters $\boldsymbol{\beta}$ and ϕ.

Several types of prior distribution are employed in practice. A *subjective prior* reflects a prior opinion about the plausible parameter values, whereas an *objective prior* is relatively uninformative, implying that the data will have a more important impact on the resulting inference. Regarding subjective priors for parameters in GLMs, Bedrick et al. (1996) consider conditional means priors, which are priors for the mean of potential observations given the explanatory variables, i.e. priors on μ_i. The idea is to assign a prior distribution to a vector $(\mu_1, \ldots, \mu_p) = (l^{-1}(\boldsymbol{x}_1^T\boldsymbol{\beta}), ..., l^{-1}(\boldsymbol{x}_p^T\boldsymbol{\beta}))$, which is random through $\boldsymbol{\beta}$, and to identify a prior on $\boldsymbol{\beta}$ through a smooth one-to-one transformation. In practice, it is assumed that we have access to p linearly independent explanatory variables, that components of the vector $(\mu_1, \ldots, \mu_p)$ are independent, and that each element is assigned a PDF. Then, we can deduce a prior on $\boldsymbol{\beta}$ by performing a change of variable. The method is useful when it is more practical to assign a prior to conditional means of observations given explanatory variables, instead than directly to the coefficients. Bedrick et al. (1996) suggest for instance to assign an inverse-gamma distribution as prior to μ_i for gamma GLMs, conditionally on ϕ. We then specify a prior on ϕ, such as an inverse-gamma distribution with parameters that reflect opinions on the response variance, in order to create a joint prior for $\boldsymbol{\beta}$ and ϕ.

Regarding objective priors, Ibrahim and Laud (1991) propose to use a *Jeffreys's prior*, whose density is proportional to the square root of the determinant of the Fisher information matrix. It is shown that, under certain conditions, the posterior moment generating function of $\boldsymbol{\beta}$ exists for any GLM, in the case where ϕ is known. For example, under the log link for gamma GLMs, the Jeffreys's prior for $\boldsymbol{\beta}$ reduces to a uniform prior on $\mathbb{R}^p$, and all posterior moments are finite under this link. In our numerical experiments, we focus on point estimation and maximum likelihood estimation for simplicity. Note that the MLE corresponds to the maximum a posteriori estimator (MAP) when the prior is set to be a uniform on the whole parameter space. In Chapter 5, we present a result stating that the posterior distribution is proper for gamma GLMs with a log link in the case where $\phi = 1/\nu$ is considered unknown and $n \geq p$, under weak assumptions on the prior distribution. Even though a proper posterior distribution is required to perform inference under the Bayesian paradigm, theoretical guarantees that it is the case are scarce.

Even though we focus on maximum likelihood/a posteriori estimation in our numerical experiments, we now briefly describe how other estimates such as posterior means and credible intervals can be computed for Bayesian GLMs. The posterior distribution defined in (1.3.13) is typically intractable. This implies that we have to resort to numerical methods, such as Markov chain Monte Carlo (MCMC) methods to compute integrals with respect to the posterior distribution. Hamiltonian Monte Carlo (Duane et al., 1987) can for instance be employed. This sampling algorithm can also be used to sample from the posterior distribution resulting from the robust approach presented in this document.

1.3.4. Gamma GLMs Estimation

Compared to the previous sections, we here consider a special case of gamma GLMs to provide details about the estimation procedure. In particular, we present the likelihood estimations, and the Hessian matrix and the Fisher information that allow to solve those equations.

Recall that in the context of gamma GLMs with a logarithmic link, $\mu_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$ and $\phi = 1/\nu$. The log-likelihood function defined in (1.3.8) is thus

$$\ell(\boldsymbol{\beta}, \nu) = \sum_{i=1}^{n} \left\{ \left[\frac{-y_i}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})} - \boldsymbol{x}_i^T\boldsymbol{\beta} \right] \nu + (\nu - 1)\log y_i + \nu \log \nu - \log(\Gamma(\nu)) \right\}.$$

This is the function that one maximizes for maximum likelihood estimation. This is also the function that one uses, after applying the exponential function, to define a posterior density for Bayesian inference (see (1.3.13)). The likelihood estimating equations regarding $\boldsymbol{\beta}$ (recall (1.3.4)) are

$$\frac{\partial \ell(\boldsymbol{\beta}, \nu)}{\partial \beta_j} = \sum_{i=1}^{n} \nu \left(\frac{y_i - \mu_i}{\mu_i} \right) x_{ij} = \sum_{i=1}^{n} \nu \left(\frac{y_i - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})} \right) x_{ij} = 0, \quad j = 1, \ldots, p, \tag{1.3.14}$$

because $\mathbb{V}ar[y_i] = \mu_i^2/\nu$ and $l'(\mu_i) = 1/\mu_i$.

As mentioned in Section 1.3.2, to solve these estimating equations, one can use Newton–Raphson or Fisher scoring method. The former uses the Hessian matrix and the latter the Fisher information. When the log link is used, the elements of the Hessian matrix evaluated at $\boldsymbol{\beta}$ are given by

$$H_{jk}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial\beta_j \partial\beta_k} = \nu \sum_{i=1}^{n} \frac{\partial}{\partial\beta_k} \left[\left(\frac{y_i - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})} \right) x_{ij} \right] = -\nu \sum_{i=1}^{n} \left(\frac{y_i}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})} \right) x_{ij}x_{ik}. \tag{1.3.15}$$

The Fisher information evaluated at $\boldsymbol{\beta}$ is given by

$$\boldsymbol{I}(\boldsymbol{\beta}) = \nu(\boldsymbol{X}^T\boldsymbol{X}),$$

since it is the minus of the expected value of the Hessian matrix in (1.3.15).

With the Fisher information, we can also find an approximate confidence interval for the true coefficient β_j^*, $j = 1, \ldots, p$, based on the asymptotic normality of MLE:

$$\beta_j^* \in \left[\hat{\beta}_j \pm z_{1-\alpha/2} \sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}/\hat{\nu}} \right]. \tag{1.3.16}$$

Based on the above expression, it is observed that the length of the confidence interval depends heavily on the estimate of ν. As we will see in Section 2.2, outliers in the data set may have a significant impact on the estimation of ν yielding a smaller estimate compared to the estimate without the outliers, which implies an overly large confidence interval. The same is true for Bayesian credible intervals.

Regarding the estimation of ν, the estimator based on the Pearson method (recall (1.3.11)) is

$$\hat{\phi} = \frac{1}{\hat{\nu}} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2} = \frac{1}{n-p} \sum_{i=1}^{n} \left(\frac{y_i - \exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})}{\exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})} \right)^2,$$

using that $w_i = 1$ and $v(\hat{\mu}_i) = \hat{\mu}_i^2$ in this case. Note that with gamma GLMs, the Pearson residuals are given by

$$r_i = \left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right) \sqrt{\hat{\nu}} = \left(\frac{y_i - \exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})}{\exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})} \right) \sqrt{\hat{\nu}}. \tag{1.3.17}$$

Chapter 2

Example of Health Care Expenditures

In this chapter, we present an analysis of a data set by using a gamma GLM. This presentation reveals clearly the issues of non-robustness against outliers mentioned in Introduction: an invalid estimation of parameters which leads to skewed inference, interpretation and predictions, as well as a masking effect which has a significant negative impact on the outlier detection. We first describe the data set in Section 2.1, and this is followed by parameter estimation and residual analysis based on a gamma GLM in Section 2.2. The non-robustness of gamma GLMs is revealed through the comparison with robust methods, which motivates the presentation of robust alternatives in the next chapters.

2.1. Data Description

The data set that will be analysed is about health care expenditures. It is known for containing outliers, and has been analysed by Marazzi and Yohai (2004) and Cantoni and Ronchetti (2006) to highlight the benefits of using robust statistical methods. This data set is about 100 patients hospitalized at the *Centre Hospitalier Universitaire Vaudois* in Lausanne, Switzerland for medical back problems during 1999. The goal is to model the response variable, which is the cost of stay in this hospital (`cost` in Swiss francs), using the following explanatory variables: length of stay (`los`, in days), age (`age`, in years), admission type (`adm`: 0, planned; 1, emergency), insurance type (`ins`: 0, regular; 1, private), sex (`sex`: 0, female; 1, male) and discharge destination (`dest`: 1, home; 0, another health institution). The data set is available in the package robmixglm (Beath, 2021) in R.

The average cost of stay is 11 126 Swiss francs, with a standard deviation equal to 7 981.35. The empirical distribution of this variable is highly right-skewed, as seen in Figure 2.1. This characteristic of the data motivates the use of a gamma GLM.

We now discuss the characteristics of the explanatory variables. The average length of stay is 12.20 days, with a standard deviation equal to 10.10. The empirical distribution of this variable is also right-skewed. In our analysis, we use this variable on log scale as did in Cantoni and Ronchetti (2006), which can help to correct the asymmetry of this variable

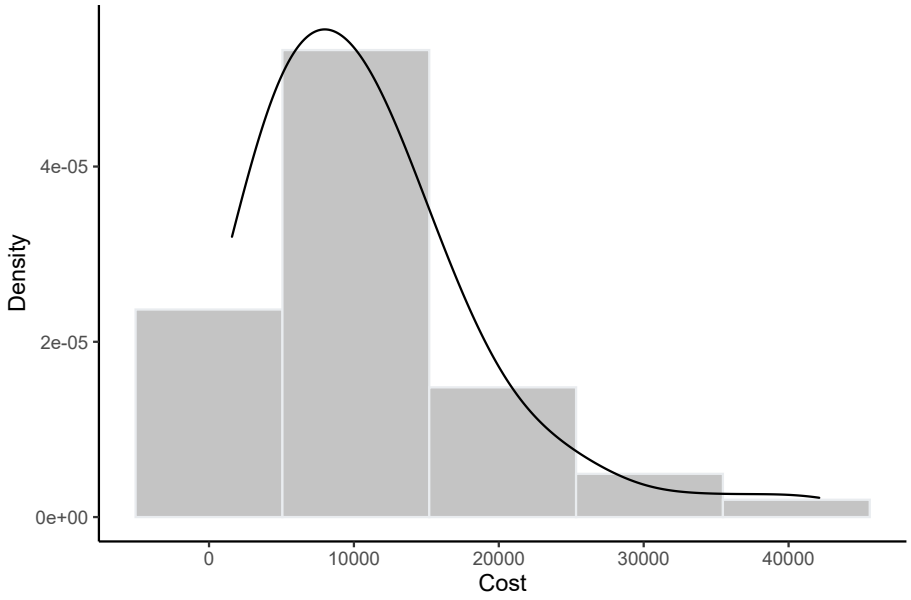


Figure 2.1. Histogram and a density estimate for the cost of stay (in Swiss francs)

and to yield a better fit of the model. The average age of patients is 57.62 years, with a standard deviation equal to 19.96; the youngest patient is 16 years old and the oldest is 93 years old. This variable is distributed symmetrically. Regarding the categorical explanatory variables, both sexes are well represented in the sample with 53 men and 47 women; the administration type is also quite balanced with 40 planned and 60 emergency. However, regarding the insurance type, only 9 patients out of 100 have private insurance. Regarding the discharge destination, 82 patients out of 100 go home after being treated, and the others go to another hospital.

Analysing such a data set is of interest for actuaries. It can, for instance, help them understand the main contributing factors, in this case, to the health cost, and provide accurate insurance pricing.

2.2. Analysis and Non-robustness Problems

To analyse the data set of health care expenditure, we use a gamma GLM with a logarithmic link. The model uses all explanatory variables and is such that

$$\mathbb{E}[\texttt{cost}] = \exp\left[\beta_1 + \beta_2 \log(\texttt{los}) + \beta_3\, \texttt{age} + \beta_4\, \texttt{adm} + \beta_5\, \texttt{ins} + \beta_6\, \texttt{sex} + \beta_7\, \texttt{dest}\right],$$

where the variables $\log(\texttt{los})$ and $\texttt{age}$ have been standardized. We standardize the continuous variables to benefit from the appealing interpretation described in Section 1.1. Recall that the relation between the mean and the variance of the response is given through the dispersion parameter $\phi = 1/\nu$ with $\mathbb{V}ar[\texttt{cost}] = \mathbb{E}[\texttt{cost}]^2/\nu$.

To highlight the non-robustness of gamma GLMs, we present an analysis by using a gamma GLM and we compare the results with those obtained using two robust alternatives: the gamma version of the robust GLMs based on M-estimators proposed by Cantoni and Ronchetti (2001), which we refer to as the *robust M method*, and our proposed robust gamma GLM. As mentioned in Introduction, the robust M method is a frequentist method, which consists of an adaptation of robust M-estimators for linear regression. At the step of estimation, the likelihood estimating equations are modified in a way to assign low weight to outliers. Our proposed method consists instead of modifying directly the distribution of the response variable: the central part of the PDF is kept as is, while the extremities are replaced by log-Pareto tails. The latter two methods will be explained in detail in Chapter 3 and Chapter 5, respectively.

We present four estimations in Table 2.1: a gamma GLM with all observations, a gamma GLM without identified outliers, the robust M method, and our proposed method.

	all observations	**without outliers**	**robust M method**	**proposed method**
$\hat{\beta}_1$	9.00	9.04	9.02	9.03
$\hat{\beta}_2$	0.68	0.71	0.70	0.70
$\hat{\beta}_3$	-0.01	-0.03	-0.02	-0.02
$\hat{\beta}_4$	0.21	0.23	0.22	0.22
$\hat{\beta}_5$	0.09	-0.03	0.01	-0.01
$\hat{\beta}_6$	0.10	0.08	0.07	0.07
$\hat{\beta}_7$	-0.10	-0.14	-0.12	-0.12
$\hat{\nu}$	20.16	41.27	41.11	41.32

Table 2.1. Parameter estimates for a gamma GLM based on all observations, a gamma GLM based on the data set excluding identified outliers, the robust alternative of Cantoni and Ronchetti (2001), and our proposed method

As observed in Table 2.1, estimation based on a gamma GLM is sensitive to outliers. Let us examine that of the coefficient β_5 as an example, and more specifically, the interpretation that results from it. Based on the analysis with a gamma GLM with all observations, if all other explanatory variables are null, the estimated expected cost for a person who has a private insurance is the base rate $\exp(\hat{\beta}_1) = 8\,113.09$, adjusted by multiplying by $\exp(\hat{\beta}_5) = 1.10$. Based on the analysis with our proposed method, the value is instead equal to the base rate $\exp(\hat{\beta}_1) = 8\,316.95$, multiplied by $\exp(\hat{\beta}_5) = 0.98$. The estimated effect of a variable on the cost of stay with or without a robust method can be very different, or even opposite.

It is observed that the biggest difference in estimates is for ν. Its estimate with a gamma GLM is almost half of that with the robust M method or our proposed method. This estimate is involved in the estimation of the response variance and uncertainty regarding the plausible

values for $\boldsymbol{\beta}$, thus also credible/confidence interval lengths. For example, if we calculate a 95% approximate confidence interval for β_7 (recall (1.3.16)), the one obtained by using a gamma GLM with all observations is [-0.24; 0.03], which contains 0. It means that at 5% level, we cannot reject the hypothesis that the variable `dest` has no effect on the cost of stay. However, if we replace the estimates involved in the calculation of the confidence interval by those of our proposed method, we obtain [-0.21; -0.02], which does not contain 0 any longer. The conclusion made can be changed to a great extent depending on whether we take care of outliers contained in the data set.

Let us now take a look at the plot of Pearson residuals against the predicted values in Figure 2.2. The Pearson residual is the response residual, i.e. $y_i - \hat{\mu}_i$, standardized with the estimated standard deviation for the observation, i.e. $\hat{\mu}_i/\sqrt{\hat{\nu}}$. It is worth noting that an outlier is a couple $(\boldsymbol{x}_i, y_i)$ such that y_i is far from its fitted value $\hat{\mu}_i = l^{-1}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$ under the applied model that reflects the trend of the bulk of the data (the latter being a desideratum of a robust model). The observation y_i does not need to be extreme, nor $\boldsymbol{x}_i$, for $(\boldsymbol{x}_i, y_i)$ to be an outlier. The two just need to be incompatible, according to the estimated model. With GLMs, a couple is identified as an outlier if its absolute value of Pearson residual exceeds 3, as proposed by Ryan (1997), which matches with the 3σ distance rule used in the normal linear regression theory. Figure 2.2 below shows the Pearson residuals against the fitted values, where the Pearson residuals are computed based on gamma GLM estimation with all observations in the left panel and based on our robust model estimation in the right panel.

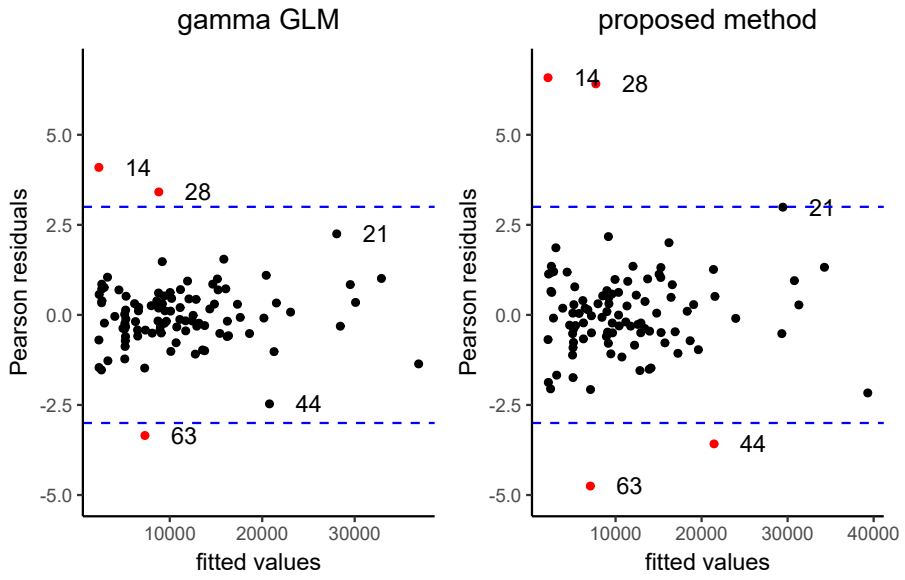


Figure 2.2. Pearson residuals with the gamma GLM and our proposed method

It is clearly observed that the Pearson residuals are overall more dispersed based on our robust model estimation than gamma GLM estimation with all observations, mainly due

to a higher $\hat{\nu}$ based on the robust model estimation. In particular, the absolute values of the identified outliers with a gamma GLM are higher with our method. For example, the 28th and 63rd Pearson residuals have values around 3.4 and -3.3 with a gamma GLMs, respectively, whereas their values are around 6.4 and -4.8 with the robust counterpart. Even more interestingly, $(\boldsymbol{x}_{44}, y_{44})$ is detected as an outlier using our approach, while it is not with the gamma GLM. The 21st couple has a value of Pearson residual equal to 2.99, which is close to the threshold beyond which couples are identified as outliers. There is an evident *masking effect* in outlier detection based on non-robust estimation. The model indeed adjusts itself for outliers which mask each other, because their residuals are distorted and appear to be less extreme than they should. The main reason is due to overestimation of the dispersion parameter, which is $1/\nu$ in our case, as presented in Table 2.1. The outlier detection based on a robust model is thus more effective: outliers do not mask each other and are effectively identified. Regarding the estimation presented in Table 2.1 for the gamma GLM without outliers, the outliers that we removed are the four outliers identified by using our proposed method.

In order to understand why these observations are extreme, let us look at the scatter plot of the response `cost` in function of log(`los`). We observe from Figure 2.3 that the outliers are data points with a relation between log(`los`) and `cost` that is significantly different from the rest of the data points. This observation appears to be coherent with the fact log(`los`) is the explanatory variable with the most impact on the calculation of Pearson residuals (recall Table 2.1). By looking at Figure 2.3, one might wonder why $(\boldsymbol{x}_{31}, y_{31})$ is not flagged as an outlier, even if it appears to be further away than $(\boldsymbol{x}_{14}, y_{14})$ to the trend of the bulk of the data. Even if $|y_{31} - \hat{\mu}_{31}| = 13\,581.4 > |y_{14} - \hat{\mu}_{14}| = 2\,314.8$, the fact that $\sqrt{\widehat{\mathbb{V}ar[y_{31}]}} = 6\,459.2$ is much larger than $\sqrt{\widehat{\mathbb{V}ar[y_{14}]}} = 351.9$ makes the 31st residual less extreme (the value is -2.1), reflecting that the model is such that the variance of the response increases with the mean. The 21st couple which is almost identified as an outlier is also relatively far away to the trend of the bulk of the data.

The analysis conducted in this section allows to show the problem of non-robustness of gamma GLMs. The estimates can heavily be influenced by outliers in the data set. Also, the outlier detection is not effective due to the masking effect under the non-robust estimation. This highly motivates us to find robust alternatives to tackle the problem of non-robustness of gamma GLMs.

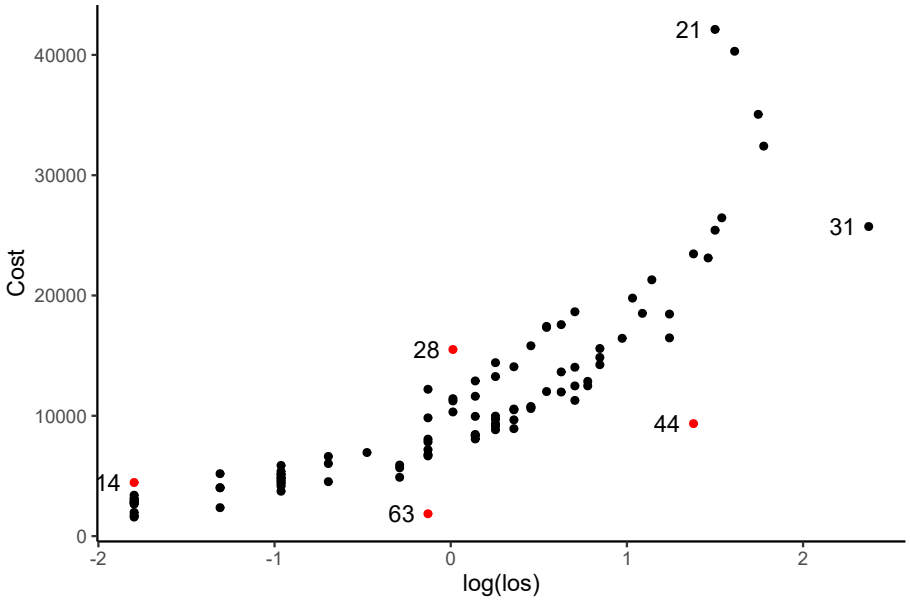


Figure 2.3. Points $(\boldsymbol{x}_{i2}, y_i)$

Chapter 3

Robust GLMs Based on M-estimators

The objective of this chapter is to present in more details the frequentist robust method for GLMs proposed by Cantoni and Ronchetti (2001), which we named previously the *robust M method* in Chapter 2. This method is applicable in general, meaning as long as the distribution of the response variable defines an exponential family, such as Poisson and gamma. As mentioned, we here are particularly interested in the case of gamma GLMs. In Section 3.1, we present the details of the robust M method for gamma GLMs, and provide a reason for the robustness of this method, by presenting the notion of the influence function. In Section 3.2, we present M-estimators in general with the goal of highlighting the connection between the robust M method (for gamma GLMs) and M-estimators, hence the name of the former.

3.1. Robust M Method for Gamma GLMs

In order to present the robust M method for gamma GLMs, it helps to recall the likelihood estimating equation with respect to $\boldsymbol{\beta}$:

$$\frac{\partial \ell(\boldsymbol{\beta}, \nu)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \nu \left(\frac{y_i - \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \right) \boldsymbol{x}_i = \sum_{i=1}^{n} \sqrt{\nu}\, r_i(\boldsymbol{\beta}, \phi) \boldsymbol{x}_i = \mathbf{0}, \tag{3.1.1}$$

where $r_i(\boldsymbol{\beta}, \phi) = (y_i - \mu_i)/\sqrt{\mathbb{V}ar[y_i]}$, which is equal to $\sqrt{\nu}\left(y_i - \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})\right) / \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$ in the case of gamma GLMs with a log link. The term $r_i(\boldsymbol{\beta}, \phi)$ can be viewed as an analogue of the Pearson residual (recall (1.3.12)). It is a function evaluated at parameters $\boldsymbol{\beta}$ and ϕ; the Pearson residual can thus be seen as $r_i = r_i(\hat{\boldsymbol{\beta}}, \hat{\phi})$. We reformulated the estimating equation by using the function $r_i(\boldsymbol{\beta}, \phi)$, because the robust M method that will be presented shortly consists of replacing $r_i(\boldsymbol{\beta}, \phi)$ by another function.

The solution of the above equation is the MLE for $\boldsymbol{\beta}$, but it is unfortunately not robust, as we have seen with the example in Section 2.2. To provide a reason for the non-robustness of gamma GLMs, we present the notion of *influence function*, which qualitatively measures the robustness of an estimator.

Proposed by Hampel (1974), the influence function is an important mathematical tool. It measures the effect of an infinitesimal change in one observation on an estimator. This measure is generally classified as a measure of qualitative robustness, because it is used to indicate whether an infinitesimal change results in a bounded effect on the function to optimize; a bounded influence function is a desirable robustness property. The influence function of an MLE is proportional to the *score function*, which is the contribution of one observation to the derivative of the log-likelihood with respect to the parameters. Therefore, if the score function is not bounded with respect to y_i and/or $\boldsymbol{x}_i$, it implies that an extreme observation in the response and/or in the explanatory variables can have a large impact on the estimation of parameters. The score function in the case of gamma GLMs with a log link is given by

$$\nu \left(\frac{y_i - \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \right) \boldsymbol{x}_i.$$

The score function is not bounded with respect to either y_i or $\boldsymbol{x}_i$, which provides a reason for the non-robustness of the MLE of $\boldsymbol{\beta}$.

To deal with the problem, Cantoni and Ronchetti (2001) propose a class of estimators that have bounded influence functions. They propose to modify directly the original likelihood estimating equation, or equivalently, the score function, so that the resulting equation is bounded with respect to y_i and $\boldsymbol{x_i}$. The robust estimator for $\boldsymbol{\beta}$ is obtained by solving the following equation:

$$\sum_{i=1}^{n} \left[\sqrt{\nu}\ \psi(r_i(\boldsymbol{\beta}, \phi), c)\ w(\boldsymbol{x}_i) \boldsymbol{x}_i - a(\boldsymbol{\beta}) \right] = \mathbf{0}, \tag{3.1.2}$$

where ψ is a function suggested by the authors, given by

$$\psi(r_i(\boldsymbol{\beta}, \phi), c) = \begin{cases} r_i(\boldsymbol{\beta}, \phi) & \text{if } |r_i(\boldsymbol{\beta}, \phi)| \leq c, \\ c\ \text{sign}(r_i(\boldsymbol{\beta}, \phi)) & \text{otherwise,} \end{cases} \tag{3.1.3}$$

$w(\boldsymbol{x}_i)$ is a weight function for $\boldsymbol{x}_i$ which helps to downweight high leverage points, $c > 0$ a tuning parameter, and $a(\boldsymbol{\beta}) = (1/n) \sum_{j=1}^{n} \mathbb{E} \left[\sqrt{\nu}\ \psi(r_j(\boldsymbol{\beta}, \phi), c)\ w(\boldsymbol{x}_j) \boldsymbol{x}_j \right]$ with the expectation taken with respect to the distribution of y_j, which is a gamma. The last term $a(\boldsymbol{\beta})$ is a correction term to ensure the Fisher consistency of the estimation of $\boldsymbol{\beta}$ (Cantoni and Ronchetti, 2001). In this context, the Fisher consistent estimator ensures that the expectation of the sum of the equivalent of the score functions for the robust M method, regarding $\boldsymbol{\beta}$, is equal to 0, i.e. $\sum_{i=1}^{n} \mathbb{E} \left[\sqrt{\nu}\ \psi(r_i(\boldsymbol{\beta}, \phi), c)\ w(\boldsymbol{x}_i) \boldsymbol{x}_i - a(\boldsymbol{\beta}) \right] = \mathbf{0}$.

The functions $w(\boldsymbol{x}_i)$ and $\psi(r_i(\boldsymbol{\beta}, \phi), c)$ are two new ingredients compared with the original likelihood estimating equation for gamma GLMs. When $w(\boldsymbol{x}_i) = 1$ and $\psi(r_i(\hat{\boldsymbol{\beta}}, \hat{\phi})) = r_i(\hat{\boldsymbol{\beta}}, \hat{\phi})$ for all i, with $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ being the solution to (3.1.2), $a(\hat{\boldsymbol{\beta}}) = 0$ because the expectation of the

score function under the gamma GLM is equal to 0. The estimator for $\boldsymbol{\beta}$ based on (3.1.2) coincides with the MLE in this case. Regarding the weight function $w(\boldsymbol{x}_i)$, the authors suggested several choices, such as $w(\boldsymbol{x}_i) = \sqrt{1-h_i}$, where h_i is the i-th diagonal element of the hat matrix $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, or a weight proportional to the inverse of the Mahalanobis distance, which is an effective distance metric where a covariance matrix is used to find the distance between data points and the centre. Regarding the function $\psi(r_i(\boldsymbol{\beta},\phi),c)$, it is an identity function if its first argument is between $-c$ and c, otherwise, ψ returns c times the sign of $r_i(\boldsymbol{\beta},\phi)$. With this choice of function ψ, combined with an appropriate weight function such as those mentioned, the influence function associated with the estimating equation (3.1.2) is bounded.

The function ψ can be also be viewed as $\psi(r_i(\boldsymbol{\beta},\phi),c) = \tilde{w}\left(r_i(\boldsymbol{\beta},\phi),c\right) r_i(\boldsymbol{\beta},\phi)$, where $\tilde{w}\left(r_i(\boldsymbol{\beta},\phi),c\right)$ is the weight of $r_i(\boldsymbol{\beta},\phi)$. With this form, (3.1.2) can be interpreted as an estimating equation weighted separately with respect to $r_i(\boldsymbol{\beta},\phi)$ and $\boldsymbol{x}_i$, and re-centred to ensure the Fisher consistency of the estimation for $\boldsymbol{\beta}$.

The tuning parameter c is typically chosen to reach a compromise between efficiency and robustness, where efficiency refers to the variance of the estimators in the context where the true model is the non-robust gamma GLM. If we let $c \to \infty$, we recover the MLE for gamma GLMs, which is the benchmark in terms of efficiency. In practice, we set the value of c between 1 and 2 to guarantee robustness with a reasonable level of efficiency (Cantoni and Ronchetti, 2006).

The solution of the estimating equation with respect to $\boldsymbol{\beta}$ in (3.1.2) now depends on ν through $r_i(\boldsymbol{\beta},\phi)$, because it cannot be put as a factor which multiplies the sum, as the case with the gamma GLM. The estimation of ν thus now has an impact on the estimation of $\boldsymbol{\beta}$. The proposed robust estimator for ν with the robust M method is the solution to

$$\sum_{i=1}^{n}\left(\psi^2(r_i(\boldsymbol{\beta},\phi),c) - \mathbb{E}\left[\psi^2(r_i(\boldsymbol{\beta},\nu),c)\right]\right) = 0, \tag{3.1.4}$$

where the expectation is to ensure the Fisher consistency of the estimation for ν. Ideally, (3.1.2) and (3.1.4) are to be solved simultaneously. However, in practice, we generally use a two-step procedure to solve alternately these two equations until convergence. Numerical approaches such as Newton–Raphson or Fisher scoring algorithms presented in Section 1.3.1 can also be used to solve (3.1.2) and (3.1.4) to provide robust parameter estimation. The correction terms of Fisher consistency in (3.1.2) and (3.1.4) should be computed explicitly. Details of all the computational aspects of the robust M method are clearly presented in the appendix of Cantoni and Ronchetti (2006).

In Chapter 2, we presented the estimation results for the health-care data set based on the robust M method. For the estimation, the tuning parameter c in the function ψ was set to 1.5, as in Cantoni and Ronchetti (2006). The weight function was set to 1, as in

Cantoni and Ronchetti (2006), meaning that all $\boldsymbol{x}_i$ have equal weight. With this choice, the contribution of a couple $(\boldsymbol{x}_i, y_i)$ to the estimating equation (3.1.2) is modified, compared to the gamma model, only if y_i is far enough from its fitted valued $\hat{\mu}_i = l^{-1}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$, or in other words, if its Pearson residual in absolute terms is large enough; it is not modified if $\boldsymbol{x}_i$ is extreme but the residual is not. The package robustbase (Maechler et al., 2020) in R, command glmrob allows to provide parameter estimates for the robust M method.

The parameter estimates were provided in Table 2.1. Comparatively to those of the gamma GLM based on the whole data set, the estimates of the robust M method are closer to the estimates of the gamma GLM without identified outliers, especially for the estimate $\hat{\nu}$. The Pearson residuals with the robust M method have similar values to those obtained with our proposed robust method (that will be explained in detail in Chapter 5): the same four couples (14th, 28th, 63rd, and 44th) are identified as outliers, and the 21st couple is close to being identified as an outlier.

3.2. Connection with M-estimators

We present M-estimators in this section, for the purpose of highlighting a connection between this class of estimators and the estimator associated with the robust M method. M-estimators are one of the most popular classes of estimators in frequentist robust estimation, and they play a crucial role in the development of modern robust statistics (Ronchetti, 2006). M-estimators can be considered as a generalization of MLEs. As explained in Section 1.3.2, we obtain MLEs by maximizing the log-likelihood function, or equivalently, by minimizing $\sum_{i=1}^{n} -\ell_i(\boldsymbol{\beta}, \phi)$, where $\ell_i(\boldsymbol{\beta}, \phi)$ is the contribution of the data point i to the log-likelihood. Huber (1964) proposed to view maximum likelihood estimation as a special case to a general estimation method:

$$\min_{\boldsymbol{\beta},\phi} \sum_{i=1}^{n} \rho(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi), \tag{3.2.1}$$

where ρ is a loss function to be chosen by the user. An estimator based on the minimization of ρ is called a *maximum likelihood type estimator (M-estimator)*, and it is seen to correspond to the MLE when $\rho = -\ell_i$, hence its name.

The development of most M-estimators is generally not based on well-defined and previously known model PDFs. However, in some cases, we can associate a distribution to a loss function of an M-estimator. The latter point will be explained in Chapter 4. The loss function ρ is typically chosen to be continuous, zero-symmetric for a function of its argument (think of the residuals in a linear-regression framework), and positive. It is also preferable for the function to be strictly convex in the parameters, as the strict convexity of ρ guarantees that there exists a unique and global solution to the minimization problem.

If ρ is differentiable with respect to $(\boldsymbol{\beta}, \phi)$, minimizing (3.2.1) can be performed by solving the following equation (the analogue of the likelihood estimating equation, see Section 1.3.2):

$$\sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) = \mathbf{0}, \tag{3.2.2}$$

where $\Psi(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) = \partial\rho(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi)/\partial(\boldsymbol{\beta}, \phi)$.

Many different choices for ρ are proposed in the literature. For an excellent reference about M-estimators and their properties of robustness and efficiency, we refer readers to Menezes et al. (2021), where 50 M-estimators are presented in the context of linear regression, including the weighted least squares estimator that is non-robust, the Huber M-estimator (Huber, 1973), and the Tukey-biweight M-estimator (Beaton and Tukey, 1974).

If we return to the robust M method in the case of gamma GLMs, recall that the estimating equation associated to $\boldsymbol{\beta}$ is given by

$$\sum_{i=1}^{n} \left[\sqrt{\nu}\ \psi(r_i(\boldsymbol{\beta}, \phi), c)\ w(\boldsymbol{x}_i)\boldsymbol{x}_i - a(\boldsymbol{\beta})\right] = \mathbf{0}.$$

The latter can be viewed as

$$-\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}}\ \rho(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) = \mathbf{0}$$

for a certain loss function ρ. It is in this sense that a connection exists in between the robust M method and M-estimators. We will see in Chapter 4 that several functions ρ can produce the same likelihood estimating equations. The approach is thus less elegant than what would follow from an M-estimator. Robust M-estimators are also more natural in robustness contexts, as they can be seen as a direct solution to a robustness problem arising with a model (through the likelihood function).

Chapter 4

Robust M-Estimators Viewed as Heavy-Tailed Distributions

In this chapter, we want to establish a connection between the robust M method presented in Chapter 3, and heavy-tailed distributions, which can be used for both Bayesian and frequentist robust analysis. However, as we have explained in Section 3.2, there is no one-to-one correspondence between the frequentist estimator and a PDF. Therefore, in Section 4.1, we take a step back by first presenting that a clear connection can be established between robust M-estimators and heavy-tailed distributions in the context of linear regression. We take the Huber M-estimator (Huber, 1973) as an example, and show that the distribution of the standardized error associated with this estimator is a modified normal PDF with tails that have been replaced by Laplace ones. In Section 4.2, we return to the context of gamma GLMs, and explain that several loss functions correspond to the modified estimating equations proposed by the robust M method. We choose one loss function in particular, then derive the corresponding density for the response variable, which is a gamma in the central part with heavier tails in the extremities. The right tail of this distribution has a polynomial decrease, whereas that of the gamma has an exponential decrease.

4.1. The Linear Regression Case

Consider that we have access to a data set of the form $(\boldsymbol{x}_i, y_i)_{i=1}^n$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{1p})^T \in \mathbb{R}^p$ represents a vector of explanatory variable data points, and $y_i \in \mathbb{R}$ represents an observation of the response variable, as defined previously in Section 1.3. The linear regression model is as follows:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma \epsilon_i, \quad i = 1, \ldots, n, \tag{4.1.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is the vector of regression coefficients, σ is a scale parameter, and $\epsilon_1, \ldots, \epsilon_n$ are standardized errors, which are assumed to be independent and identically distributed with $\epsilon_i \sim f$. In the normal linear regression model, $f = \mathcal{N}(0, 1)$. To find MLEs

for the parameters $\boldsymbol{\beta}$ and σ, we need to maximize the log-likelihood function, which is defined as

$$\ell(\boldsymbol{\beta}, \sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2. \tag{4.1.2}$$

Maximizing (4.1.2) with respect to $\boldsymbol{\beta}$ is equivalent to the minimization of the following loss function with respect to $\boldsymbol{\beta}$:

$$\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2. \tag{4.1.3}$$

Thus, the MLE for $\boldsymbol{\beta}$ in normal linear regression coincides with the OLS estimator, which consists of minimizing the sum of squared residuals. The scale parameter σ does not influence the estimation for $\boldsymbol{\beta}$, because σ is in fact a factor that multiplies the sum in (4.1.3). Therefore, the optimization problem can be solved by first finding $\boldsymbol{\beta}$ which minimizes the sum of squared residuals, and then by using the obtained value to find σ that maximizes (4.1.2), similarly as with the gamma GLM.

As mentioned in Section 3.2, a bounded influence function is a desirable property for an estimator when there is a potential presence of outliers in the data set. If we take the derivative of (4.1.2) with respect to $\boldsymbol{\beta}$, we obtain the score function associated with $\boldsymbol{\beta}$, which is equal to $\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)/\sigma^2$. This function is clearly not bounded with respect to either y_i or $\boldsymbol{x}_i$.

The idea of Huber (1973) was to modify the sum of squared standardized residuals to produce less extreme values when some residuals are extreme. The modified loss function in linear regression is often quadratic (or similar to a quadratic function) below a certain threshold, but then grows more slowly beyond that threshold. For example, the Huber M-estimator proposes to minimize

$$n\log(\sigma) + \frac{1}{2}\sum_{i=1}^{n}\varrho\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right),$$

and to use the *Huber loss function* (Huber, 1964), defined as

$$\varrho(\epsilon) = \begin{cases} \frac{1}{2}\epsilon^2 & \text{if } |\epsilon| \leq k, \\ k|\epsilon| - \frac{1}{2}k^2 & \text{otherwise,} \end{cases} \tag{4.1.4}$$

where k is a tuning parameter chosen by the user to reach a compromise between efficiency and robustness. In particular, $k = 1.345$ allows the estimator to produce 95-percent efficiency, meaning that, asymptotically, the variance of $\hat{\boldsymbol{\beta}}$ corresponds to that of the OLS estimator to which we add a factor of 1.05, if the true model is a normal linear regression. The penalization by the Huber loss function is quadratic, which is the same as in normal linear

regression, between $-k$ and k; otherwise, the penalization is linear, which is more moderate. Note that the term $-k^2/2$ in (4.1.4) is to ensure that ϱ is continuous.

If we take the derivative of ϱ with respect to $\boldsymbol{\beta}$, we obtain

$$\frac{\partial}{\partial \boldsymbol{\beta}} \varrho\left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma}\right) = \begin{cases} -\dfrac{\boldsymbol{x}_i^T (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})}{\sigma^2} & \text{if } \left|\dfrac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma}\right| \leq k, \\ -\dfrac{\boldsymbol{x}_i^T}{\sigma}\, k \text{ sign}\left(\dfrac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma}\right) & \text{otherwise,} \end{cases} \tag{4.1.5}$$

which is similar to the derivative with the robust M method (recall (3.1.2) and (3.1.3)). Note that the function in (4.1.5) is bounded with respect to the residual.

Figure 4.1 shows the functions ϱ associated with the OLS estimator and the Huber M-estimator with $k = 1.345$.

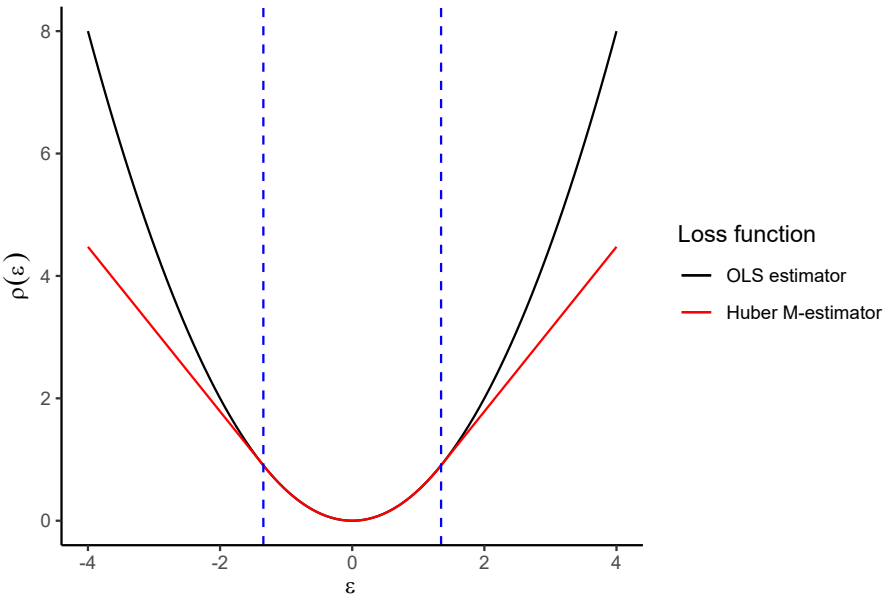


Figure 4.1. Loss functions $\varrho(\epsilon)$ associated with the OLS estimator and the Huber M-estimator ($k = 1.345$)

From a perspective of modelling, it is helpful to have a precise characterization of the model associated with robust M-estimators, if possible, i.e. to associate PDFs to the estimators, as with the MLE and the model in (4.1.1). With robust M-estimators, the associated PDFs are no longer normal distributions, since the likelihood is modified through the loss function. In order to establish a connection between robust M-estimators and heavy-tailed distributions, let us rewrite the log-likelihood for the linear regression model by using $f(\epsilon) = g(\epsilon)/m$, where m is a normalizing constant:

$$\ell(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^{n} \log\left(\frac{1}{\sigma} f\left(\frac{y_i - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right)\right) = -n\log(\sigma) + \sum_{i=1}^{n} \log\left(f\left(\frac{y_i - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right)\right)$$
$$= -n\log(\sigma) - n\log(m) + \sum_{i=1}^{n} \log\left(g\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)\right).$$

In the case of normal linear regression,

$$g = \exp\left\{-\frac{1}{2}\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)^2\right\} \quad \text{and} \quad m = \sqrt{2\pi}.$$

By presenting the log-likelihood of the regression model in this form, it is observed that for Huber M-estimator, the Huber loss function $\varrho(\epsilon)$ in (4.1.4) can be viewed as $-\log(g(\epsilon))$, where $g(\epsilon)$ is defined as

$$g(\epsilon) = \begin{cases} \exp\left(-\frac{1}{2}\epsilon^2\right) & \text{if } |\epsilon| \leq k, \\ \exp\left(-k|\epsilon| + \frac{1}{2}k^2\right) & \text{otherwise.} \end{cases}$$

After the normalization, the PDF associated with the Huber M-estimator is equal to $g(\epsilon)/m$, where $m = 2\exp(-k^2/2)/k + \sqrt{2\pi}(2\Phi(k) - 1)$ with $\Phi(k)$ the cumulative distribution function (CDF) of a standard normal distribution evaluated at k. The equation to minimize with the Huber M-estimator thus corresponds to a likelihood function, where $f(\epsilon)$ is such that the density in the central part is proportional to a standard normal distribution, and the tails of the density behave like $\exp(-k|\epsilon|)$, which corresponds to a Laplace distribution. As k increases, f approaches a standard normal, meaning that the mass of the heavy tails decreases. As we can see in Figure 4.2, the density has a slower decrease than a standard normal after the threshold k.

The Huber M-estimator is a proper example to present a connection between robust M-estimators and heavy-tailed distributions. However, not all robust M-estimators have a clear correspondence with a model. For example, it is not possible to find a model associated with the Tukey-biweight M-estimator (Beaton and Tukey, 1974). Indeed, the loss function is constant beyond a certain threshold, thus yields an improper distribution.

4.2. The Gamma GLM Case

In the linear regression case, we had an expression for ρ for the Huber M-estimator, which is a loss function that can be seen as being the negative of the log density associated with the distribution of the response variable (recall the discussion in Section 3.2). The difference between the gamma GLM case and the linear regression one is that there is no expression for ρ in the former case, but rather an expression for Ψ which is the derivative

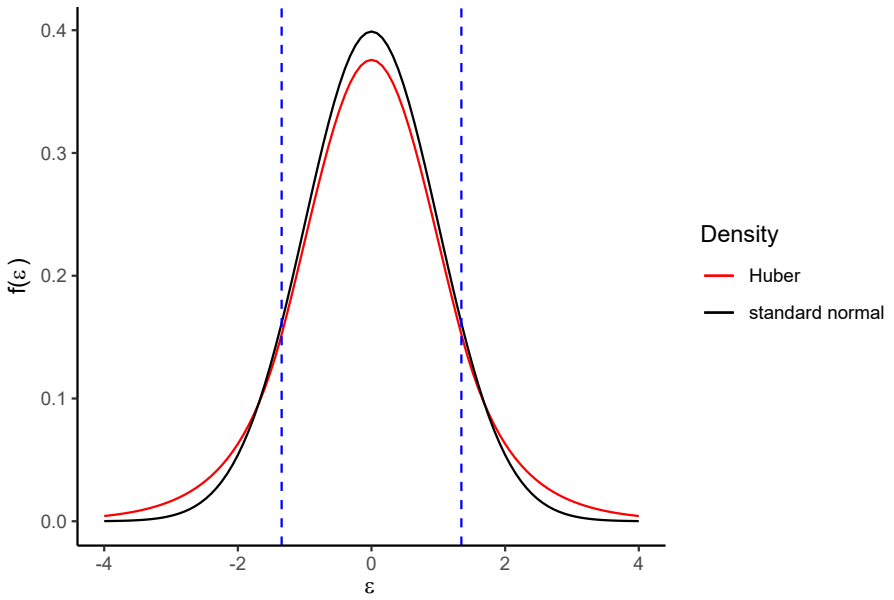


Figure 4.2. Density of the $\mathcal{N}(0,1)$ and that corresponding to the Huber M-estimator with $k = 1.345$

of ρ. In this section, we make an attempt at establishing a connection between the robust M method and heavy-tailed distributions. The first step is to identify a function ρ which yields the proposed Ψ by the robust M method. As we will present shortly, the choice of ρ is not unique, which makes the connection not as clear as with robust linear regression. We will identify a function ρ which, in our opinion, corresponds to the most natural choice. The second step is to connect this function ρ with a PDF for the response variable, as we performed in the linear regression case. We then compare this density with the original gamma one to see if this PDF has a slower decrease than that of gamma, i.e. whether it is a heavy-tailed distribution.

Let us recall the proposed estimating equation by the robust M method. To simplify, we consider that the weight function is such that $w(\boldsymbol{x}_i) = 1$ for all i, and we omit the Fisher consistency term $a(\boldsymbol{\beta})$. The estimating equation regarding $\boldsymbol{\beta}$ (recall (3.1.2) and (3.1.3)) with this simplification is given by

$$-\sum_{i=1}^{n} \Psi(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) = \mathbf{0},$$

where

$$\Psi(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi) = \begin{cases} -\sqrt{\nu}\ r_i(\boldsymbol{\beta}, \phi)\ \boldsymbol{x}_i & \text{if } |r_i(\boldsymbol{\beta}, \phi)| \leq c, \\ -\sqrt{\nu}\ c\ \text{sign}(r_i(\boldsymbol{\beta}, \phi))\ \boldsymbol{x}_i & \text{otherwise,} \end{cases} \tag{4.2.1}$$

with $\phi = 1/\nu$. Recall that with gamma GLMs, $r_i(\boldsymbol{\beta}, \phi) = \sqrt{\nu}(y_i - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))/\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$. We will write ν instead of ϕ as an argument in following functions to avoid confusion.

As mentioned in Section 3.2, Ψ can be viewed as the partial derivative of ρ with respect to $\boldsymbol{\beta}$. If we set ρ as following:

$$\rho(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \nu) = \begin{cases} -\ell_i(\boldsymbol{\beta}, \nu) & \text{if } |r_i(\boldsymbol{\beta}, \nu)| \leq c, \\ c\sqrt{\nu}\left(h(y_i) - \boldsymbol{x}_i^T\boldsymbol{\beta}\right) + a_1(\nu) & \text{if } r_i(\boldsymbol{\beta}, \nu) > c, \\ -c\sqrt{\nu}\left(h(y_i) - \boldsymbol{x}_i^T\boldsymbol{\beta}\right) + a_2(\nu) & \text{if } r_i(\boldsymbol{\beta}, \nu) < -c, \end{cases} \tag{4.2.2}$$

where $\ell_i(\boldsymbol{\beta}, \nu)$ is the contribution of the data point i to the log-likelihood in gamma GLMs, i.e.

$$\ell_i(\boldsymbol{\beta}, \nu) = -\nu(y_i/\mu_i + \log \mu_i) + (\nu - 1)\log y_i + \nu \log \nu - \log(\Gamma(\nu)),$$

we can verify that the derivative of (4.2.2) with respect to $\boldsymbol{\beta}$ is equal to $\Psi(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \phi)$ in (4.2.1). The terms $a_1(\nu)$ and $a_2(\nu)$ will be used to ensure that this loss function is continuous. There are many possible loss functions that can result in the estimating equation above, because $h(y_i)$ which does not depend on $\boldsymbol{\beta}$ has no influence on the derivative of ρ with respect to $\boldsymbol{\beta}$. Different choices for h can thus yield the same estimating equation.

In order to establish a connection between the loss function in (4.2.2) and a heavy-tailed distribution as we achieved in the linear regression case, we consider a natural choice for h, as we now explain.

When $y_i \to \infty$, with $\boldsymbol{\beta}$ and ν fixed, the dominant term of $\ell_i(\boldsymbol{\beta}, \nu)$ in gamma GLMs is

$$-\nu(y_i/\mu_i) = -\nu \exp\left\{\log(y_i) - \boldsymbol{x}_i^T\boldsymbol{\beta}\right\}.$$

To retrieve a similar form in the function in (4.2.2) when $r_i(\boldsymbol{\beta}, \nu) > c$ (which is the part of the function that is activated when y_i is large, and $\boldsymbol{\beta}$ and ν are fixed), h should be set to be the log function. With this function, the PDF of the response variable y_i based on $\rho(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \nu)$ in (4.2.2) is given by

$$f_{\boldsymbol{\beta},\nu,c}(y_i) = \exp(-\rho(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \nu)) = \frac{1}{\mu_i} f_{\nu,c}\left(\frac{y_i}{\mu_i}\right) \propto \frac{1}{\mu_i} g_{\nu,c}\left(\frac{y_i}{\mu_i}\right),$$

where $g_{\nu,c}$ is defined as

$$g_{\nu,c}(z) := \begin{cases} g_{\text{mid}}(z) := \exp\{-\nu z\}\, z^{\nu-1}\nu^\nu/\Gamma(\nu) & \text{if } |\sqrt{\nu}(z-1)| \leq c, \\ g_{\text{right}}(z) := z^{-c\sqrt{\nu}} \exp(a_1(\nu)) & \text{if } \sqrt{\nu}(z-1) > c, \\ g_{\text{left}}(z) := z^{c\sqrt{\nu}} \exp(a_2(\nu)) & \text{if } \sqrt{\nu}(z-1) < -c. \end{cases} \tag{4.2.3}$$

After the normalization, $f_{\nu,c}(z) = g_{\nu,c}(z)/m(\nu)$, where the normalizing constant $m(\nu)$ depends on ν. The explicit forms of $a_1(\nu)$, $a_2(\nu)$, and $m(\nu)$ are presented in Appendix A. The density $f_{\nu,c}$ can be viewed as a standardized version of $f_{\boldsymbol{\beta},\nu,c}$, as it does not depend on $\boldsymbol{\beta}$ any longer. Unfortunately, the parameter ν controls the shape of the gamma PDF, thus it is not possible to further standardize the random variable, contrarily to what can be done in the linear regression case.

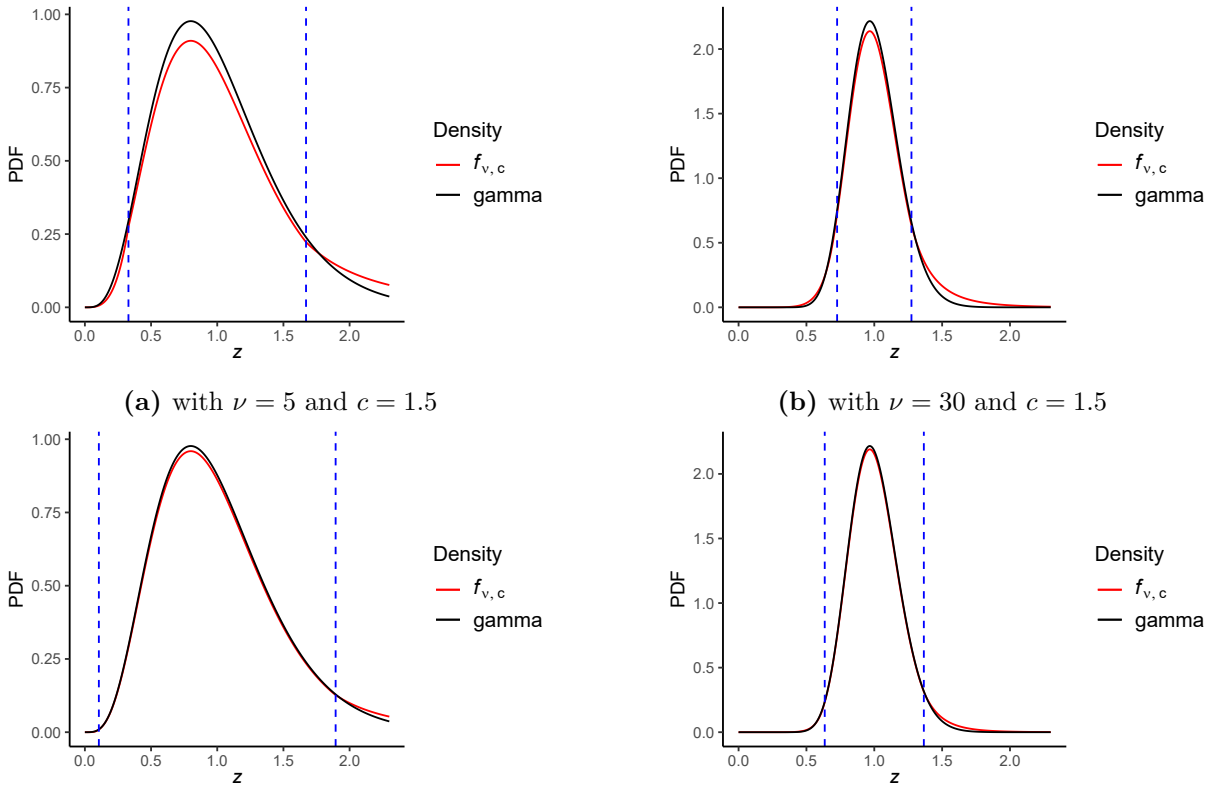
The random variable z is the response variable divided by the mean of the gamma model; its PDF is proportional to that of the gamma distribution with $\mu = 1$, when evaluated at z belonging to a closed interval. When z is outside of this interval, the function decreases polynomially.

The function g may not be integrable. It is integrable if $c\sqrt{\nu} > 1$. Moreover, the left tail of g may not exist. Since $z > 0$, it means that to have a left tail, $-c/\sqrt{\nu} + 1 > 0$ must be satisfied. The first condition $c\sqrt{\nu} > 1$ needs to be satisfied in order for the density to be a PDF. If we combine it with the second one, in order to obtain a PDF with a left part g_{left}, c must satisfy $1/\sqrt{\nu} < c < \sqrt{\nu}$ and $\nu > 1$. This condition makes sense, because the original gamma PDF does not converge to 0 as $z \to 0$ when $\nu \leq 1$ (it converges to a constant when $\nu = 1$ and goes to infinity when $\nu < 1$), meaning that the gamma PDF has, in a sense, no left tail in this case.

In order to understand better the difference with gamma GLMs in terms of tail behaviour, let us take a closer look at the two tails of $f_{\nu,c}$ separately. For simplicity, we study $g_{\nu,c}$ instead, as the normalizing constant $m(\nu)$ does not influence the tail behaviour. On the right side, when $z \to \infty$, with ν fixed, the dominant term of the gamma PDF is $\exp(-\nu z)$, whose decrease is exponential, which is thus a faster decrease compared with that of g_{right}. On the left side, when $z \to 0$, with ν fixed, the dominant term of the gamma PDF is $z^{\nu-1}$, which has a polynomial decrease. The dominant term of g_{left} is $z^{c\sqrt{\nu}}$, which also decreases polynomially. Different from the right side, the gamma PDF and g_{left} have both a polynomial decrease when z tends to 0.

Comparisons between $f_{\nu,c}$ and gamma PDF with different values of ν and c are shown in Figure 4.3. As we can observe, when c and/or ν increases, $f_{\nu,c}$ approaches more and more a gamma PDF. We understand why it is the case when c increases by looking at (3.1.2). When ν increases, it is because the right tail has a faster decrease as it behaves like $z^{c\sqrt{\nu}}$; the left tails of the two densities have similar behaviour. The left tail of $f_{\nu,c}$ is nearly the same as that of a gamma when $\nu = 30$ and $c = 2$, whereas the difference for the right tails of the two densities is still visible.

The analysis conducted in this section to establish a connection between the estimator associated with the robust M method and a heavy-tailed distribution is more complicated than that in the linear regression framework in Section 4.1. The main reason is because, with the robust M method, the estimating equation is modified, instead of the log-likelihood function. That in turn is because the Pearson residual in the likelihood function of gamma GLMs is not retrieved, contrarily to the standardized residual $(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})/\sigma$ in the linear-regression framework. We saw that modifying the estimating equation yields a robust estimator that does not allow to establish a one-to-one correspondence with a response PDF, which is not appealing from a modelling perspective. We nevertheless managed to establish a natural connection with a specific heavy-tailed model, whose PDF is similar to that of the gamma



(a) with $\nu = 5$ and $c = 1.5$

(b) with $\nu = 30$ and $c = 1.5$

(c) with $\nu = 5$ and $c = 2$

(d) with $\nu = 30$ and $c = 2$

Figure 4.3. Comparison between gamma PDFs and $f_{\nu,c}$ for different values of ν and c

on the central part. That motivates the introduction of robust alternatives based directly on a modified response PDF, with similar desirable characteristics. Our proposed method presented in detail in Chapter 5 represents such a robust alternative.

Chapter 5

Proposed Robust Gamma GLMs

In this chapter, we will present our proposed robust method for gamma GLMs. As we mentioned in Introduction and Chapter 4, the idea comes from recent Bayesian approaches (Desgagné, 2015; Gagnon et al., 2020): we adapt the original gamma GLMs to a robust version by replacing the tails by heavier ones, while keeping the central part of the PDF as is. In Section 5.1, we present the definition of our model, and the differences between this one and the one identified in Section 4.2. In Section 5.2, we present theoretical results that characterize our model: first, we provide sufficient conditions under which the posterior distribution is proper, guaranteeing that a Bayesian analysis can be conducted; second, we present an asymptotic result about the behaviour of the posterior distribution as outliers move further and further away from the bulk of the data. We present also simulation results that support those theoretical results.

5.1. Model Definition

Recall that with gamma GLMs (see (1.2.1)), the density of a response variable y_i is given by

$$f_{\beta,\nu}(y_i) = \frac{1}{\mu_i} f_\nu \left(\frac{y_i}{\mu_i} \right),$$

where f_ν is given by

$$f_\nu(z) = \exp\{-\nu z\}\, z^{\nu-1} \nu^\nu / \Gamma(\nu).$$

We have shown with an example in Chapter 2 that gamma GLMs are not robust against outliers. In Chapter 3, we presented the robust M method proposed by Cantoni and Ronchetti (2001), which is the most popular frequentist method for robust GLMs. It consists in a robust estimator, which is similar, in essence, to M-estimators. We stress that the difference with our approach is that we adapt the model to the potential presence of outliers, rather than estimators. As mentioned, the advantage is that it is easier to understand the difference

with gamma GLMs from a modelling point of view, and it can be applied to both frequentist and Bayesian analyses.

We consider a new density to replace $f_\nu(z)$, which consists of an adaptation of the log-Pareto-tailed normal (LPTN) distribution proposed by Desgagné (2015) and used by Gagnon et al. (2020) in the context of linear regression. The errors ϵ_i in linear regression follow an LPTN distribution, whose central part is a standard normal, but with tails that have been replaced by log-Pareto ones, which behave like $(1/|x|)(1/\log|x|)^\lambda$. In the context of gamma GLMs, we assume that the response variable y_i follows no longer a gamma distribution: f_ν is replaced by $f_{\nu,c}$, whose central part is that of a gamma PDF, whereas the tails are replaced by log-Pareto ones.

There are two main differences between our proposed model and the one identified in Section 4.2 (see (4.2.3)). Firstly, the central part of our proposed distribution $f_{\nu,c}$ matches exactly a gamma PDF, rather than being proportional to it as in (4.2.3), which aims to improve efficiency. Secondly, both left and right tails in our model are log-Pareto, implying that they are heavier than polynomial ones. Models with log-Pareto tails have better robustness properties than models with polynomial tails, at least in a linear regression context (Gagnon et al., 2020).

The PDF of our proposed distribution $f_{\nu,c}$ is defined as

$$f_{\nu,c}(z) := \begin{cases} f_{\text{mid}}(z) := \exp\{-\nu z\}\, z^{\nu-1}\nu^\nu/\Gamma(\nu) & \text{if } z_{\text{l}} \le z \le z_{\text{r}}, \\ f_{\text{right}}(z) := f_{\text{mid}}(z_{\text{r}})\dfrac{z_{\text{r}}}{z}\left(\dfrac{\log(z_{\text{r}})}{\log(z)}\right)^{\lambda_{\text{r}}} & \text{if } z > z_{\text{r}}, \\ f_{\text{left}}(z) := f_{\text{mid}}(z_{\text{l}})\dfrac{z_{\text{l}}}{z}\left(\dfrac{\log(z_{\text{l}})}{\log(z)}\right)^{\lambda_{\text{l}}} & \text{if } 0 < z < z_{\text{l}}, \end{cases} \tag{5.1.1}$$

where z_{r}, λ_{r}, z_{l}, λ_{l} are functions of ν and c given by

$$z_{\text{r}} := 1 + c/\sqrt{\nu}, \quad z_{\text{l}} := \begin{cases} 1 - c/\sqrt{\nu} & \text{if } \nu > 1 \\ 0 & \text{if } \nu \le 1 \end{cases},$$

$$\lambda_{\text{r}} := 1 + \frac{f_{\text{mid}}(z_{\text{r}})\log(z_{\text{r}})\, z_{\text{r}}}{\mathbb{P}[Z_{\text{gamma}} > z_{\text{r}}]}, \quad \text{and} \quad \lambda_{\text{l}} := 1 - \frac{f_{\text{mid}}(z_{\text{l}})\log(z_{\text{l}})\, z_{\text{l}}}{\mathbb{P}[0 < Z_{\text{gamma}} < z_{\text{l}}]},$$

with Z_{gamma} being a random variable following a gamma distribution whose mean and dispersion parameter are given by 1 and $1/\nu$, respectively.

The tuning parameter c with this model is considered to be a positive constant, and thus $0 < c/\sqrt{\nu} < \infty$. This implies that $z_{\text{r}} > 1$ and thus the log terms in f_{right} are positive. Also, f_{left} exists when $z_{\text{l}} > 0$, i.e. when $c < \sqrt{\nu}$ and $\nu > 1$, and z_{l} is upper bounded by 1. This implies that both log terms in f_{left} are negative and thus that $f_{\text{left}}(z) > 0$ when $0 < z < z_{\text{l}}$.

We now make two remarks about the tuning parameter c. First or all, it plays the same role as the parameter with the same notation c in the robust M method (recall (4.2.3)): the conditions in (5.1.1) to determine which part of the function is activated can be rewritten as

in (4.2.3). Furthermore, there is a correspondence between the value of c and the mass under $f_{\nu,c}$ assigned to the part where the density exactly matches the gamma PDF. For example, when $c = 1.35$ and $\nu = 41.32$ (the value used for c and estimated for ν in the real-data example in Chapter 2), the mass of the central part is $\mathbb{P}[-1.35 \leq Z_{\text{gamma}} \leq 1.35] \approx 0.83$. We can use this correspondence to guide the choice of c, if one has prior belief about ν. If, for instance, one believes that ν should take values around 40, and one wants 90% of the mass to be assigned to the central part where the density matches the gamma PDF, one could set c to 1.65. In Chapter 2, we presented the estimation results for the health-care data set based on our proposed method. The parameter c was set to 1.35, which provides good results. Studies about how to choose objectively and effectively this tuning parameter are left for future work.

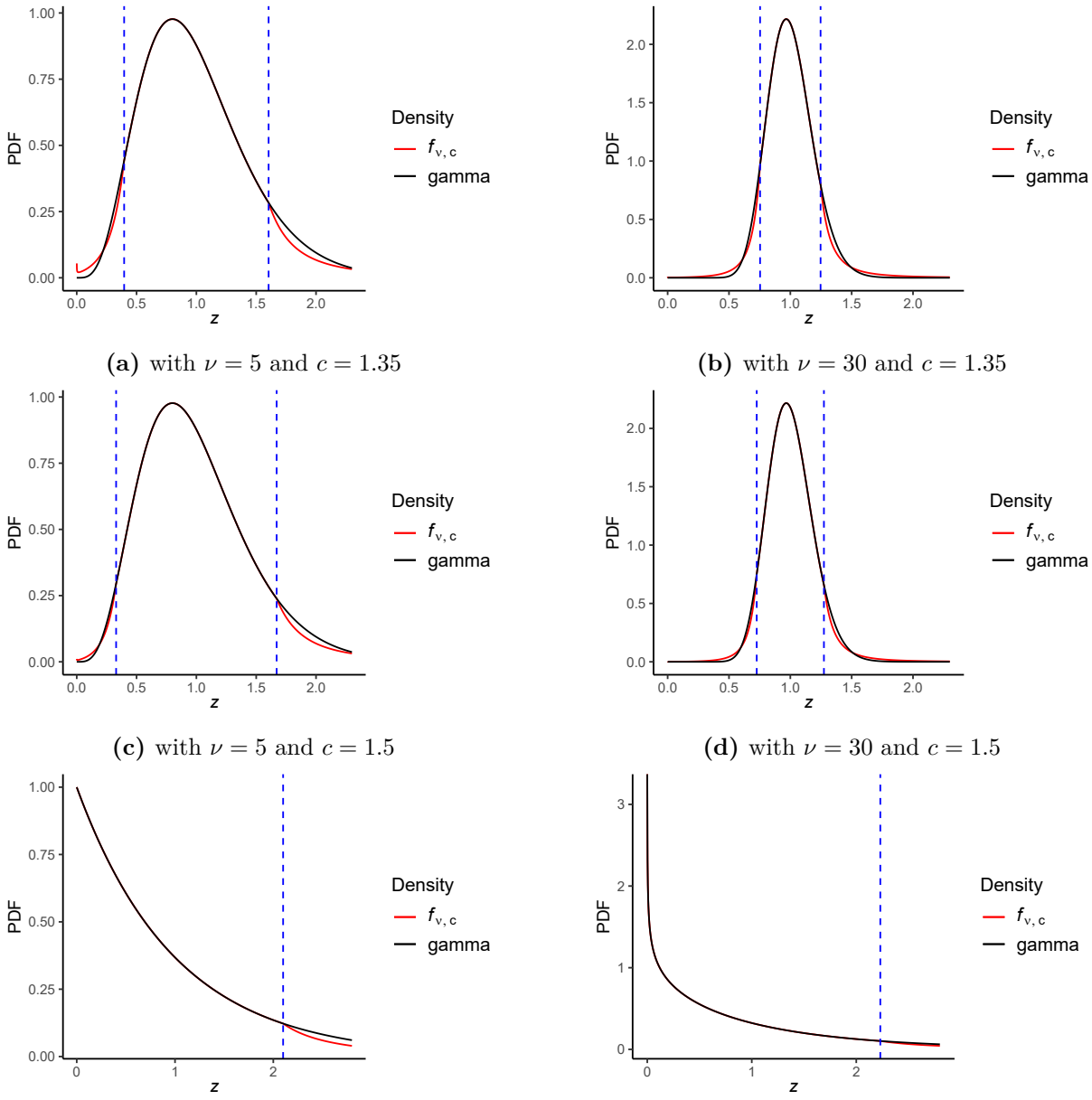
The terms z_{l} and z_{r}, depending on ν and c, control which part of the function is activated. The terms $f_{\text{mid}}(z_{\text{r}})$, z_{r} and $\log(z_{\text{r}})$ in f_{right}, as well as $f_{\text{mid}}(z_{\text{l}})$, z_{l} and $\log(z_{\text{l}})$ in f_{left} ensure that the PDF is continuous. The function $f_{\nu,c}$ is integrable for all $c, \nu > 0$. It goes to $+\infty$ when $z \to 0$, when f_{left} exists. This behaviour close to 0 allows to have integrals that are similar to those on the right tails, and that are to be contrasted with those under the original gamma PDF given that the latter goes to 0 as $z \to 0$. Indeed, integrals from 0 to small values a can be rewritten as

$$\int_0^a f_{\text{mid}}(z_{\text{l}}) \frac{z_{\text{l}}}{z} \left(\frac{\log(z_{\text{l}})}{\log(z)} \right)^{\lambda_{\text{l}}} \mathrm{d}z = \int_{1/a}^{\infty} f_{\text{mid}}(z_{\text{l}}) \frac{z_{\text{l}}}{u} \left(\frac{\log(1/z_{\text{l}})}{\log(u)} \right)^{\lambda_{\text{l}}} \mathrm{d}u.$$

After the change of variables, the mass associated to the left tail can be viewed as an integral from $1/a$ to ∞ with respect to a function which is similar to f_{right}, but with a different normalizing constant and a different power term. In other words, the behaviour of f_{left} is analogous to that of f_{right}, up to a change of variables.

The constraint that $z_{\text{l}} = 0$ if $\nu \leq 1$ is to ensure that f_{left} is never activated when the original gamma PDF does not have a left tail. The idea of the model is to replace the tails (when they exist) by heavier ones.

Comparisons between gamma PDFs and $f_{\nu,c}$ with different values of ν and c are shown in Figure 5.1. In **(e)** and **(f)**, f_{left} is not activated since $c \geq \sqrt{\nu}$. In **(b)**, **(c)** and **(d)**, we do not see that $f_{\nu,c}(z) \to \infty$ when $z \to 0$ because this explosive behaviour happens too close to 0 to be observed. Similar to the density in (4.2.3), $f_{\nu,c}$ approaches more and more a gamma PDF when c increases. We understand the reason by looking at (5.1.1). When ν increases, f_{mid} becomes more pointed, and both tails of the gamma PDF and those of $f_{\nu,c}$ have a more rapid decrease, although the latter always decrease more slowly; it is thus more difficult to visually distinguish the gamma PDF from $f_{\nu,c}$ when ν is large. Different from the LPTN distribution, $f_{\nu,c}$ is not symmetric; the mass associated with f_{right} and f_{left} can be very unbalanced, which depends on the value of ν.



(a) with $\nu = 5$ and $c = 1.35$

(b) with $\nu = 30$ and $c = 1.35$

(c) with $\nu = 5$ and $c = 1.5$

(d) with $\nu = 30$ and $c = 1.5$

(e) with $\nu = 1$ and $c = 1.1$

(f) with $\nu = 0.8$ and $c = 1.1$

Figure 5.1. Comparison between gamma PDFs and $f_{\nu,c}$ for different values of ν and c

The proposed model can be estimated by the maximum likelihood method. The estimation results shown in Chapter 2 were produced using this method. By writing the log-likelihood function associated with $f_{\nu,c}$, the MLE can be viewed as a robust M-estimator of the gamma GLM according to the definition in (3.2.1). As mentioned, one of the advantages of our approach is that it can be also applied to perform robust Bayesian analyses. MCMC methods can be employed to obtain posterior means, medians, credible intervals, and so on. In the next section, we will present theoretical results which characterize the behaviour of the posterior distribution resulting from our proposed model.

5.2. Theoretical Results

The theoretical results presented in this section assume that all explanatory variables are continuous to simplify. To use the proposed model in Bayesian analysis, we need to select a prior distribution for $\boldsymbol{\beta}$ and ν, denoted by π. Importantly, we have to make sure that the posterior distribution is proper. The posterior density with a prior π is such that (recall (1.3.13)):

$$\pi(\boldsymbol{\beta}, \nu \mid \mathbf{y}) = m(\mathbf{y})^{-1} \pi(\boldsymbol{\beta}, \nu) \prod_{i=1}^{n} \left(\frac{1}{\mu_i}\right) f_{\nu,c}\left(\frac{y_i}{\mu_i}\right). \tag{5.2.1}$$

We will use $\pi(\boldsymbol{\beta} \mid \nu)$ to represent the conditional (prior) density of $\boldsymbol{\beta}$ given ν, and $\pi(\nu)$ to represent the marginal (prior) density of ν.

Proposition 5.2.1.

Assume that $\pi(\boldsymbol{\beta} \mid \nu)$ is bounded, and that $\pi(\nu)$ is a proper PDF such that $\int_0^\infty \pi(\nu)\nu^{(n-p)/2}\,\mathrm{d}\nu < \infty$. If $n \geq p \geq 1$, the posterior distribution is proper.

Proof.

See Appendix B.1. □

The assumptions on the prior are weak, which explains why we require $n \geq p$, a condition similar to that for frequentist inference. The condition on $\pi(\boldsymbol{\beta} \mid \nu)$ is satisfied by any continuous PDF and by Jeffreys prior. The condition on $\pi(\nu)$ is satisfied if the prior is a gamma distribution with any shape and scale parameters. We believe that our assumption on $\pi(\nu)$ can be weakened. Indeed, our assumption seems to be a consequence of our proof technique, but we did not manage to find a more effective technique.

We now state a result characterizing the robustness of our model against outliers. The result is asymptotic, and more precisely, about the behaviour of the posterior distribution under an asymptotic regime, where outliers are considered to be further and further from the bulk of the data. As we explained in Chapter 2, an outlier is defined as a couple $(\boldsymbol{x}_i, y_i)$ whose components are incompatible with the trends in the bulk of the data. The analogue of the Pearson residual $r_i(\boldsymbol{\beta}, \nu)$ (recall (3.1.1)) can be used to evaluate this incompatibility. It can be extreme because, for a given $\boldsymbol{x}_i$, the value of y_i makes it extreme or because, for a given y_i, the value of $\boldsymbol{x}_i$ makes it extreme. We mathematically represent such extreme situations by considering an asymptotic scenario where the outliers move away from the bulk of the data along particular paths (see Figure 5.2). More precisely, we consider that the outliers $(\boldsymbol{x}_i, y_i)$ are such that $y_i \to \infty$ or $y_i \to 0$ with $\boldsymbol{x}_i$ being kept fixed (but perhaps extreme). Our result states that, for the outliers with fixed $\boldsymbol{x}_i$, there exist y_i values such that the posterior distribution is similar to one which excludes the PDF terms of the outliers.

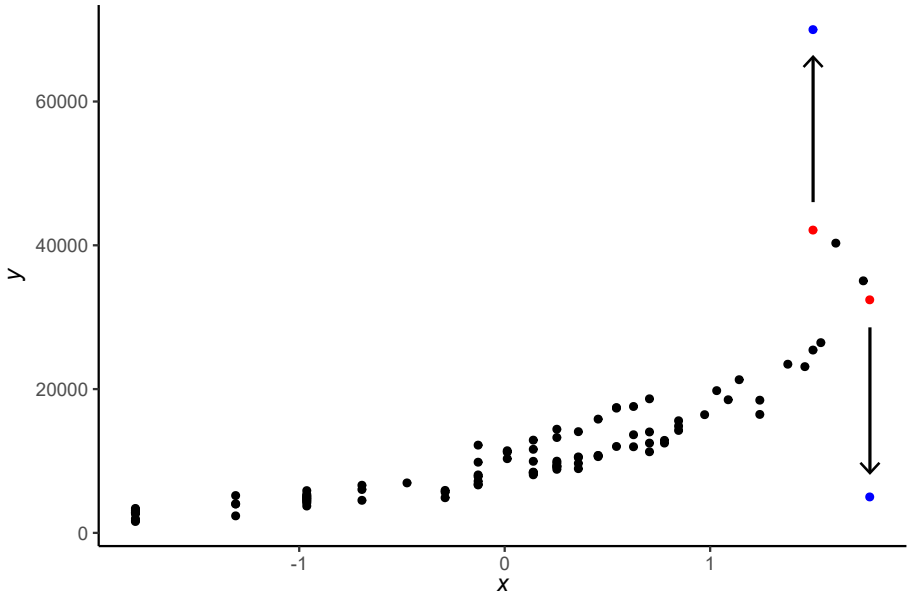


Figure 5.2. A data set with two outliers where one can be seen as having a $y_i \to \infty$ and the other one can be seen as having a $y_i \to 0$

Central to the characterization of the robustness of our proposed model is the limiting behaviour of the PDF evaluated at an outlying point. The following proposition is about this limiting behaviour.

Proposition 5.2.2.

Consider c, ν and μ fixed. We have

$$\lim_{y\to\infty} \frac{f_{\nu,c}(y/\mu)/\mu}{f_{\nu,c}(y)} = 1.$$

If $c < \sqrt{\nu}$ and $\nu > 1$,

$$\lim_{y\to 0} \frac{f_{\nu,c}(y/\mu)/\mu}{f_{\nu,c}(y)} = 1.$$

Proof.

See Appendix B.2. □

Figure 5.3 shows the ratio $(1/\mu)f_{\nu,c}(y/\mu)/f_{\nu,c}(y)$ as a function of y, with $\nu = 30$, $c = 1.35$ and $\mu = 2$. We take the same values of c and ν as in Figure 5.1b. As we observe, this ratio does converge to 1 but slowly. The speed of the convergence is logarithmic, and it depends on ν, μ and c. The greater the value of ν, μ and c, the slower the convergence.

To obtain a theoretical result about the asymptotic behaviour of the posterior distribution, we simplify the context and consider that the parameter ν is a fixed constant; the unknown parameter is thus considered to be only $\boldsymbol{\beta}$ for the rest of the section. The prior

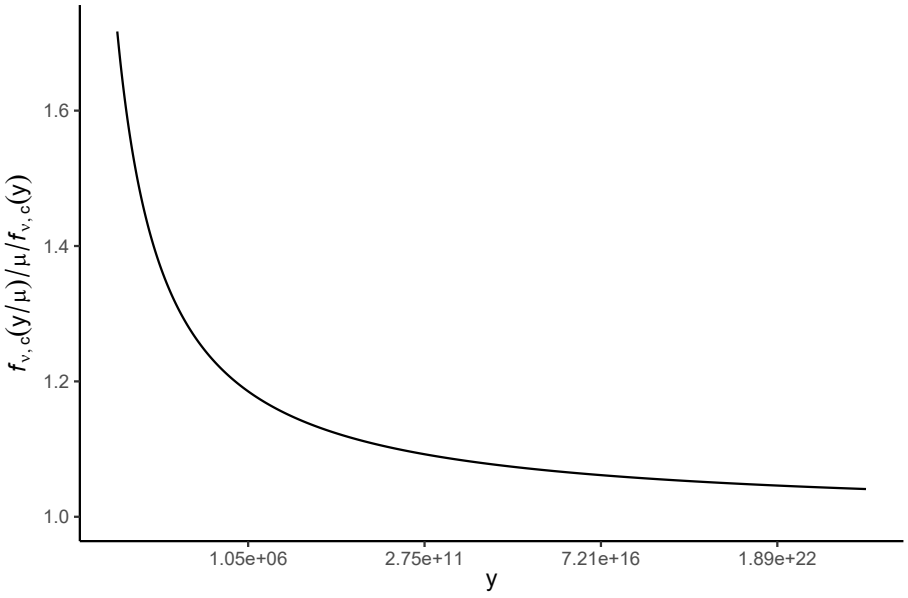


Figure 5.3. The ratio $(1/\mu)f_{\nu,c}(y/\mu)/f_{\nu,c}(y)$ as a function of y, with $\nu = 30$, $c = 1.35$ and $\mu = 2$

and posterior are thus about this parameter only. We further simplify by considering that ν is such that $c < \sqrt{\nu}$ and $\nu > 1$ to ensure the existence of both tails, which corresponds to the shape that often is sought for and supported by the data. We call a couple $(\boldsymbol{x}_i, y_i)$ where y_i goes to ∞, a *big outlier*, and a couple $(\boldsymbol{x}_i, y_i)$ where y_i goes to 0, a *small outlier*. The y_i itself is called a *big/small outlying observation*. More precisely, we consider that each y_i goes to ∞ or 0 at its own specific rate. In particular, for a big outlying observation, $y_i = b_i\omega$, whereas $y_i = b_i/\omega$ for a small outlying observation, and we let $\omega \to \infty$. For a non-outlying observation, we assume that $y_i = a_i$, where $a_i \in \mathbb{R}$. Among the n observations $y_1, \ldots, y_n$, we assume that k of them form a group of non-outlying observations, s of them form a group of small outlying observations, and r of them form a group of big outlying observations. We denote the set of non-outlying observations, small outlying observations, and big outlying observations as $\mathbf{y}_k, \mathbf{y}_s, \mathbf{y}_r$, respectively. For $i = 1, \ldots, n$, we define the binary functions k_i, s_i and r_i as follows: $k_i = 1$ if y_i is a non-outlying observation, $s_i = 1$ if it is a small outlying observation, and $r_i = 1$ if it is a big outlying observation. These functions take the value of 0 otherwise. Therefore, we have $k_i + s_i + r_i = 1$ for $i = 1, \ldots, n$, with $\sum_{i=1}^n k_i = k$, $\sum_{i=1}^n s_i = s$, and $\sum_{i=1}^n r_i = r$.

In the simplified context described above, Proposition 5.2.2 suggests that the PDF term of an outlier in the posterior density behaves in the limit like $f(y_i) \propto 1$. This conflicting information is thus wholly rejected as its source becomes increasingly remote (West, 1984). The model is thus said to be *wholly robust*. Note that this is case in the simplified context

where ν is considered fixed. If it is considered unknown, the model is *partially robsut*, as the limiting term $f(y_i)$ does not depend on $\boldsymbol{\beta}$, but depends on ν.

The theoretical result that we demonstrate is a convergence of the posterior distribution towards $\pi(\cdot \mid \mathbf{y}_k)$, which has a density defined as follows:

$$\pi(\boldsymbol{\beta} \mid \mathbf{y}_k) := \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \left[f_{\nu,c}(y_i/\mu_i)/\mu_i \right]^{k_i} \Big/ m(\mathbf{y}_k), \quad \boldsymbol{\beta} \in \mathbb{R}^p,$$

where

$$m(\mathbf{y}_k) := \int_{\mathbb{R}^p} \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \left[f_{\nu,c}(y_i/\mu_i)/\mu_i \right]^{k_i} \mathrm{d}\boldsymbol{\beta}.$$

Theorem 5.2.1.

Suppose that ν is a fixed constant such that $c < \sqrt{\nu}$ and $\nu > 1$. Assume that π is bounded. If $k \geq d(r+s) + 2p - 1$, i.e. $n \geq (d+1)(s+r) + 2p - 1$, where $d = \max\{\lambda_\mathrm{l}/\lambda_\mathrm{r}, \lambda_\mathrm{r}/\lambda_\mathrm{l}\}$, then as $\omega \to \infty$,

(a) the asymptotic behaviour of the marginal distribution is:

$$\frac{m(\mathbf{y})}{\prod_{i=1}^{n} [f(y_i)]^{s_i+r_i}} \to m(\mathbf{y}_k);$$

(b) the posterior density converges pointwise: for any $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\pi(\boldsymbol{\beta} \mid \mathbf{y}) \to \pi(\boldsymbol{\beta} \mid \mathbf{y}_k);$$

(c) the posterior distribution converges: $\pi(\cdot \mid \mathbf{y}) \to \pi(\cdot \mid \mathbf{y}_k)$.

Proof.

See Appendix B.3.

□

Figure 5.4 shows the value of d as a function of ν for $\nu > c^2$ when $c = 1.35$. Numerically, we observe that $\lambda_\mathrm{l} > \lambda_\mathrm{r}$, and that the ratio is monotonically decreasing from a value of 2.27 when $\nu = 2$ (we take a value of ν close to $c^2 = 1.35^2 = 1.82$) and converges to 1 as $\nu \to \infty$.

For our proposed model, once the parameter ν is considered as a fixed constant, and that the prior distribution has been set such that $\pi(\boldsymbol{\beta})$ is bounded, it is seen that Theorem 5.2.1 holds as long as the number of non-outliers is large enough. A sufficient number of non-outliers is $d(r+s) + 2p - 1$, where values of d are shown numerically in function of ν in Figure 5.4. This condition suggests that the breakdown point, generally defined as the proportion of outliers $(r+s)/n$ that an estimator can handle, is $1/(d+1) - (2p-1)/(n(d+1))$, which is close to $1/(d+1)$ if n is large relatively to p.

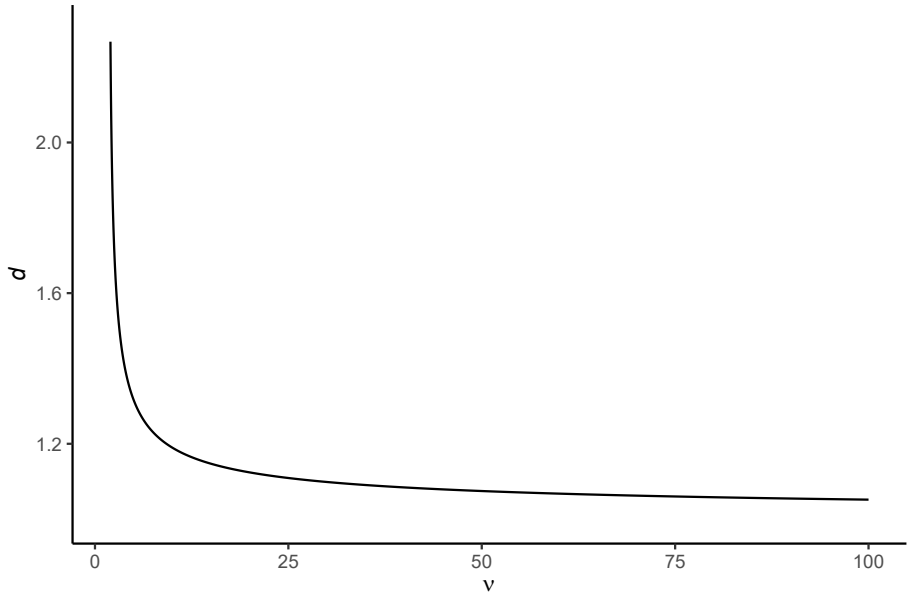


Figure 5.4. The ratio d as a function of ν with $c = 1.35$

In Theorem 5.2.1, result (a) represents the centrepiece; it leads relatively easily to the other results of the theorem, but its demonstration requires considerable work. The convergence of the posterior density in result (b) enables to state that the MAP is wholly robust. Given that this estimate corresponds to the MLE when the prior is proportional to 1, the frequentist estimate is, as a result, also wholly robust. This allows establishing a connection between Bayesian and frequentist robustness. Result (c) indicates that any estimation of $\boldsymbol{\beta}$ based on posterior quantiles (e.g. using posterior medians or Bayesian credible intervals) is wholly robust to outliers. All these results characterize the limiting behaviour of a variety of Bayes estimators.

We now perform a simulation study in order to show the empirical behaviour of estimates when one outlier is more and more extreme. In the simulation, we set $n = 20$, and we consider a model with an intercept and an explanatory variable. We set x_{i2} from 2 to 6 with equal distances for $i = 1, \ldots, 20$, and the true coefficients are $\beta_0 = -1$ and $\beta_1 = 1$. The observations are generated from a gamma GLM with a logarithmic link with $\nu = 40$. Therefore, y_i follows a gamma distribution with parameters $\mu_i = \exp(-1 + x_{i2})$ and $\nu = 40$. To generate an outlier, we consider two ways: either generate an extreme observation y_i with an ordinary values of x_{i2}, or generate an extreme explanatory variable x_{i2} with an ordinary values of y_i. We consider both ways because in practice, with a fixed data set, both $\boldsymbol{x}$ and y can make a point an outlier.

Let us consider for now the first way. We gradually increase the value of y_{20}, from 170 to 500. The associated $r_i(\boldsymbol{\beta}, \nu)$ for the outlying point varies from 0.92 to 14.98. For each data set, we perform an analysis by using our proposed model in (5.1.1) with $c = 1.35$ and $\nu = 40$,

as if we knew the true value of ν. We estimate the parameters β_1, β_2 with the maximum likelihood method, based on $y_1, \ldots, y_{19}$ and the 20th outlying observation. We also perform an analysis by removing completely y_{20}.

Figure 5.5 shows the estimates for β_1 and β_2 as y_{20} gets more and more extreme. The black points are estimates by using the maximum likelihood method. The red one indicates the estimate if y_{20} were completely removed from the data set. In the first plot, the MLE of β_1 first decreases, then increases, and finally becomes stable as y_{20} becomes more and more extreme. The estimates converge to that obtained without y_{20}. In the second plot, we observe a similar pattern for $\hat{\beta}_2$ as in the first one: the estimate increases at first, then decreases, and finally converges to the estimate without the outlying point.

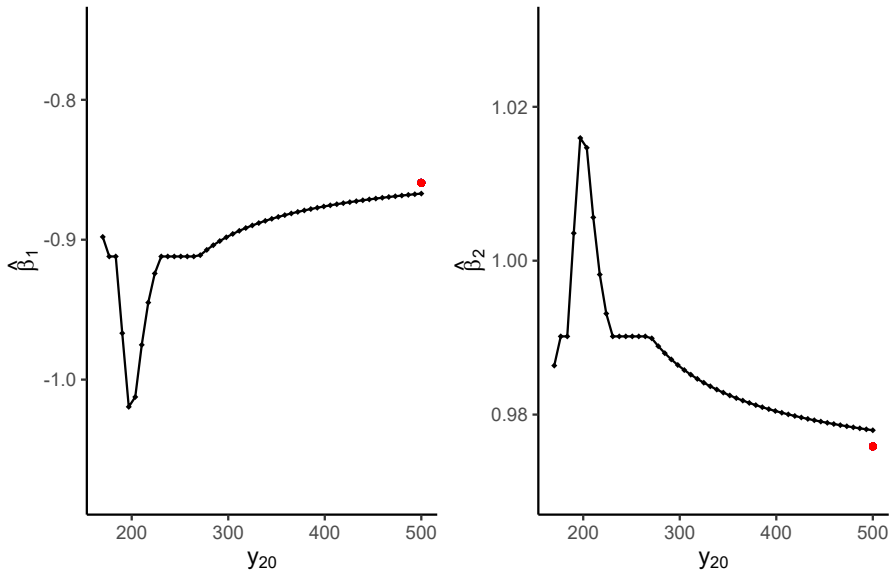


Figure 5.5. Estimates of β_1 and β_2 as a function of y_{20}

Let us consider the second way of generating an outlier. We gradually decrease the value of $x_{20,2}$ from 6 to 4.52 such that the associated $r_i(\boldsymbol{\beta}, \nu)$ vary from -1.44 to 15.13, whose values approximately match those when we moved y_{20}. The results are presented in Figure 5.6. To improve readability, the x-axis is the difference between the maximum of $x_{20,2}$ and $x_{20,2}$, which we denote as $\operatorname{diff} x_{20,2}$, so that the limiting case is on the right side of the x-axis. Figure 5.6 illustrates the estimates for β_1 and β_2 when $\operatorname{diff} x_{20,2}$ gets more and more extreme. Because of the log link of the gamma GLM, a small change in the covariates yields a big change in $r_i(\boldsymbol{\beta}, \nu)$. We observe a similar pattern as in Figure 5.5, but the estimates seem to converge to values that, while being close to estimates obtained without the 20th point, are not equal to these.

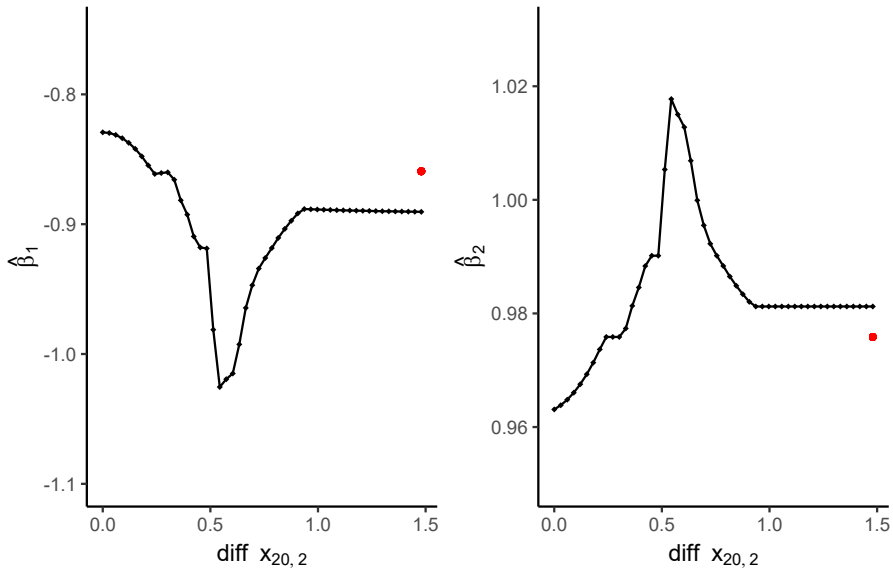


Figure 5.6. Estimates of β_1 and β_2 as a function of diff $x_{20,2}$

Conclusion

The objective of this thesis is mainly to propose an approach with a precise characterization of the model that can be used in both frequentist and Bayesian analyses, in order to solve the problem of non-robustness for GLMs. With an analysis of a real health-care expenditure data set by using a gamma GLM, one can clearly see the difference between estimation with and without a robust method; the outlier detection is also negatively impacted due to the masking effect, if no robust approach is applied. Before presenting our approach, we study the most commonly used method for robust GLMs, which is the frequentist method proposed by Cantoni and Ronchetti (2001). There is a connection between their proposed estimators and the famous robust M-estimators through likelihood estimating equations, however, it is not possible to establish a clear correspondence between their proposed estimators and a model. In the linear regression context, some robust M-estimators can be seen as MLEs with a regression model where its error term follows a heavy-tailed distribution. For example, we presented the connection of the Huber M-estimator and the distribution whose central part is still normal, but the tails have been replaced by Laplace ones. With robust estimators of Cantoni and Ronchetti (2001) for gamma GLMs, although we still identified a distribution which can be considered as heavy-tailed, i.e. its density is gamma in the central part, and tails have been replaced by polynomial ones, this connection is neither natural nor unique. Therefore, it highly motives the introduction of our model. We mainly focus on gamma GLMs, but the approach is seen to be valid for other GLMs with distributions having tails.

Our proposed model consists of directly using a heavy-tailed distribution whose central part is a gamma, while the extremities have been replaced by log-Pareto ones, to replace the original distribution of the response variable. The theoretical results that characterize the model, with regard to the properness of the posterior distribution under weak conditions and the asymptotic behaviour of the posterior distribution, have been the main contribution of this thesis. The presented simulation study supports these theoretical results.

During the study of the asymptotic behaviour of the posterior distribution, we simplified the context by considering that one of the parameter for the gamma distribution, which is the inverse of the dispersion parameter, is a fixed constant. It would be better, in the future study, that this parameter would be considered as random, which is often the case in

practice, as it is improbable that one knows the value of this parameter before the analysis. Moreover, it would be also interesting to perform a study of the tuning parameter, which compromises the efficiency with the robustness of the model. This study would help a user to choose his tuning parameter in an analysis in a more appropriate way.

Appendix

Appendix A: Supplementary Material for Chapter 4

Recall the function $g_{\nu,c}$ in (4.2.3):

$$g_{\nu,c}(z) := \begin{cases} g_{\text{mid}}(z) := \exp\{-\nu z\}\, z^{\nu-1}\nu^{\nu}/\Gamma(\nu) & \text{if } |\sqrt{\nu}(z-1)| \le c, \\ g_{\text{right}}(z) := z^{-c\sqrt{\nu}} \exp(a_1(\nu)) & \text{if } \sqrt{\nu}(z-1) > c, \\ g_{\text{left}}(z) := z^{c\sqrt{\nu}} \exp(a_2(\nu)) & \text{if } \sqrt{\nu}(z-1) < -c. \end{cases}$$

We define $z_{\text{r}} := c/\sqrt{\nu}+1$ and $z_{\text{l}} := -c/\sqrt{\nu}+1$, which signify the right and left thresholds for z that determine which function is activated. Since the function $g_{\nu,c}$ is continuous, the terms $\exp(a_1(\nu))$ and $\exp(a_1(\nu))$ should be such that $g_{\text{mid}}(z_{\text{r}}) = g_{\text{right}}(z_{\text{right}})$, and $g_{\text{mid}}(z_{\text{l}}) = g_{\text{left}}(z_{\text{l}})$. Thus, we have

$$\exp(a_1(\nu)) = \frac{g_{\text{mid}}(z_{\text{r}})}{z_{\text{r}}^{-c\sqrt{\nu}}} = \frac{\exp\{-\nu z_{\text{r}}\}\left(z_{\text{r}}^{\nu+c\sqrt{\nu}-1}\right)\nu^{\nu}}{\Gamma(\nu)},$$

$$\exp(a_2(\nu)) = \frac{g_{\text{mid}}(z_{\text{l}})}{z_{\text{l}}^{c\sqrt{\nu}}} = \frac{\exp\{-\nu z_{\text{l}}\}\left(z_{\text{l}}^{\nu-c\sqrt{\nu}-1}\right)\nu^{\nu}}{\Gamma(\nu)}.$$

The PDF for z is $f_{\nu,c}(z) = g_{\nu,c}(z)/m(\nu)$. In order to find the normalizing constant $m(\nu)$, we need to first calculate the mass for g_{mid}, g_{right} and g_{left} separately, then add them together to find the normalizing constant.

The mass associated with $g_{\text{mid}}(z)$ is given by

$$\text{mass}_{\text{m}}(\nu) = \int_{z_{\text{l}}}^{z_{\text{r}}} \frac{\exp\{-\nu z\}\, z^{\nu-1}\nu^{\nu}}{\Gamma(\nu)}\, \mathrm{d}z = \mathbb{P}\left[Z_{\text{gamma}} < z_{\text{r}}\right] - \mathbb{P}\left[Z_{\text{gamma}} < z_{\text{l}}\right],$$

where Z_{gamma} follows a gamma distribution with $\mu = 1$. For the mass associated with g_{right}, we can obtain an analytical form. It is given by

$$\begin{aligned} \text{mass}_{\text{r}}(\nu) &= \int_{z_{\text{r}}}^{\infty} z^{-c\sqrt{\nu}} \exp(a_1(\nu))\, \mathrm{d}z = \frac{z_{\text{r}}^{-c\sqrt{\nu}+1}}{c\sqrt{\nu}-1} \exp(a_1(\nu)) \\ &= \frac{(\nu z_{\text{r}})^{\nu}\exp(-\nu z_{\text{r}})}{\Gamma(\nu)(c\sqrt{\nu}-1)} = \frac{\left(\nu + c\sqrt{\nu}\right)^{\nu}\exp(-\nu - c\sqrt{\nu})}{\Gamma(\nu)(c\sqrt{\nu}-1)}. \end{aligned}$$

Analogously, the mass associated with g_{left} is given by

$$\begin{aligned}\text{mass}_{\text{l}}(\nu) &= \int_0^{z_{\text{l}}} z^{c\sqrt{\nu}} \exp(a_2(\nu))\,\mathrm{d}z = \frac{z_{\text{l}}^{c\sqrt{\nu}+1}}{c\sqrt{\nu}+1}\exp(a_2(\nu)) \\ &= \frac{(\nu z_{\text{l}})^{\nu}\exp(-\nu z_{\text{l}})}{\Gamma(\nu)(c\sqrt{\nu}+1)} = \frac{\left(\nu - c\sqrt{\nu}\right)^{\nu}\exp(-\nu + c\sqrt{\nu})}{\Gamma(\nu)(c\sqrt{\nu}+1)}.\end{aligned}$$

The normalizing constant $m(\nu)$ is thus given by

$$m(\nu) = \text{mass}_{\text{m}}(\nu) + \text{mass}_{\text{r}}(\nu) + \text{mass}_{\text{r}}(\nu).$$

Appendix B: Supplementary Material for Chapter 5

Appendix B.1: Proof of Proposition 5.2.1

We first present and prove two lemmas that will be used in the proof of Proposition 5.2.1.

Lemma 1.

Viewed as a function of μ, $f_{\nu,c}(y/\mu)/\mu$ is bounded above by $(e^{-1}\nu)^{\nu}/(y\Gamma(\nu))$, for all ν, c and y.

Proof.

Based on (5.1.1),

$$f_{\nu,c}(y/\mu)/\mu = \begin{cases} f_{\text{mid}}(y/\mu)/\mu = \exp\{-\nu y/\mu\}\,\mu^{-\nu}y^{\nu-1}\nu^{\nu}/\Gamma(\nu) & \text{if } z_{\text{l}} \le y/\mu \le z_{\text{r}}, \\ f_{\text{right}}(y/\mu)/\mu = f_{\text{mid}}(z_{\text{r}})\dfrac{z_{\text{r}}}{y}\left(\dfrac{\log(z_{\text{r}})}{\log(y/\mu)}\right)^{\lambda_{\text{r}}} & \text{if } y/\mu > z_{\text{r}}, \\ f_{\text{left}}(y/\mu)/\mu = f_{\text{mid}}(z_{\text{l}})\dfrac{z_{\text{l}}}{y}\left(\dfrac{\log(z_{\text{l}})}{\log(y/\mu)}\right)^{\lambda_{\text{l}}} & \text{if } 0 < y/\mu < z_{\text{l}}, \end{cases}$$

We analyse the three parts of the function (of μ) separately. We consider that all three parts exist; otherwise, one part (with f_{left}) has to be skipped.

We first consider that $\mu \in (0, y/z_{\text{r}})$. In this case, $f_{\nu,c}(y/\mu)/\mu = f_{\text{right}}(y/\mu)/\mu$. We have that

$$f_{\text{right}}(y/\mu)/\mu = f_{\text{mid}}(z_{\text{r}})\frac{z_{\text{r}}}{y}\left(\frac{\log(z_{\text{r}})}{\log(y/\mu)}\right)^{\lambda_{\text{r}}} \propto \left(\frac{1}{\log(y/\mu)}\right)^{\lambda_{\text{r}}}.$$

This function (of μ) is strictly increasing because $\lambda_{\text{r}} > 0$. Thus, for $\mu \in (0, y/z_{\text{r}})$,

$$f_{\text{right}}(y/\mu)/\mu \le f_{\text{mid}}(z_{\text{r}})z_{\text{r}}/y.$$

Analogously, we consider that $\mu \in (y/z_{\mathrm{l}}, \infty)$. In this case, $f_{\nu,c}(y/\mu)/\mu = f_{\mathrm{left}}(y/\mu)/\mu$. We have that

$$f_{\mathrm{left}}(y/\mu)/\mu = f_{\mathrm{mid}}(z_{\mathrm{l}})\frac{z_{\mathrm{l}}}{y}\left(\frac{\log(z_{\mathrm{l}})}{\log(y/\mu)}\right)^{\lambda_{\mathrm{l}}} \propto \left(\frac{-1}{\log(y/\mu)}\right)^{\lambda_{\mathrm{l}}}.$$

This function (of μ) is strictly decreasing because $\lambda_{\mathrm{l}} > 0$. Thus, for $\mu \in (0, y/z_{\mathrm{r}})$,

$$f_{\mathrm{left}}(y/\mu)/\mu \leq f_{\mathrm{mid}}(z_{\mathrm{l}})z_{\mathrm{l}}/y.$$

Finally, consider that $\mu \in (y/z_{\mathrm{l}}, y/z_{\mathrm{r}})$. In this case, $f_{\nu,c}(y/\mu)/\mu = f_{\mathrm{mid}}(y/\mu)/\mu$. We consider a larger domain $\mu \in (0, \infty)$ to find an upper bound for the function $f_{\mathrm{mid}}(y/\mu)/\mu$. We have

$$f_{\mathrm{mid}}(y/\mu)/\mu = f_{\mathrm{mid}}(y/\mu)(y/\mu)/y,$$

and thus maximizing this function with respect to μ is equivalent to maximizing $f_{\mathrm{mid}}(z)z/y$ with respect to $z \in (0, \infty)$. The derivative of the log of $f_{\mathrm{mid}}(z)\, z/y$ with respect to z is given by

$$\begin{aligned}\frac{\partial}{\partial z}\log\left[f_{\mathrm{mid}}(z)\, z/y\right] &= \frac{\partial}{\partial z}\left[-\nu z + \nu\log z + \nu\log\nu - \log(\Gamma(\nu)) - \log(y)\right] \\ &= -\nu + \frac{\nu}{z}.\end{aligned}$$

The root of this function is $z = 1$. If $z < 1$, $-\nu + \nu/z > 0$, meaning that the function $f_{\mathrm{mid}}(z)\, z/y$ is increasing from $z \in (0, 1)$. If $z > 1$, $-\nu + \nu/z < 0$, and the function is decreasing from $z \in (1, \infty)$. Thus, $f_{\mathrm{mid}}(1)/y$ is the maximum of $f_{\mathrm{mid}}(z)z/y$. The former is clearly larger than $f_{\mathrm{mid}}(z_{\mathrm{r}})z_{\mathrm{r}}/y$ and $f_{\mathrm{mid}}(z_{\mathrm{l}})z_{\mathrm{l}}/y$, which are the bounds for $f_{\mathrm{right}}(y/\mu)/\mu$ and $f_{\mathrm{left}}(y/\mu)/\mu$ on the parts of the domain where these functions are activated, respectively. Therefore, the upper bound for $f_{\nu,c}(y/\mu)/\mu$ is given by $f_{\nu,c}(1)/y = (e^{-1}\nu)^{\nu}/(y\Gamma(\nu))$.

□

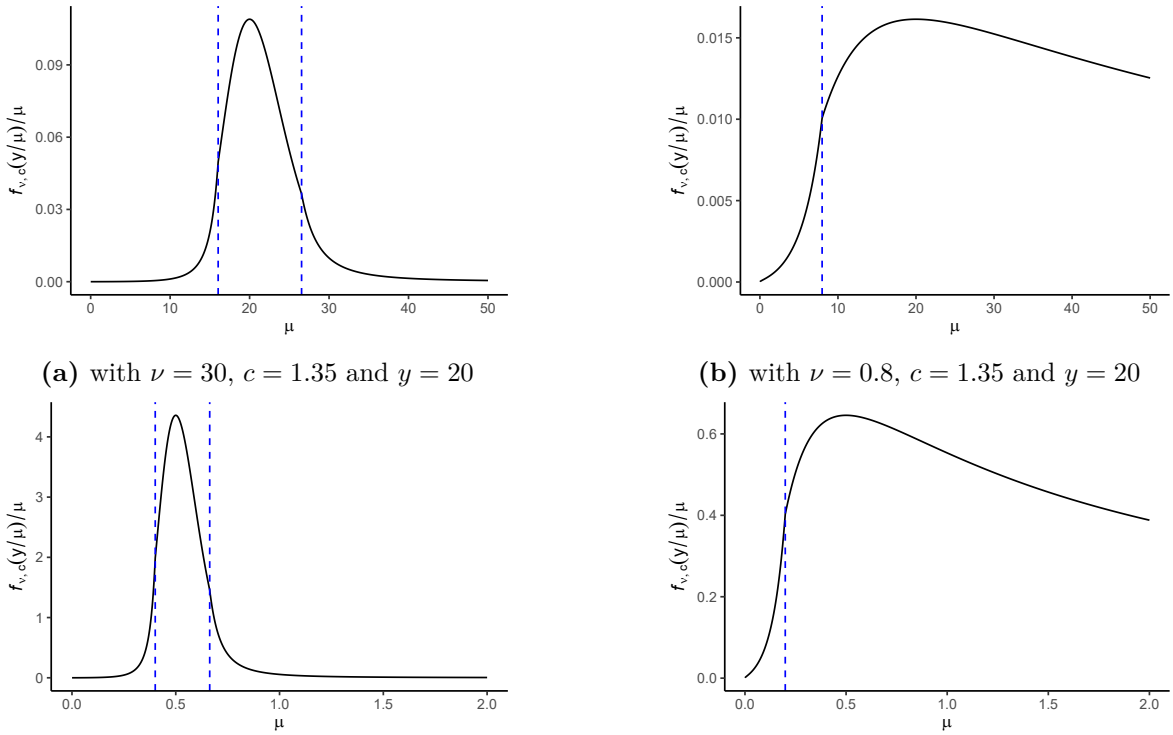
Figure 5.7 shows the function (of μ) $f_{\nu,c}(y/\mu)/\mu$ with $c = 1.35$, and with different values of ν and y. In **(b)** and **(d)**, f_{left} is never activated because $\nu < \sqrt{c}$.

Lemma 2.

If $\int_0^{\infty} \pi(\nu)\nu^{(n-p)/2}\,\mathrm{d}\nu < \infty$, *then* $\int_0^{\infty} \pi(\nu)\left[\frac{(e^{-1}\nu)^{\nu}}{\Gamma(\nu)}\right]^{n-p}\,\mathrm{d}\nu < \infty$.

Proof.

We will separate the integral into two parts: from 0 to a large positive constant ν^*, and from ν^* to ∞. The function $(e^{-1}\nu)^{\nu}/\Gamma(\nu)$ is strictly increasing, thus this function is bounded on $0 < \nu \leq \nu^*$ by $(e^{-1}\nu^*)^{\nu^*}/\Gamma(\nu^*)$. As ν gets large, $\Gamma(\nu)$ can be approximated by *Stirling's*



(a) with $\nu = 30$, $c = 1.35$ and $y = 20$

(b) with $\nu = 0.8$, $c = 1.35$ and $y = 20$

(c) with $\nu = 30$, $c = 1.35$ and $y = 0.5$

(d) with $\nu = 0.8$, $c = 1.35$ and $y = 0.5$

Figure 5.7. The function $f_{\nu,c}(y/\mu)/\mu$ with $c = 1.35$, and with different values of ν and y

formula, given by

$$\Gamma(\nu) \approx S(\nu) = \frac{\sqrt{2\pi}(\nu/e)^\nu}{\sqrt{\nu}}.$$

We thus have $(e^{-1}\nu)^\nu/\Gamma(\nu) \approx (e^{-1}\nu)^\nu/S(\nu)$. More precisely,

$$\frac{S(\nu)}{\Gamma(\nu)} \to 1 \Leftrightarrow \frac{(e^{-1}\nu)^\nu}{(e^{-1}\nu)^\nu}\frac{S(\nu)}{\Gamma(\nu)} \to 1 \Leftrightarrow \frac{\sqrt{2\pi}}{\sqrt{\nu}}\frac{(e^{-1}\nu)^\nu}{\Gamma(\nu)} \to 1.$$

Therefore, for all $\delta > 0$, we can find a ν^* such that for all $\nu \geq \nu^*$,

$$\frac{(e^{-1}\nu)^\nu}{\Gamma(\nu)} = \frac{(e^{-1}\nu)^\nu}{\Gamma(\nu)}\frac{\sqrt{2\pi}}{\sqrt{\nu}}\frac{\sqrt{\nu}}{\sqrt{2\pi}} \leq (1+\delta)\frac{\sqrt{\nu}}{\sqrt{2\pi}}.$$

Thus,

$$\begin{aligned}
\int_0^\infty \pi(\nu)\left[\frac{(e^{-1}\nu)^\nu}{\Gamma(\nu)}\right]^{n-p}\mathrm{d}\nu &= \int_0^{\nu^*}\pi(\nu)\left[\frac{(e^{-1}\nu)^\nu}{\Gamma(\nu)}\right]^{n-p}\mathrm{d}\nu + \int_{\nu^*}^\infty \pi(\nu)\left[\frac{(e^{-1}\nu)^\nu}{\Gamma(\nu)}\right]^{n-p}\mathrm{d}\nu \\
&\le \left[\frac{(e^{-1}\nu^*)^{\nu^*}}{\Gamma(\nu^*)}\right]^{n-p}\int_0^{\nu^*}\pi(\nu)\,\mathrm{d}\nu + \int_{\nu^*}^\infty \pi(\nu)\left[(1+\delta)\frac{\sqrt{\nu}}{\sqrt{2\pi}}\right]^{n-p}\mathrm{d}\nu \\
&= \left[\frac{(e^{-1}\nu^*)^{\nu^*}}{\Gamma(\nu^*)}\right]^{n-p}\int_0^{\nu^*}\pi(\nu)\,\mathrm{d}\nu + \frac{\int_{\nu^*}^\infty \pi(\nu)(1+\delta)\nu^{(n-p)/2}\,\mathrm{d}\nu}{2\pi^{(n-p)/2}} \\
&\le \left[\frac{(e^{-1}\nu^*)^{\nu^*}}{\Gamma(\nu^*)}\right]^{n-p}\int_0^{\nu^*}\pi(\nu)\,\mathrm{d}\nu + \frac{\int_0^\infty \pi(\nu)(1+\delta)\nu^{(n-p)/2}\,\mathrm{d}\nu}{2\pi^{(n-p)/2}} \\
&< \infty.
\end{aligned}$$

The last step is due to the condition $\int_0^\infty \pi(\nu)\nu^{(n-p)/2} < \infty$, and because $\pi(\nu)$ is a PDF. □

Proof of Proposition 5.2.1.

To prove this proposition, it suffices to show that the marginal $m(\mathbf{y})$ is finite, i.e.

$$\iint \pi(\boldsymbol{\beta},\nu)\prod_{i=1}^n\left(\frac{1}{\mu_i}\right)f_{\nu,c}\left(\frac{y_i}{\mu_i}\right)\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\nu < \infty.$$

To prove this, we first split the data points into two parts. The first part contains p data points, which will be used to perform a change of variables from $\boldsymbol{\beta}$ to $z_i = y_i/(\boldsymbol{x}_i^T\boldsymbol{\beta})$ for $i = 1,\ldots,p$. Without loss of generality, we choose the first p data points. For the rest of the $n-p$ data points, we bound $\prod_{i=1}^{n-p} f_{\nu,c}(y_i/\mu_i)/\mu_i$ by a function depending on ν, then we show that this bound, multiplied by $\pi(\nu)$, is integrable with respect to ν. We thus use the condition that $n \ge p$. When $n = p$, the proof is seen to be more simple, because the part

with the rest of the $n-p$ data points does not actually exist. We have

$$
\begin{aligned}
m(\mathbf{y}) &= \int_0^\infty \int_{\mathbb{R}^p} \pi(\boldsymbol{\beta}, \nu) \prod_{i=1}^{n} \frac{f_{\nu,c}(y_i/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta}))}{\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})} \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\nu \\
&= \int_0^\infty \int_{\mathbb{R}^p} \pi(\boldsymbol{\beta} \mid \nu)\, \pi(\nu) \prod_{i=1}^{p} \frac{f_{\nu,c}(y_i/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta}))}{\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})} \prod_{i=p+1}^{n} \frac{f_{\nu,c}(y_i/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta}))}{\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})} \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\nu \\
&\overset{a}{\leq} B \int_0^\infty \left(\int_{\mathbb{R}^p} \prod_{i=1}^{p} \frac{f_{\nu,c}(y_i/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta}))}{\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})} \mathrm{d}\boldsymbol{\beta} \right) \pi(\nu) \prod_{i=p+1}^{n} \frac{(e^{-1}\nu)^\nu}{y_i \Gamma(\nu)} \mathrm{d}\nu \\
&\overset{b}{=} B \int_0^\infty \left(\left| \det \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_p^T \end{pmatrix} \right|^{-1} \prod_{i=1}^{p} \int_0^\infty \frac{f_{\nu,c}(z_i)}{y_i} \mathrm{d}z_i \right) \pi(\nu) \prod_{i=p+1}^{n} \frac{(e^{-1}\nu)^\nu}{y_i \Gamma(\nu)} \mathrm{d}\nu \\
&\overset{c}{=} B \left| \det \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_p^T \end{pmatrix} \right|^{-1} \prod_{i=1}^{n} \frac{1}{y_i} \int_0^\infty \pi(\nu) \left[\frac{(e^{-1}\nu)^\nu}{\Gamma(\nu)} \right]^{n-p} \mathrm{d}\nu \\
&\overset{d}{<} \infty.
\end{aligned}
$$

In step a, we split the data points into two parts as we explained previously, and we use that $\pi(\boldsymbol{\beta} \mid \nu) \leq B$ with B a positive constant, and we bound the product of $f(y_i/\mu_i)/\mu_i$ for $i = p+1, \ldots, n$ by using Lemma 1. In step b, we perform a change of variables from $\boldsymbol{\beta}$ to $z_i = y_i/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})$, for $i = 1, \ldots, p$. For each i, we have $\left|\mathrm{d}z_i/\mathrm{d}\boldsymbol{\beta}\right| = \left|y_i \boldsymbol{x_i}^T/\exp(\boldsymbol{x_i}^T\boldsymbol{\beta})\right|$. The determinant is non-null because all explanatory variables are continuous. Indeed, consider the case $p = 2$ for instance; the determinant is different from 0 provided that $x_{12} \neq x_{22}$, which happens with probability 1. When any type of explanatory variables is considered, we need to be able to select p observations, say those with $\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_p}$, such that the matrix with rows $\boldsymbol{x}_{i_1}^T, \ldots, \boldsymbol{x}_{i_p}^T$ has a non-null determinant. In step c, we used that $f_{\nu,c}$ is a PDF. In step d, we used Lemma 2.

□

Appendix B.2: Proof of Proposition 5.2.2

Proof.

We first consider that $y \to \infty$. In this case, f_{right} is activated, and we have

$$
f_{\text{right}}(y) = f_{\text{mid}}(z_{\text{r}}) \frac{z_{\text{r}}}{y} \left(\frac{\log(z_{\text{r}})}{\log(y)} \right)^{\lambda_{\text{r}}},
$$

where z_{r}, $f_{\mathrm{mid}}(z_{\mathrm{r}})$, and λ_{r} depend only on ν and c. As $y/\mu \to \infty$, f_{right} is also activated in $f_{\nu,c,}(y/\mu)/\mu$. Thus,

$$\begin{aligned}\frac{f_{\mathrm{right}}(y/\mu)/\mu}{f_{\mathrm{right}}(y)} &= f_{\mathrm{mid}}(z_{\mathrm{r}})\frac{z_{\mathrm{r}}}{y}\left(\frac{\log(z_{\mathrm{r}})}{\log(y/\mu)}\right)^{\lambda_{\mathrm{r}}} \Big/ f_{\mathrm{mid}}(z_{\mathrm{r}})\frac{z_{\mathrm{r}}}{y}\left(\frac{\log(z_{\mathrm{r}})}{\log(y)}\right)^{\lambda_{\mathrm{r}}} \\ &= \left(\frac{\log(y)}{\log(y)-\log(\mu)}\right)^{\lambda_{\mathrm{r}}} \to 1, \quad \text{as} \quad y \to \infty,\end{aligned}$$

because μ is assumed to be a constant.

We consider now that $y \to 0$, with condition $c < \sqrt{\nu}$ and $\nu > 1$. In this case, the function f_{left} exists and is activated in both $f_{\nu,c}(y)$ and $f_{\nu,c}(y/\mu)/\mu$. We have

$$\begin{aligned}\frac{f_{\mathrm{left}}(y/\mu)/\mu}{f_{\mathrm{left}}(y)} &= f_{\mathrm{mid}}(z_{\mathrm{l}})\frac{z_{\mathrm{l}}}{y}\left(\frac{\log(z_{\mathrm{l}})}{\log(y/\mu)}\right)^{\lambda_{\mathrm{l}}} \Big/ f_{\mathrm{mid}}(z_{\mathrm{l}})\frac{z_{\mathrm{l}}}{y}\left(\frac{\log(z_{\mathrm{l}})}{\log(y)}\right)^{\lambda_{\mathrm{l}}} \\ &= \left(\frac{\log(y)}{\log(y)-\log(\mu)}\right)^{\lambda_{\mathrm{l}}} \to 1, \quad \text{as} \quad y \to 0.\end{aligned}$$

□

Appendix B.3: Proof of Theorem 5.2.1

Proof of Theorem 5.2.1.

We start with the proof of Result (a), which is quite lengthy. We next turn to the proofs of Results (b) and (c) which are shorter.

Let us assume for now that $m(\mathbf{y}) < \infty$ for all ω, and $m(\mathbf{y}_k) < \infty$. This will be proved later. We first observe that

$$\begin{aligned}\frac{m(\mathbf{y})}{m(\mathbf{y}_k)\prod_{i=1}^n [f_{\nu,c}(y_i)]^{s_i+r_i}} &= \frac{m(\mathbf{y})}{m(\mathbf{y}_k)\prod_{i=1}^n [f_{\nu,c}(y_i)]^{s_i+r_i}} \int_{\mathbb{R}^p} \pi(\boldsymbol{\beta} \mid \mathbf{y}_n)\,\mathrm{d}\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \frac{\pi(\boldsymbol{\beta})\prod_{i=1}^n [f_{\nu,c}(y_i/\mu_i)/\mu_i]^n}{m(\mathbf{y}_k)\prod_{i=1}^n [f_{\nu,c}(y_i)]^{s_i+r_i}}\,\mathrm{d}\boldsymbol{\beta} \\ &= \int_{\mathbb{R}^p} \pi(\boldsymbol{\beta} \mid \mathbf{y}_k)\prod_{i=1}^n \left[\frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)}\right]^{s_i+r_i}\mathrm{d}\boldsymbol{\beta}.\end{aligned}$$

We show that the last integral converges to 1 as $\omega \to \infty$. If we use Lebesgues's dominated convergence theorem to interchange the limit and the integral, we obtain that

$$\begin{aligned}
&\lim_{\omega\to\infty} \int_{\mathbb{R}^p} \pi(\boldsymbol{\beta} \mid \mathbf{y}_k) \prod_{i=1}^{n} \left[\frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)} \right]^{s_i+r_i} \mathrm{d}\boldsymbol{\beta} \\
&= \int_{\mathbb{R}^p} \lim_{\omega\to\infty} \pi(\boldsymbol{\beta} \mid \mathbf{y}_k) \prod_{i=1}^{n} \left[\frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)} \right]^{s_i+r_i} \mathrm{d}\boldsymbol{\beta} \\
&\overset{a}{=} \int_{\mathbb{R}^p} \pi(\boldsymbol{\beta} \mid \mathbf{y}_k) \times 1 \, \mathrm{d}\boldsymbol{\beta} \overset{b}{=} 1.
\end{aligned}$$

In step a, we use Proposition 5.2.2. In step b, we use that $\pi(\boldsymbol{\beta} \mid \mathbf{y}_k)$ is proper. Indeed, we notice in the proof of Proposition 5.2.1 that, if ν is fixed, the posterior distribution (of $\boldsymbol{\beta}$) is proper if the prior is bounded and if $k \geq p$. These conditions are satisfied because we assume that $\pi(\boldsymbol{\beta})$ is bounded, and that $k \geq d(r+s) + 2p - 1 \geq p$. Note that this implies that $m(\mathbf{y}_k) < \infty$ and $m(\mathbf{y}) < \infty$ for all ω.

However, to use Lebesgue's dominated convergence theorem, we need to prove that the intergral is bounded by an integrable function of $\boldsymbol{\beta}$ that does not depend on ω. Therefore, we need to show that

$$\begin{aligned}
&\pi(\boldsymbol{\beta} \mid \mathbf{y}_k) \prod_{i=1}^{n} \left[\frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)} \right]^{s_i+r_i} \propto \frac{\pi(\boldsymbol{\beta}) \prod_{i=1}^{n} [f_{\nu,c}(y_i/\mu_i)/\mu_i]^n}{\prod_{i=1}^{n} f_{\nu,c}(y_i)^{s_i+r_i}} \\
=&\pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \left[\frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)} \right]^{s_i} \prod_{i=1}^{n} \left[\frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)} \right]^{r_i} \prod_{i=1}^{n} \left[f_{\nu,c}(y_i/\mu_i)/\mu_i \right]^{k_i} \\
\leq& B \prod_{i=1}^{n} \left[\frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)} \right]^{s_i} \prod_{i=1}^{n} \left[\frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)} \right]^{r_i} \prod_{i=1}^{n} \left[f_{\nu,c}(y_i/\mu_i)/\mu_i \right]^{k_i} \\
=& g(\boldsymbol{\beta}) h(\omega),
\end{aligned}$$

with $g(\boldsymbol{\beta})$ an integrable function and $h(\omega)$ a bounded function, where we use $\pi(\boldsymbol{\beta}) \leq B$; the functions g and h are defined below.

According to the condition of the proposition, there are at least $d(r+s) + 2p - 1$ non-outliers in the data set. Without loss of generality, assume that the first $d(r+s)+2p-1$ points are non-outliers, i.e. $k_1, \ldots, k_{d(s+r)+2p-1} = 1$.

Step 1: We first choose p points among the non-outliers. Without loss of generality, we choose $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_p, y_p)$. We want to show that $g(\boldsymbol{\beta}) := B \prod_{i=1}^{p} f_{\nu,c}(y_i/\mu_i)/\mu_i$ is an

integrable function. We have

$$\int_{\mathbb{R}^p} B \prod_{i=1}^{p} \frac{f_{\nu,c}(y_i/\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}\,\mathrm{d}\boldsymbol{\beta} = B\left|\det\begin{pmatrix}\boldsymbol{x}_1^T\\ \vdots\\ \boldsymbol{x}_p^T\end{pmatrix}\right|^{-1} \prod_{i=1}^{p}\frac{1}{y_i}\prod_{i=1}^{p}\int_0^\infty f_{\nu,c}(z_i)\,\mathrm{d}z_i$$

$$= B\left|\det\begin{pmatrix}\boldsymbol{x}_1^T\\ \vdots\\ \boldsymbol{x}_p^T\end{pmatrix}\right|^{-1} \prod_{i=1}^{p}\frac{1}{y_i} < \infty$$

We use the change of variables $z_i = y_i/\mu_i = y_i/\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$, for $i = 1,\ldots,p$. The determinant term is different from 0 because $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p$ are linearly independent (because the covariates are continuous). Since these p observations are non-outlying, $\prod_{i=1}^{p}\frac{1}{y_i}$ is bounded and independent of ω.

Step 2: We want to show that the rest of the product, i.e.

$$h(\omega) := \prod_{i=d(s+r)+2p}^{n}\left[\frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)}\right]^{s_i}\prod_{i=d(s+r)+2p}^{n}\left[\frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)}\right]^{r_i}\prod_{i=p+1}^{n}\left[f_{\nu,c}(y_i/\mu_i)/\mu_i\right]^{k_i}$$

is bounded, and that the bound does not depend on $\boldsymbol{\beta}$ or ω.

In order to show this, let us split the domain of $\boldsymbol{\beta}$ as follows:

$$\begin{aligned}\mathbb{R}^p = {}& [\cap_i \mathcal{O}_i^c] \cup \left[\cup_i(\mathcal{O}_i \cap (\cap_{i_1}\mathcal{F}_{i_1}^c))\right] \cup \left[\cup_{i,i_1}(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap (\cap_{i_2\neq i_1}\mathcal{F}_{i_2}^c))\right]\\ &\cup\ldots\cup\left[\cup_{i,i_1,\ldots,i_{p-1}(i_j\neq i_s,\forall i_j,i_s\ \text{s.t.}\,j\neq s)}\left(\mathcal{O}_i\cap\mathcal{F}_{i_1}\cap\ldots\cap\mathcal{F}_{i_{p-1}}\cap\left(\cap_{i_p\neq i_1,\ldots,i_{p-1}}\mathcal{F}_{i_p}^c\right)\right)\right]\\ &\cup\left[\cup_{i,i_1,\ldots,i_p(i_j\neq i_s,\forall i_j,i_s\ \text{s.t.}\,j\neq s)}\left(\mathcal{O}_i\cap\mathcal{F}_{i_1}\cap\ldots\mathcal{F}_{i_p}\right)\right],\end{aligned}$$

where

$$\mathcal{O}_i := \begin{cases}\boldsymbol{\beta} : \log(b_i\omega) - \boldsymbol{x}_i^T\boldsymbol{\beta} < \log(\omega)/2 & \text{if } i\in\mathcal{I}_{\mathcal{R}},\\ \boldsymbol{\beta} : \boldsymbol{x}_i^T\boldsymbol{\beta} - \log\left(\frac{b_i}{\omega}\right) < \log(\omega)/2 & \text{if } i\in\mathcal{I}_{\mathcal{S}},\end{cases}$$

$$\mathcal{F}_i := \left\{\boldsymbol{\beta} : |\boldsymbol{x}_i^T\boldsymbol{\beta}| < \log(\omega)/\gamma\right\}\ \text{ if }\ i\in\mathcal{I}_{\mathcal{F}},$$

γ being a positive constant that will be defined. The sets $\mathcal{I}_{\mathcal{R}}$, $\mathcal{I}_{\mathcal{S}}$, and $\mathcal{I}_{\mathcal{F}}$ are defined as follows:

$$\begin{aligned}\mathcal{I}_{\mathcal{R}} &:= \{i : i\in\{d(s+r)+2p,\ldots,n\}\text{ and } r_i = 1\},\\ \mathcal{I}_{\mathcal{S}} &:= \{i : i\in\{d(s+r)+2p,\ldots,n\}\text{ and } s_i = 1\},\\ \mathcal{I}_{\mathcal{F}} &:= \{p+1,\ldots,d(s+r)+2p-1\}.\end{aligned}$$

Remember that the first p observations, which are non-outliers, have already been used for the purpose of integration in step 1. Thus, the index of non-outliers begins from $p+1$.

The set $\mathcal{O}_i$ represents the hyperplanes $\boldsymbol{x}_i^T\boldsymbol{\beta}$ characterized by different values of $\boldsymbol{\beta}$ satisfying $\log(b_i\omega) - \boldsymbol{x}_i^T\boldsymbol{\beta} < \log(\omega)/2$ for $i \in \mathcal{I}_{\mathcal{R}}$, and $\log(b_i/\omega) - \boldsymbol{x}_i^T\boldsymbol{\beta} < \log(\omega)/2$ for $i \in \mathcal{I}_{\mathcal{S}}$. The points $(\boldsymbol{x}_i, \log(b_i\omega))$ and $(\boldsymbol{x}_i, \log(b_i/\omega))$ can be seen as log transformations of big outliers and of small outliers, respectively, since $\omega \to \infty$.

Now we claim that $\mathcal{O}_i \cap \mathcal{F}_{i_1} \cdots \cap \mathcal{F}_{i_p} = \varnothing$ for all $i, i_1, \ldots, i_p$ with $i_j \neq i_s, \forall i_j, i_s$ such that $j \neq s$. To prove this, we use the fact that $\boldsymbol{x}_i$ (a vector of dimension p) can be expressed as a linear combination of $\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_p}$. This is true because all explanatory variables are continuous, therefore the space spanned by the vectors $\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_p}$ has dimension p.

As a result, if $\boldsymbol{\beta} \in \mathcal{F}_{i_1} \cap \ldots \cap \mathcal{F}_{i_p}$ and $\boldsymbol{x}_i = \sum_{s=1}^{p} a_s \boldsymbol{x}_{i_s}$, for some $a_1, \ldots, a_p \in \mathbb{R}$,

- if $i \in \mathcal{I}_R$,

$$\log(b_i\omega) - \boldsymbol{x}_i^T\boldsymbol{\beta} = \log(b_i\omega) - \left(\sum_{s=1}^{p} a_s \boldsymbol{x}_{i_s}\right)^T \boldsymbol{\beta} \overset{a}{\geq} \log(\omega) - \sum_{s=1}^{p} a_s \boldsymbol{x}_{i_s}^T\boldsymbol{\beta}$$
$$\overset{b}{>} \log(\omega) - \frac{\log(\omega)}{\gamma} \sum_{s=1}^{p} a_s \overset{c}{\geq} \log(\omega)/2;$$

- if $i \in \mathcal{I}_S$,

$$\log(b_i/\omega) - \boldsymbol{x}_i^T\boldsymbol{\beta} = \log(b_i/\omega) - \left(\sum_{s=1}^{p} a_s \boldsymbol{x}_{i_s}\right)^T \boldsymbol{\beta} < -\log(\omega) - \sum_{s=1}^{p} a_s \boldsymbol{x}_{i_s}^T\boldsymbol{\beta}$$
$$\overset{d}{<} -\log(\omega) + \left(\frac{\log(\omega)}{\gamma} \sum_{s=1}^{p} a_s\right) \overset{e}{\leq} -\log(\omega)/2.$$

In step a, we use that $b_i \geq 0$ and we simplify the form of the linear combination. In step b, because $\boldsymbol{\beta} \in \mathcal{F}_{i_1} \cap \ldots \cap \mathcal{F}_{i_p}$, we have $\boldsymbol{x}_i^T\boldsymbol{\beta} < \log(\omega)/\gamma$ for all $i \in \{i_1, \ldots, i_p\}$. Thus, $-\sum_{s=1}^{p} a_s \boldsymbol{x}_{i_s}^T\boldsymbol{\beta} > -(\log(\omega)/\gamma)\sum_{s=1}^{p} a_s$. In step c, we define the constant γ such that $\gamma \geq 2\sum_{s=1}^{p} a_s$ (we define γ such that it satisfies this inequality for any combination of i and $i_1, \ldots i_p$; without loss of generality, we consider that $\gamma \geq 1$). The proof is analogous in the case where $i \in \mathcal{I}_{\mathcal{S}}$. In step d, we use the fact that $\boldsymbol{x}_i^T\boldsymbol{\beta} > -\log(\omega)/\gamma$, thus $-\sum_{s=1}^{p} a_s \boldsymbol{x}_{i_s}^T\boldsymbol{\beta} < \log(\omega)/\gamma$. In step e, we use that γ is such that $\gamma \geq 2\sum_{s=1}^{p} a_s$. Therefore, we have that if $\boldsymbol{\beta} \in \mathcal{F}_{i_1} \cap \ldots \cap \mathcal{F}_{i_p}$, then $\boldsymbol{\beta} \notin \mathcal{O}_i$. This proves that $\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \ldots \cap \mathcal{F}_{i_p} = \varnothing$ for all $i, i_1, \ldots, i_p$ with $i_j \neq i_s, \forall i_j, i_s$ such that $j \neq s$. This result in turn implies that the domain of $\boldsymbol{\beta}$ can be

written as

$$\mathbb{R}^p = [\cap_i \mathcal{O}_i^c] \cup \left[\cup_i (\mathcal{O}_i \cap (\cap_{i_1} \mathcal{F}_{i_1}^c))\right] \cup \left[\cup_{i,i_1} (\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap (\cap_{i_2 \neq i_1} \mathcal{F}_{i_2}^c))\right]$$
$$\cup \ldots \cup \left[\cup_{i,i_1,\ldots,i_{p-1}(i_j \neq i_s \text{s.t.} j \neq s)} \left(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \ldots \cap \mathcal{F}_{i_{p-1}} \cap \left(\cap_{i_p \neq i_1,\ldots,i_{p-1}} \mathcal{F}_{i_p}^c\right)\right)\right].$$

This decomposition of $\mathbb{R}^p$ consists of $1 + \sum_{i=0}^{p-1} \binom{d(s+r)+p-1}{i}$ mutually exclusive sets given by $\cap_i \mathcal{O}_i^c$, $\cup_i (\mathcal{O}_i \cap (\cap_{i_1} \mathcal{F}_{i_1}^c))$,$\cup_i (\mathcal{O}_i \cap \mathcal{F}_i \cap (\cap_{i_1} \mathcal{F}_{i_2 \neq i_1}^c))$ for $i_1 \in \mathcal{I}_{\mathcal{F}}$, and so on.

We find an upper bound on each of these subsets. Because there is a finite number of subsets, we will be able to bound h by the maximal bound. Recall that

$$h(\omega) = \prod_{i=d(s+r)+2p}^{n} \left[\frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)}\right]^{r_i} \prod_{i=d(s+r)+2p}^{n} \left[\frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)}\right]^{s_i} \prod_{i=p+1}^{n} \left[f_{\nu,c}(y_i/\mu_i)/\mu_i\right]^{k_i}$$
$$:= A \times B \times C,$$

where each of A, B and C represents one of the products above.

Case 1: if $\boldsymbol{\beta} \in \cap_i \mathcal{O}_i^c$, we have

$$A \overset{a}{=} \prod_{i=d(s+r)+2p}^{n} \left(\frac{\log(y_i)}{\log(y_i) - \log(\mu_i)}\right)^{r_i \lambda_{\mathrm{r}}} \overset{b}{\leq} \prod_{i=d(s+r)+2p}^{n} \left(\frac{\log(b_i) + \log(\omega)}{\log(\omega)/2}\right)^{r_i \lambda_{\mathrm{r}}}$$
$$= \prod_{i=d(s+r)+2p}^{n} \left(2 + 2\frac{\log(b_i)}{\log(\omega)}\right)^{r_i \lambda_{\mathrm{r}}} \overset{c}{\leq} 3^{r\lambda_{\mathrm{r}}} \overset{d}{<} \infty.$$

In step a, if $\boldsymbol{\beta} \in \cap_i \mathcal{O}_i^c$, it means that $y_i/\mu_i = b_i\omega/\mu_i \geq \sqrt{\omega}$ for $i \in \mathcal{I}_{\mathcal{R}}$. We are thus sure that $b_i\omega/\mu_i$ and $b_i\omega$ are both on the right tail of $f_{\nu,c}$, i.e. f_{right}. In step b, we use $\log(b_i\omega) - \boldsymbol{x}_i^T\boldsymbol{\beta} \geq \log(\omega)/2$. In step c, we have $2\log(b_i)/\log(\omega) \leq 1$ for large enough ω. In step d, for any fixed ν, $3^{r\lambda_r}$ is finite given that λ_r, which depends only on c and ν, is finite.

For the part B, we have

$$B \overset{a}{=} \prod_{i=d(s+r)+2p}^{n} \left(\frac{\log(y_i)}{\log(y_i) - \log(\mu_i)}\right)^{s_i \lambda_{\mathrm{l}}} \overset{b}{\leq} \prod_{i=d(s+r)+2p}^{n} \left(\frac{\log(b_i) - \log(\omega)}{-\log(\omega)/2}\right)^{s_i \lambda_{\mathrm{l}}}$$
$$= \prod_{i=d(s+r)+2p}^{n} \left(2 - 2\frac{\log(b_i)}{\log(\omega)}\right)^{s_i \lambda_{\mathrm{l}}} < 2^{s\lambda_{\mathrm{l}}} \overset{c}{<} \infty.$$

The proof is analogous to the part B. In step a, if $\boldsymbol{\beta} \in \cap_i \mathcal{O}_i^c$, it means that $y_i/\mu_i = (b_i/\omega)/\mu_i \leq 1/\sqrt{\omega}$ for $i \in \mathcal{I}_{\mathcal{S}}$. We are thus sure that $(b_i/\omega)/\mu_i$ and b_i/ω are both on the left tail of $f_{\nu,c}$, i.e. f_{left}. In step b, we use $\log(y_i) - \boldsymbol{x}_i^T\boldsymbol{\beta} \leq -\log(\omega)/2$, and it is noted that $\log(y_i) = \log(b_i) - \log(\omega) < 0$. In step c, $2^{s\lambda_{\mathrm{l}}}$ is finite given that λ_{l} depending on c and ν is finite.

For the part C, we have

$$\prod_{i=p+1}^{n}\left[f_{\nu,c}(y_i/\mu_i)/\mu_i\right]^{k_i} \overset{a}{\leq} \prod_{i=p+1}^{n}\left[\frac{(e^{-1}\nu)^{\nu}}{y_i\Gamma(\nu)}\right]^{k_i} < \infty.$$

In step a, according to Lemma 1, $f_{\nu,c}(y/\mu)/\mu$ is upper bounded by $(e^{-1}\nu)^{\nu}/(y\Gamma(\nu))$ for any value of μ, when y is considered fixed.

We conclude that in this case, $A \times B \times C$ is bounded.

Case 2: consider now that $\boldsymbol{\beta}$ belongs to one of the $\sum_{i=1}^{p-1}\binom{d(s+r)+p-1}{i}$ mutually exclusive sets $\cup_i(\mathcal{O}_i \cap (\cap_{i_1}\mathcal{F}_{i_1}^c))$, $\cup_i(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap (\cap_{i_2\neq i_1}\mathcal{F}_{i_2}^c))$ for $i_1 \in \mathcal{I}_{\mathcal{F}}$, etc.

Consider part A, B and C separately, we have

$$\begin{aligned}
A &= \left[\prod_{i=d(s+r)+2p}^{n}\frac{1}{\mu_i}f_{\nu,c}(y_i/\mu_i)\right]^{r_i}\prod_{i=d(s+r)+2p}^{n}\left[\frac{1}{f_{\nu,c}(y_i)}\right]^{r_i} \\
&\propto \left[\prod_{i=d(s+r)+2p}^{n}\frac{1}{\mu_i}f_{\nu,c}(y_i/\mu_i)\right]^{r_i}\prod_{i=d(s+r)+2p}^{n}\left[y_i\left(\log(y_i)\right)^{\lambda_{\mathrm{r}}}\right]^{r_i} \\
&= \left[\prod_{i=d(s+r)+2p}^{n}\frac{y_i}{\mu_i}f_{\nu,c}(y_i/\mu_i)\right]^{r_i}\prod_{i=d(s+r)+2p}^{n}\left(\log(b_i\omega)\right)^{r_i\lambda_{\mathrm{r}}} \\
&\overset{a}{\leq} \left(\frac{(e^{-1}\nu)^{\nu}}{\Gamma(\nu)}\right)^{r}\prod_{i=d(s+r)+2p}^{n}\left(\log(b_i\omega)\right)^{r_i\lambda_{\mathrm{r}}}.
\end{aligned} \tag{5.2.2}$$

In step a, we can deduce from Lemma 1 that, viewed as a function of μ, $(y/\mu)f_{\nu,c}(y/\mu)$ is bounded by $(e^{-1}\nu)^{\nu}/\Gamma(\nu)$, for all ν, c, and y.

Analogously, for part B we have

$$\begin{aligned}
B &= \left[\prod_{i=d(s+r)+2p}^{n}\frac{1}{\mu_i}f_{\nu,c}(y_i/\mu_i)\right]^{s_i}\prod_{i=d(s+r)+2p}^{n}\left[\frac{1}{f_{\nu,c}(y_i)}\right]^{s_i} \\
&= \left[\prod_{i=d(s+r)+2p}^{n}\frac{1}{\mu_i}f_{\nu,c}(y_i/\mu_i)\right]^{s_i}\prod_{i=d(s+r)+2p}^{n}\left[\frac{y_i}{z_{\mathrm{l}}}\left(\frac{\log(y_i)}{\log(z_l)}\right)^{\lambda_{\mathrm{l}}}\right]^{s_i} \\
&\overset{a}{\propto} \left[\prod_{i=d(s+r)+2p}^{n}\frac{y_i}{\mu_i}f_{\nu,c}(y_i/\mu_i)\right]^{s_i}\prod_{i=d(s+r)+2p}^{n}\left(-\log(b_i/\omega)\right)^{s_i\lambda_{\mathrm{r}}} \\
&\overset{b}{\leq} \left(\frac{(e^{-1}\nu)^{\nu}}{\Gamma(\nu)}\right)^{s}\prod_{i=d(s+r)+2p}^{n}\left(\log(\omega/b_i)\right)^{s_i\lambda_{\mathrm{l}}} \\
&\overset{c}{\leq} \left(\frac{(e^{-1}\nu)^{\nu}}{\Gamma(\nu)}\right)^{s}\prod_{i=d(s+r)+2p}^{n}\left(\log(b_i\omega)\right)^{s_i\lambda_{\mathrm{l}}}.
\end{aligned} \tag{5.2.3}$$

In step a, we change the sign of $\log(b_i/\omega)$ because $\log(z_{\mathrm{l}}) < 0$. In step b, we bound $(y/\mu)f_{\nu,c}(y/\mu)$ by $(e^{-1}\nu)^\nu/\Gamma(\nu)$. In step c, we use the fact that $b_i \geq 1$.

Let us discuss now the part C. We have shown previously that in any of the sets to which $\boldsymbol{\beta}$ can belong, there are at most $p-1$ non-outlying points such that $|\boldsymbol{x}_i^T\boldsymbol{\beta}| < \log(\omega)/\gamma$. The case where that upper bound is attained is that where $\boldsymbol{\beta} \in \cup_i(\mathcal{O}_i \cap \mathcal{F}_{i_1} \cap \ldots \cap \mathcal{F}_{i_{p-1}} \cap (\cap_{i_p \neq i_1,\ldots,i_{p-1}}\mathcal{F}_{i_p}^c))$. Without loss of generality, suppose that all non-outlying points such that $|\boldsymbol{x}_i^T\boldsymbol{\beta}| < \log(\omega)/\gamma$ have index i belonging to $\{p+1,...,2p-1\}$. There are thus at least $d(r+s)+p-1-(p-1) = d(r+s)$ remaining non-outlying points such that $|\boldsymbol{x}_i^T\boldsymbol{\beta}| \geq \log(\omega)/\gamma$, with $i = 2p,\ldots,d(r+s)+2p-1$. Therefore, for these points, we are sure that y_i/μ_i is on the extremities of $f_{\nu,c}$, as either $y_i/\mu_i = \exp(\log(a_i) - \boldsymbol{x}_i^T\boldsymbol{\beta})) \geq a_i\omega^{1/\gamma} > z_{\mathrm{r}}$ or $y_i/\mu_i \leq a_i/\omega^{1/\gamma} < z_{\mathrm{l}}$, for large enough ω.

In the situation where y_i/μ_i is on the right side of the tails, i.e. $y_i/\mu_i \geq a_i\omega^{1/\gamma}$, we have

$$\begin{aligned}\frac{1}{\mu_i}f_{\nu,c}(y_i/\mu_i) &\propto \frac{1}{a_i}\left(\frac{\log(z_{\mathrm{r}})}{\log(a_i)-\log(\mu_i)}\right)^{\lambda_{\mathrm{r}}} \propto \left(\frac{1}{\log(a_i)-\log(\mu_i)}\right)^{\lambda_{\mathrm{r}}} \\ &\overset{a}{\leq} \left(\frac{1}{\log(a_i)+\log(\omega)/\gamma}\right)^{\lambda_{\mathrm{r}}} \overset{b}{\leq} \left(\frac{2\gamma}{\log(\omega)}\right)^{\lambda_r}.\end{aligned} \tag{5.2.4}$$

In step a, since $y_i/\mu_i \geq a_i\omega^{1/\gamma}$, we have $\mu_i \leq \omega^{-1/\gamma}$, thus $-\log(\mu_i) \geq \log(\omega)/\gamma$. In step b, we use the fact that $\log(a_i) \geq -\log(\omega)/(2\gamma)$ for large enough ω. We thus have $\log(a_i) + \log(\omega)/\gamma \geq \log(\omega)/(2\gamma)$.

In the situation where y_i/μ_i is on the left side of the tails, i.e. $y_i/\mu_i \leq a_i/\omega^{1/\gamma}$, we have

$$\begin{aligned}\frac{1}{\mu_i}f_{\nu,c}(y_i/\mu_i) &\propto \frac{1}{a_i}\left(\frac{\log(z_{\mathrm{l}})}{\log(a_i)-\log(\mu_i)}\right)^{\lambda_{\mathrm{l}}} \overset{a}{\propto} \left(\frac{1}{\log(\mu_i)-\log(a_i)}\right)^{\lambda_{\mathrm{l}}} \\ &\leq \left(\frac{1}{\log(\omega)/\gamma-\log(a_i)}\right)^{\lambda_{\mathrm{l}}} \overset{b}{\leq} \left(\frac{2\gamma}{\log(\omega)}\right)^{\lambda_{\mathrm{l}}}.\end{aligned} \tag{5.2.5}$$

In step a, we change the sign because $\log(z_{\mathrm{l}}) < 0$. In step b, we use the fact that $\log(a_i) \leq \log(\omega)/(2\gamma)$ for large enough ω.

The reason we consider these two cases is that we want to use a density of an "extreme non-outlier", i.e. $f_{\nu,c}(y_j/\mu_j)/\mu_j$ with j such that $\boldsymbol{\beta} \in \mathcal{F}_j^c$, to cancel each $\log(b_i\omega)$ at some power for $i \in \mathcal{I}_{\mathcal{R}} \cup \mathcal{I}_{\mathcal{S}}$ that appears in the bounds of A and B. As we explained, there are at least $d(r+s)$ extreme non-outliers that can be used. However, the major problem here is that we do not know how many extreme non-outliers among $d(r+s)$ are such that y_j/μ_j are on the right tail, and how many are on the left tail, depending on the value of $\boldsymbol{\beta}$. We thus have to consider all possible scenarios, including the worst scenario. We now present clearly how we bound each $\log(b_i\omega)$ for $i \in \mathcal{I}_{\mathcal{R}} \cup \mathcal{I}_{\mathcal{S}}$, by using the densities of extreme non-outliers.

Let us consider first a big outlying observation y_i, i.e. $i \in \mathcal{I}_{\mathcal{R}}$. Recall that we define $d = \max\{\lambda_{\mathrm{l}}/\lambda_{\mathrm{r}}, \lambda_{\mathrm{r}}/\lambda_{\mathrm{l}}\}$. This definition is possible for all ν fixed, because λ_{l} and λ_{r} can be

computed, and are positive and finite. We take d non-outliers among the $d(r+s)$ extreme non-outliers that are thus such that $\boldsymbol{\beta} \in \cap_{j\in\{i_1,\ldots,i_d\}}\mathcal{F}_j^c$. In other words, all these points are such that $|\boldsymbol{x}_j^T\boldsymbol{\beta}| \geq \log(\omega)/\gamma$. There are two possible cases.

Case 1: there is at least one point among the d points such that $y_j/\mu_j \geq a_j\omega^{1/\gamma}$, implying that the density is evaluated on the right tail. In this case, we have

$$\begin{aligned}
&\log(b_i\omega)^{\lambda_\mathrm{r}} \prod_{j=i_1}^{i_d} \frac{1}{\mu_j} f_{\nu,c}(y_j/\mu_j) \\
&\overset{a}{\leq} \left[\log(b_i\omega)^{\lambda_\mathrm{r}}\right] \left(\frac{2\gamma}{\log(\omega)}\right)^{\lambda_\mathrm{r}} \\
&= \left(\frac{2\gamma\log(b_i\omega)}{\log(\omega)}\right)^{\lambda_\mathrm{r}} \overset{b}{\leq} (4\gamma)^{\lambda_\mathrm{r}} \overset{c}{<} \infty.
\end{aligned}$$

In step a, we take one point such that $y_j/\mu_j \geq a_j\omega^{1/\gamma}$. We bound $f_{\nu,c}(y_j/\mu_j)/\mu_j$ by the bound presented in (5.2.4), and we bound the rest of the points by 1 using the bounds in (5.2.4) and (5.2.5), given that $2\gamma/\log(\omega) \leq 1$ for large enough ω. In step b, we have that $\log(b_i\omega)/\log(\omega) = (\log(b_i)+\log(\omega))/\log(\omega) \leq 2$, as $\log(b_i)/\log(\omega) \leq 1$ for large enough ω. In step c, since every term is a well-defined constant, it is finite.

Case 2: no point among the d points is such that $y_j/\mu_j \geq a_j\omega^{1/\gamma}$, implying that the density of every point is evaluated on the left tail. In this case, we have

$$\begin{aligned}
&\log(b_i\omega)^{\lambda_\mathrm{r}} \prod_{j=i_1}^{i_d} \frac{1}{\mu_j} f_{\nu,c}(y_j/\mu_j) \\
&\overset{a}{\leq} \left[\log(b_i\omega)^{\lambda_\mathrm{r}}\right] \left(\frac{2\gamma}{\log(\omega)}\right)^{d\lambda_\mathrm{l}} \\
&= \left(\frac{2\gamma\log(b_i\omega)}{\log(\omega)}\right)^{\lambda_\mathrm{r}} \left(\frac{2\gamma}{\log(\omega)}\right)^{d\lambda_\mathrm{l}-\lambda_\mathrm{r}} \\
&\overset{b}{\leq} (4\gamma)^{\lambda_\mathrm{r}} < \infty.
\end{aligned}$$

In step a, we bound every term $f_{\nu,c}(y_j/\mu_j)/\mu_j$ by the bound in (5.2.5). In step b, since we have $d = \max\{\lambda_\mathrm{l}/\lambda_\mathrm{r}, \lambda_\mathrm{r}/\lambda_\mathrm{l}\}$, then $d\lambda_\mathrm{l} \geq \lambda_\mathrm{r}$. We also use that $2\gamma/\log(\omega) \leq 1$, and $\log(b_i\omega)/\log(\omega) = (\log(b_i)+\log(\omega))/\log(\omega) \leq 2$, as $\log(b_i)/\log(\omega) \leq 1$ for large enough ω.

We showed that we can use the product of the densities of d extreme non-outliers to offset $\log(b_i\omega)^{\lambda_\mathrm{r}}$ for $i \in \mathcal{I}_\mathcal{R}$, so that the whole product is bounded. The approach is analogous for

small outliers, i.e. $i \in \mathcal{I}_{\mathcal{S}}$. Therefore, if we multiply now A, B and C, we obtain that

$$
\begin{aligned}
A \times B \times C &\overset{a}{\leq} \left[\left(\frac{(e^{-1}\nu)^{\nu}}{\Gamma(\nu)} \right)^{s+r} \prod_{i=d(s+r)+2p}^{n} \left(\log(b_i\omega)\right)^{r_i\lambda_{\mathrm{r}}+s_i\lambda_{\mathrm{l}}} \right] \prod_{i=p+1}^{n} \left(\frac{f_{\nu,c}(y_i/\mu_i)}{\mu_i} \right)^{k_i} \\
&\overset{b}{\leq} \left[\left(\frac{(e^{-1}\nu)^{\nu}}{\Gamma(\nu)} \right)^{s+r} \prod_{i=d(s+r)+2p}^{n} \left(\log(b_i\omega)\right)^{r_i\lambda_{\mathrm{r}}+s_i\lambda_{\mathrm{l}}} \right] \prod_{i=p+1}^{2p-1} \left(\frac{(e^{-1}\nu)^{\nu}}{y_i\Gamma(\nu)} \right) \prod_{i=2p}^{n} \left(\frac{f_{\nu,c}(y_i/\mu_i)}{\mu_i} \right)^{k_i} \\
&\overset{c}{\leq} \left[\left(\frac{(e^{-1}\nu)^{\nu}}{\Gamma(\nu)} \right)^{s+r} \right] \left[\prod_{i=p+1}^{2p-1} \left(\frac{(e^{-1}\nu)^{\nu}}{y_i\Gamma(\nu)} \right) \right] \left[\prod_{i=d(r+s)+2p}^{n} \left(\frac{f_{\nu,c}(y_i/\mu_i)}{\mu_i} \right)^{k_i} \right] \left[(4\gamma)^{r\lambda_{\mathrm{r}}+s\lambda_{\mathrm{l}}} \right] \\
&\overset{d}{\leq} \left[\left(\frac{(e^{-1}\nu)^{\nu}}{\Gamma(\nu)} \right)^{n-p-d(r+s)} \left(\prod_{i=p+1}^{2p-1} \frac{1}{y_i} \right) \left(\prod_{i=d(r+s)+2p}^{n} \frac{1}{y_i} \right)^{k_i} \right] \left[(4\gamma)^{r\lambda_{\mathrm{r}}+s\lambda_{\mathrm{l}}} \right] \\
&\overset{e}{=} \left[\left(\frac{(e^{-1}\nu)^{\nu}}{\Gamma(\nu)} \right)^{n-p-d(r+s)} \left(\prod_{i=p+1}^{2p-1} \frac{1}{a_i} \right) \left(\prod_{i=d(r+s)+2p}^{n} \frac{1}{a_i} \right)^{k_i} \right] \left[(4\gamma)^{r\lambda_{\mathrm{r}}+s\lambda_{\mathrm{l}}} \right] \\
&\overset{f}{<} \infty.
\end{aligned}
$$

In step a, we bound the part A and B by expressions that are previously shown (see (5.2.2) and (5.2.3)). In step b, we bound the density of y_i, for $i = p+1, \ldots, 2p-1$, by $(e^{-1}\nu)^{\nu}/(y_i\Gamma(\nu))$ (see Lemma 1). Recall that these are non-outliers such that $|\boldsymbol{x}_i^T\boldsymbol{\beta}| < \log(\omega)/\gamma$. In step c, we simplify each $\log(b_i\omega)^{\lambda_{\mathrm{r}}}$ and $\log(b_i\omega)^{\lambda_{\mathrm{l}}}$ by multiplying by the product of densities of d non-extreme outliers, as we have explained earlier. In step d, we bound the rest of the non-outliers by $(e^{-1}\nu)^{\nu}/(y_i\Gamma(\nu))$. If we consider all the k non-outliers, there are p non-outliers that are taken for the change of variables and to integrate over $\boldsymbol{\beta}$ at the beginning, $d(r+s)$ are used to offset the outliers, and we bound $p-1$ others that are not extreme. There are thus still $k - p - (p-1) - d(r+s) = k - 2p - d(r+s) + 1$ non-outliers left, that need to be considered. The proof is simpler and is still valid if there is no point left. The condition of this theorem $k \geq d(r+s) + 2p - 1$ is to make sure that we have enough non-outlying points to bound the whole product. In step e, every y_i in the expression is a non-outlying observation, thus is equal to a_i. Finally, in step f, the whole expression is bounded since all terms are constant.

Therefore, $h(\omega) = A \times B \times C$ is bounded. This completes our proof of Result (a).

We now turn to the proof of Result (b). We have that

$$
\pi(\boldsymbol{\beta} \mid \mathbf{y}) = \pi(\boldsymbol{\beta} \mid \mathbf{y}_k) \frac{m(\mathbf{y}_k) \prod_{i=1}^{n} f_{\nu,c}(y_i)^{s_i+r_i}}{m(\mathbf{y})} \left[\prod_{i=1}^{n} \frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)} \right]^{s_i+r_i},
$$

and

$$\frac{m(\mathbf{y}_k)\prod_{i=1}^{n} f_{\nu,c}(y_i)^{s_i+r_i}}{m(\mathbf{y})}\left[\prod_{i=1}^{n}\frac{f_{\nu,c}(y_i/\mu_i)/\mu_i}{f_{\nu,c}(y_i)}\right]^{s_i+r_i} \to 1,$$

as $\omega \to \infty$, for any $\boldsymbol{\beta} \in \mathbb{R}^p$, using Result (a) and Proposition 5.2.2. We also showed that $\pi(\boldsymbol{\beta} \mid \mathbf{y}_k)$ is proper. This concludes the proof of Result (b).

We now finish with the proof of Result (c). This result is a direct consequence of Result (b) using Scheffé's theorem (Scheffé, 1947).

□

Figure 5.8 shows λ_l and λ_r as a function of ν with $c = 1.35$. The numerical calculation suggests that $\lambda_\text{l} \geq \lambda_\text{r}$ for all ν. Moreover, the ratio $d = \lambda_\text{l}/\lambda_\text{r}$ seems to strictly decrease as ν increases.

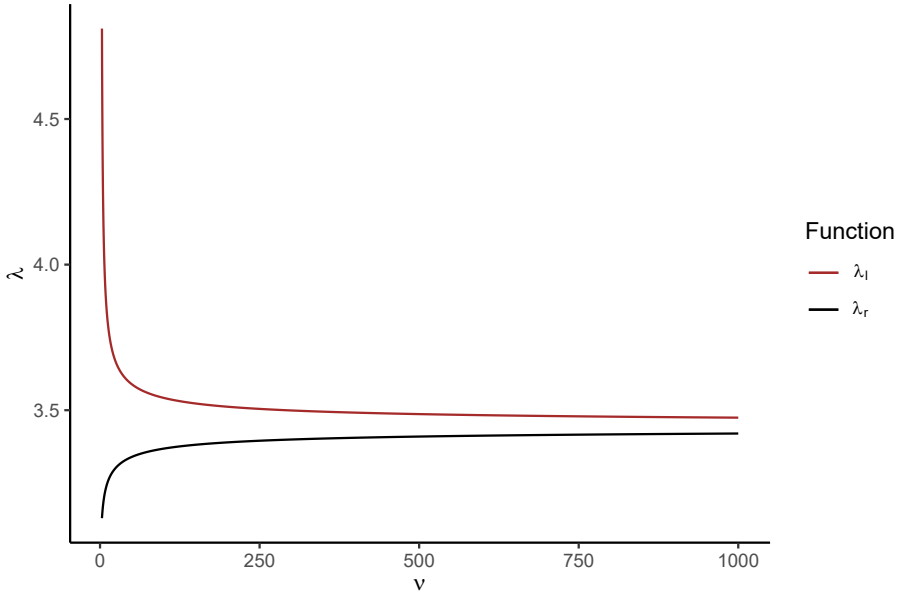


Figure 5.8. Comparison between λ_l and λ_r as a function of ν with $c = 1.35$

References

Beath, K. 2021, *robmixglm: Robust Generalized Linear Models (GLM) using Mixtures*. R package version 1.2-1.

Beaton, A. E. and J. W. Tukey. 1974, The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data, *Technometrics*, vol. 16, 2, p. 147–185.

Bedrick, E. J., R. Christensen and W. Johnson. 1996, A New Perspective on Priors for Generalized Linear Models, *Journal of the American Statistical Association*, vol. 91, 436, p. 1450–1460.

Bissiri, P. G., C. Holmes and S. Walker. 2013, A General Framework for Updating Belief Distributions, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78.

Cantoni, E. and E. Ronchetti. 2001, Robust Inference for Generalized Linear Models, *Journal of American Statistical Association*, vol. 96, 455, p. 1022–1030.

Cantoni, E. and E. Ronchetti. 2006, A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures, *Journal of Health Economics*, vol. 25, 2, p. 198–213.

Casals, M., M. Girabent-Farrés and J. L. Carrasco. 2014, Methodological Quality and Reporting of Generalized Linear Mixed Models in Clinical Medicine (2000–2012): A Systematic Review, *PLoS ONE*, vol. 9, 11.

Desgagné, A. 2015, Robustness to Outliers in Location-Scale Parameter Model Using Log-Regularly Varying Distributions, *The Annals of Statistics*, vol. 43, 4, p. 1568–1595.

Dobson, A. and A. Barnett. 2018, An Introduction to Generalized Linear models, Taylor Francis Group.

Duane, S., A. D. Kennedy, B. J. Pendleton and D. Roweth. 1987, Hybrid monte carlo, *Physics Letters B*, vol. 195, p. 216–222.

Gagnon, P., A. Desgagné and M. Bédard. 2020, A New Bayesian Approach to Robustness Against Outliers in Linear Regression, *Bayesian Analysis*, vol. 15, 2, p. 389–414.

Goldburd, M., A. Khare, D. Tevet and D. Guller. 2019, Generalized Linear Models for Insurance Rating, Casualty Actuarial Society.

Hadi, A. S. and J. S. Simonoff. 1993, Procedures for the Identification of Multiple Outliers in Linear Models, *Journal of the American Statistical Association*, vol. 88, 424, p. 1264–1272.

Hampel, F. R. 1974, The Influence Curve and Its Role in Robust Estimation, *Journal of the American Statistical Association*, vol. 69, 346, p. 383–393.

Huber, P. J. 1964, Robust Estimation of a Location Parameter, *The Annals of Mathematical Statistics*, vol. 35, 1, p. 73–101.

Huber, P. J. 1973, Robust Regression: Asymptotics, Conjectures and Monte Carlo, *The Annals of Statistics*, vol. 1, 5, p. 799–821.

Ibrahim, J. G. and P. W. Laud. 1991, On Bayesian Analysis of Generalized Linear Models Using Jeffreys's Prior, *Journal of the American Statistical Association*, vol. 86, 416, p. 981–986.

Kunsch, H. R., L. A. Stefanski and R. J. Carroll. 1989, Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, with Applications to Generalized Linear Models, *Journal of the American Statistical Association*, vol. 84, 406, p. 460–466.

Maechler, M., P. Rousseeuw, C. Croux, V. Todorov, M. Ruckstuhl, A.and Salibian-Barrera, T. Verbeke, M. Koller, E. L. T. Conceicao and M. Anna di Palma. 2020, *robustbase: Basic Robust Statistics*. R package version 0.93-6.

Marazzi, A. and V. J. Yohai. 2004, Adaptively truncated maximum likelihood regression with asymmetric errors, *Journal of Statistical Planning and Inference*, vol. 122, p. 271–291.

McCullagh, P. and J. A. Nelder. 1983, Generalized Linear Models, London: Chapman and Hal.

Menezes, D., D. Prata, A. Secchi and J. Pinto. 2021, A Review on Robust M-Estimators for Regression Analysis, *Computers & Chemical Engineering*, vol. 147, p. 107 254.

Nelder, J. A. and R. W. M. Wedderburn. 1972, Generalized Linear Models, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 135, 3, p. 370–384.

Pregibon, D. 1982, Resistant Fits for Some Commonly Used Logistic Models with Medical Applications, *Biometrics*, vol. 38, 2, p. 485–498.

R Core Team. 2021, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ronchetti, E. 2006, The historical development of robust statistics, in *Proceedings of the 7th International Conference on Teaching Statistics (ICOTS-7)*, p. 2–7.

Ryan, T. P. 1997, Modern Regression Methods, New York: Wiley.

Scheffé, H. 1947, A Useful Convergence Theorem for Probability Distributions, *The Annals of Mathematical Statistics*, vol. 18, 3, p. 434–438.

Stefanski, L. A., R. J. Carroll and D. Ruppert. 1986, Optimally Bounded Score Functions for Generalized Linear Models with Applications to Logistic Regression, *Biometrika*, vol. 73, 2, p. 413–424.

West, M. 1984, Outlier Models and Prior Distributions in Bayesian Linear Regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 46, 3, p. 431–439.