Doctoral Thesis

Novel Regularization Models for Dynamic and Discrete Response Data

Author:
Hamed Haseli Mashhadi

Supervisor:
Dr. Veronica Vinciotti

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Department of Mathematics

February 2017

Declaration of Authorship

I, Hamed Haseli Mashhadi, declare that this thesis titled, "Novel Regularization Models for Dynamic and Discrete Response Data" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism."

Dave Barry

"Everything is related to everything else, but near things are more related than distant things."

Waldo Tobler

Abstract

Regularized regression models have gained popularity in recent years. The addition of a penalty term to the likelihood function allows parameter estimation where traditional methods fail, such as in the $p \gg n$ case. The use of an l_1 penalty in particular leads to simultaneous parameter estimation and variable selection, which is rather convenient in practice. Moreover, computationally efficient algorithms make these methods really attractive in many applications. This thesis is inspired by this literature and investigates the development of novel penalty functions and regression methods within this context.

In particular, Chapter § 2 deals with linear models for time-dependent response and explanatory variables. This is beyond the independent framework which is common to many of the developed regularized regression models. We propose to account for the time dependency in the data by explicitly adding autoregressive terms to the response variable together with an autoregressive process for the residuals. In addition, the use of a l_1 penalized likelihood approach for parameter estimation leads to automatic order and variable selection and makes this method feasible for high-dimensional data. Theoretical properties of the estimators are provided and an extensive simulation study is performed. Finally, we show the application of the model on air pollution and stock market data and discuss its implementation in the R package `DREGAR`, which is freely available in CRAN.

In Chapter § 3, we develop a new penalty function. Despite all the advantages of the l_1 penalty, this penalty is not differentiable at zero, and neither are the alternatives that are proposed in the literature. The only exception is the ridge penalty, which does not lead to variable selection. Motivated by this gap, and noting the advantages that a differentiable penalty can give, such as increased computational efficiency in some cases and the derivation of more accurate model selection criteria, we develop a new penalty function based on the error function. We study the theoretical properties of this function and of the estimators obtained in a regularized regression context. Finally, we perform a simulation study and we use the new penalty to analyse a diabetes and prostate cancer dataset. The new method is implemented in the R package `DLASSO`, that is freely available in CRAN.

Finally, Chapter § 4 deals with regression models for discrete response data, which is frequently collected in many application areas. In particular, we consider a discrete Weibull regression model that has recently been introduced in the literature. In this chapter, we propose the first Bayesian implementation of this model. We consider a general parametrization, where both parameters of the discrete Weibull distribution can be conditioned on the predictors, and show theoretically how, under a uniform non-informative prior, the posterior distribution is proper with finite moments. In addition, we consider closely the case of Laplace priors for parameter shrinkage and variable selection. A simulation study and the analysis of four real datasets of medical records show the applicability of this approach to the analysis of count data. The method is implemented in the R package `BDWreg`, which is freely available in CRAN.

Acknowledgements

I take this opportunity to express my very best gratitude to my first supervisor, Dr. Veronica Vinciotti for her countless hours of reflecting, reading, encouraging, and most of all patience throughout the entire process. Also, my special thanks go to Dr. Keming Yu for all his support and helping as a second supervisor.

I would like to acknowledge and thank the members of staff for allowing me to conduct my research and providing any assistance requested. Special thanks go to Dr. Paresh Date for his continued support, kindness and encouragement.

Finally, I would also like to thank my parents for their wise counsel and sympathetic ear. You are always there for me.

Dedication

I dedicate this work to my mother. A special feeling of gratitude to my loving parents.
I also dedicate this work to my supervisors, Dr. Veronica Vinciotti and Dr. Keming Yu, that supported me throughout the process. I will always appreciate all they have done.
I dedicate this work and give special thanks to my best friend Dr.Amir Ali Mohagheghi and my wonderful sister Mrs. Nazanin Haseli Mashhadi for being there for me throughout the entire doctorate program. Both of you have been my best cheerleaders.

Contents

List of Figures

List of Tables

List of Acronyms

ACF ..... Autocorrelation function

AIC ..... Akaike information criterion

AR ..... Autoregressive

ARMA ..... Autoregressive-moving averages

BDWreg ..... Bayesian discrete Weibull regression

BF ..... Bays factor

BIC ..... Bayesian information criteria

BLUE ..... Best linear unbiased estimator

BPIC ..... Bayesian predictive information criteria

CAIC ..... Consistent Akaike information criterion

CHN ..... Cumulative half normal distribution

CI ..... Credible interval

CV ..... Cross validation

DIC ..... Deviance information criteria

DREGAR ..... Dynamic regression in the presence of autocorrelated residuals

DW ..... Discrete Weibull

EV ..... Error in variables

FN ..... False negative

FP ..... False positive

GCV ..... Generalized cross validation

GIC ..... Generalized information criteria

GLM ..... Generalized linear model

GVS ..... Gibbs variable selection

HN ..... Half normal distribution

HPDI ..... Highest posterior density interval

IFF ..... If and only if

IID ..... Independent and identically distributed

LARS ..... Least angle regression

LASSO ..... Least absolute shrinkage and selection operator

LP ..... Linear programming

LS ..... Least squares

LSE ..... Least square errors

MA ..... Moving averages

MCMC ..... Markov chain Monte Carlo

MDS ..... Martingale difference sequence

MH ..... Metropolis Hastings

MLE ..... Maximum likelihood estimator

MSE ..... Mean squared errors

MSPE ..... Mean squared prediction error

MVN ..... Multivariate normal distribution

NB ..... Negative Binomial

NMMAPS ..... National mortality morbidity and air pollution study

OLS ..... Ordinary least squares

PACF ..... Partial autocorrelation function

PPD ..... Prior predictive density

QIC ..... Quasi-likelihood information criteria

REGAR ..... Regression in the presence of AR residuals

REGARMA ..... Regression in the presence ARMA residuals

RHS ..... Right hand side

RJMCMC ..... Reversible jumps Markov chain Monte Carlo

RJMH ..... Reversible jumps Metropolis Hastings

SCAD ..... Smoothly clipped absolute deviation

SSVS ..... Stochastic search variable selection

TNR ..... True negative rate

TP ..... True positive

WLLN ..... Weak law of large numbers

Chapter 1

Introduction

1.1 Introduction

Since the first application of linear regression in [Galton, 1894], there have been so many publications in almost any field that make *regression* an essential part of statistical modelling. For example, a quick search on Google at the end of June 2015 revealed about $4,170,000$ results including the word "regression" and the number of publication is increasing as it is shown in Table (1.1).

Month	**June 2015**	**July**	**August**	**September**	**October**	**November**	**December**
Results	4,170,000	4,230,000	4,270,000	4,310,000	4,330,000	4,380,000	4,420,000

TABLE 1.1: Google scholar results for the word "regression" during seven consecutive months starting from June, 30 2015.

Let the general form of regression be $y = f(x) + e$ where y, x, f and e are response, covariates, link function and unknown error respectively. Imposing linearity on the link function f and assuming that x is a vector of r mutually independent variables and that the errors are independent result in the well-known linear regression model,

$$y = x\beta + e. \tag{1.1}$$

Then, given a data matrix, X of $T(> r)$ observations on the covariates and a noisy column vector response y, the least-squares (LS) method gives the Best Linear Unbiased Estimator (BLUE) of the parameters, provided the noise elements are independent and identically normally distributed, $e_i \sim N(0, \sigma^2 < \infty), i = 1, 2, \ldots, T$. Rewriting the problem using matrices and norms, LS minimizes the second norm of the errors with respect to the parameters that is

$$\arg\min_{\beta} ||y - X\beta||_2^2,$$

where $||z||_2 = \sqrt[2]{\sum_i^T z_i^2}$ for any vector $z = (z_1, z_2, \ldots, z_T)$. Solving this minimization problem for β leads to $\hat{\beta} = (X'X)^{-1}X'y$ and $\mathbb{V}ar(\hat{\beta}) = (X'X)^{-1}\sigma^2$ where $(')$ denotes the transpose of a matrix.

Various extensions of the traditional regression model have been developed in literature. For instance, replacing linearity with a *general* function results in so called non-parametric regression, see [Hollander et al., 2013, Gibbons and Chakraborti, 2003, Wasserman, 2006]. Allowing measurement error in the predictors is the subject of Error in Variables (EV) methods, see [Fuller, 2009, Carroll et al., 2006, Gustafson, 2003]. Moreover, if one looks at the estimation aspect of the problem, there are a number of inference procedures such as Robust, Minimax, Quantile etc, see [Du and Pardalos, 2013, Davino et al., 2013, Bloomfield and Steiger, 2012, Koenker, 2005, Lawrence and Arthur, 1990, Nawata, 1988, Wu, 1997] and references therein for a complete discussion about the corresponding methods.

Among different types of regression, we concentrate on a relatively young class, namely models and inference procedures for high-dimensional data. By this, we mean data that contain more variables than observations. It is well understood that linear models and classical multivariate methods do not properly handle problems with more variables than observations. This is due to the fact that they rely heavily on the inverse of $X'X$, that can be singular or ill-conditioned in high-dimensional settings. For example in the linear regression presented in (1.1) we get,

$$MSE(\hat{\beta}) = \sigma^2 tr\{(X'X)^{-1}\},$$

where $tr\{\cdot\}$ denotes the trace of a matrix. It is important to recognise that inverse of a matrix can be significantly high on diagonals if singular values are small. This is a common case in high-dimensional setting because there is not enough information to identify the space of the parameters. In other words, small, or zero singular values convert the problem to an infinite solution problem. As a result, β is estimated with significantly high variation.
It should be noted that increasing or decreasing T compared to r has very different and opposite effect on the statistical inferences. In general, multivariate methods try to make statistical inference about dependencies among variables so that increasing T has the effect of improving the accuracy of the inferred parameters, whereas increasing r has the opposite effect of reducing accuracy.

1.2 Penalized approaches to regression

Reviewing the literature reveals several methods that can cope with high-dimensional data. The majority of these methods rely on imposing an extra term, namely a penalty term, on the likelihood to convert the maximum likelihood estimation (MLE) from an infinite solution problem to a tuned-solution one. In other words, the classical multivariate methods solve the unconstrained likelihood whereas penalized MLE constrains the estimations to lie in some geometric shapes centred around the origin. For instance, Ridge regression [Hoerl and Kennard, 1970] imposes an l_2 norm penalty on the (log)likelihood to regularize estimations. That is, ridge regression solves the following minimization problem, which is equivalent to maximizing an l_2 constrained likelihood,

$$||y - X\beta||_2^2 + \lambda||\beta||_2^2, \qquad \lambda \geq 0.$$

[Frank and Friedman, 1993] extend Ridge regression by introducing $l_\alpha, \alpha \geq 1$ norm penalty in Bridge regression. Then, the underlying problem in Bridge is to find the solution to the following minimization

problem,

$$||y - X\beta||_2^2 + \lambda||\beta||_\alpha^\alpha \qquad \alpha \geq 1, \lambda \geq 0,$$

where $||\beta||_\alpha = \sqrt[\alpha]{\sum_i \beta_i^\alpha}$. Least Absolute Shrinkage and Selection Operator (LASSO) [Tibshirani, 1996] considers a special case of Bridge penalty when $\alpha = 1$,

$$||y - X\beta||_2^2 + \lambda||\beta||_1, \qquad \lambda \geq 0.$$

This penalty leads to interesting properties including automatic parameter estimation and variable selection which are the main subject of this thesis. By variable selection, we mean estimating some coefficients exactly equal to zero. Figure (1.1) provides an illustrative view of penalized LS under l_1, l_2 and l_4 norm penalties. From this figure, one can see that increasing the norm index α results in less sharp geometry on the axis, that is less probability of getting zero for estimations.

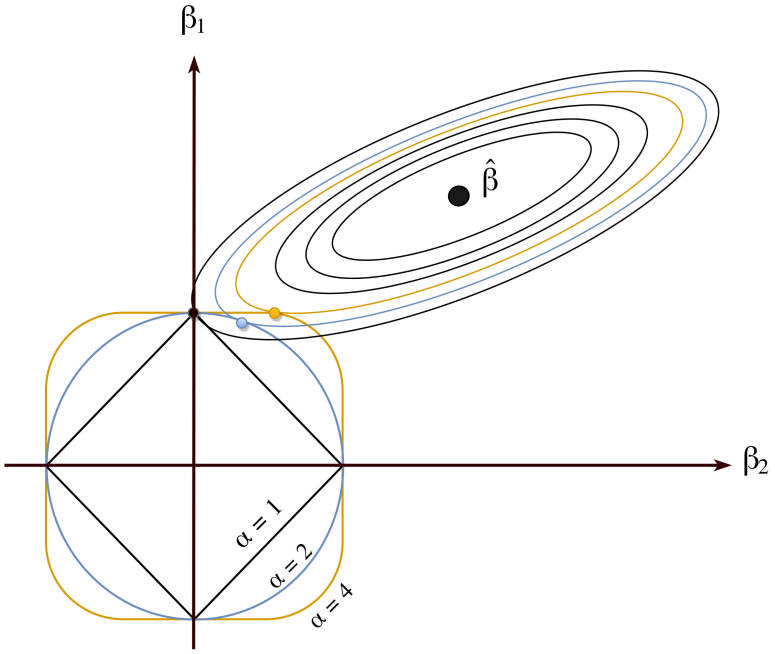

FIGURE 1.1: Contour lines of penalized LSE where $\hat{\beta}$ is LS estimation, and l_1, l_2 and l_4 correspond to lasso, ridge, and bridge penalties.

Due to the usefulness of l_1 and l_2 norm penalties, some authors take the advantage of combining norms for certain purposes. For example [Zou and Hastie, 2005] suggest a weighted function of l_1 and l_2 norm penalties under the name of Elastic-net,

$$Penalty_{\lambda,a}^{(Elastic-net)} = a\lambda||\beta||_2^2 + (1-a)\lambda||\beta||_1, \qquad a \in [0,1], \lambda \geq 0.$$

Clearly setting a to the two extremes leads to lasso and ridge penalties, respectively. This configuration is particularly useful when strongly correlated predictors tend to be in or out of the model together. Similar to elastic-net, [Hebiri, 2008] suggests Smooth-Lasso by assuming the l_2 norm to be on the difference of

consecutive parameters,

$$Penalty_{\lambda,a}^{(Smooth-Lasso)} = a\lambda||\nabla\beta||_2^2 + (1-a)\lambda||\beta||_1, \qquad a \in [0,1], \lambda \geq 0,$$

where $\nabla(.)$ is the differencing operator acting on a vector of length r so that $\nabla(\beta_j) = \beta_j - \beta_{j-1}, j = 2,\ldots,r$. This setting is particularly useful when the variations between successive coefficients of the unknown parameter of the regression are small. Similar to Smooth-Lasso, [Tibshirani et al., 2005] propose Fused-Lasso, by assuming the l_1 norm for both terms in the Smooth-Lasso penalty,

$$Penalty_{\lambda,a}^{(Fused-lasso)} = a\lambda||\nabla\beta||_1 + (1-a)\lambda||\beta||_1, \qquad a \in [0,1], \lambda \geq 0.$$

This configuration is specifically designed for problems with features that can be ordered in some meaningful way. [Fan, 2001] propose the Smoothly Clipped Absolute Deviation (SCAD) penalty, by considering a quadratic spline function with knots at λ and a,

$$Penalty_{\lambda\geq 0,a}^{SCAD} = \begin{cases} \lambda|\beta_j| & |\beta_j| \leq \lambda \\ -\frac{\beta_j^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)} & \lambda < |\beta_j| \leq a\lambda \\ \frac{a+1}{2}\lambda^2 & |\beta_j| > a\lambda \end{cases},$$

where $a = 3.7$ is recommended by [Fan, 2001]. This penalty is non-concave and is capable of producing a sparse set of solutions and approximately unbiased estimations for large coefficients. Alternatively, Dantzing selector [Candes and Tao, 2007] tries to find the solution to the following regularization problem,

$$\min_{\beta\in\mathbb{R}^r} |\beta|_1 \qquad \text{subject to} \qquad ||X'(y - X\beta)||_\infty \leq (1 + t^{-1})\sqrt{2\log r}\,\sigma, \qquad t > 0.$$

This configuration is particularly important because it results in a simple convex function that can be optimized by convenient linear programming (LP) methods.

1.3 L_1 penalized likelihood

In this section, we concentrate on lasso and its theoretical properties. In the context of linear regression, $y = X\beta + e$, lasso estimation of the parameters is the solution to minimizing l_1 penalized (log)likelihood with respect to the parameters,

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{1.2}$$

where $\lambda \geq 0$ is a tuning parameter. Setting $\lambda = 0$ leads to ordinary least squares (OLS) error regression whereas a very large λ shrinks all the coefficients toward zero and results in a null model. We show this fact by means of a simple example.

Let $y = \beta x + \epsilon$ be a linear function and $x = 1$. Then, we theoretically minimize the following constrained problem with respect to β,

$$l = (y - \beta)^2 + \lambda|\beta|.$$

Note that $\lambda = 0$ results in $(y-\beta)^2$ that is minimized at $\beta_{\min} = y$. On the other hand, for $\lambda > 0$ it results in

$$\begin{aligned} 0 = \frac{dl}{d\beta} &= -2(y-\beta) + \lambda \operatorname{sign}(\beta) \\ &= -2(y-\beta) + \lambda \operatorname{sign}(y), \end{aligned}$$

where the final term holds because of $\operatorname{sign}(y) = \operatorname{sign}(\beta)$. Consequently,

$$\beta_{\min} = \operatorname{sign}(y)\left[|y| - \frac{\lambda}{2}\right]_+, \tag{1.3}$$

where $[z]_+$ is zero for negative values of z. From the last equation, one can see that increasing the value of the penalty λ results in a zero value for $\beta_{\min}$.

Generally, lasso has two advantages over other subset selection methods, which make it a popular and widely applicable technique. Firstly, model selection and parameter estimation in lasso is simultaneous compared to other subset selection methods like forward selection, backward elimination and so on, see [Miller, 2002, Ch.3]. The second advantage is the successful application of lasso, as well as its feasibility, to the analysis of high-dimensional data [Huang, 2008]. These features make lasso a popular method that is widely studied in much of the recent literature, see e.g., [Zhou, 2009], [Knight and Fu, 2000], [Yoon, 2012], [Kyung et al., 2010a], [Morten Arendt Rasmussen, 2012], [Y. Nardi, 2011], [Zou, 2006], [Leng, 2006], [Meinshausen and Bühlmann, 2006], [Park and Casella, 2008] [Pourahmadi, 2013] and references therein. In what follows, we briefly review some theoretical properties of lasso.

1.3.1 Consistency of lasso

Beside other properties of lasso, consistency in variable selection as well as in parameter estimation, which together are called oracle property of the estimator, are two very important ones. Both of these properties are studied in details by some authors, see e.g., [Zhao, 2006, Knight and Fu, 2000, Zou, 2006, Fan and Li, 2001]. We start reviewing the consistency of lasso by defining the concept of consistency for a variable selection method.

Definition 1.1. A variable selection procedure is said to be consistent if the probability that the procedure correctly selects the set of significant variables approaches to one as the sample size increases.

In the context of linear models this is equivalent to correctly selecting the true set of non-zero coefficients. [Fan and Li, 2001, Knight and Fu, 2000, Sourav, 2013, Zhao, 2006] prove the consistency of lasso in selecting the true underlying model. In particular, [Zhao, 2006] proves that a single condition, called "Irrepresentable Condition" is almost necessary and sufficient for lasso to select the true model in the fixed r or as a function of the sample size, $r = h(n)$ for some $h(.)$. Assuming regression coefficients are divided into two parts, non-zero coefficients (β_s°) and zero coefficients ($\beta_{s^c}^\circ$), where s and s^c are two sets of indices corresponding to non-zero and zero coefficients respectively. Having written X_s and X_{s^c} as variables corresponding to s and s^c respectively, *irrepresentable* condition defines by

$$\left|\frac{1}{n}(X_{s^c}' X_s)\left(\frac{1}{n}X_s' X_s\right)^{-1} \operatorname{sign}(\beta_s^\circ)\right| \leq \mathbf{1} - \nu, \tag{1.4}$$

where **1** is a vector of 1's with the length of $(|s| - |s^c|)$ and $|.|$ denotes cardinality of a vector as well as $\nu \geq 0$. [Zhao, 2006] assumes that the inverse of the matrix in the middle of (1.4) exists. If $\nu = 0$, latter equation is called *weak irrepresentable condition*; otherwise, it is called *strong irrepresentable condition*.

Alternatively, [Sourav, 2013] studies the consistency of lasso with respect to mean squared prediction error, $\text{MSPE}(\hat{\beta}) = \mathbb{E}(\hat{Y} - Y)^2$, under minimal assumptions. The paper concludes that for the loss function considered in [Tibshirani, 1996], lasso is consistent under almost no assumptions.
On the other hand, the second challenge of lasso is consistency in parameter estimation. We should stress that an estimator that consistently selects the true underling model is not necessarily consistent in terms of parameter estimation. [Zou, 2006] studies the low consistency of lasso in terms of prediction accuracy. Then paper considers the variable selection and estimation properties of lasso and proposes a set of conditions under which lasso enjoys the oracle property in low-dimensional cases, see also [Zou, 2006]. These conditions also were discovered by [Meinshausen and Bühlmann, 2006] and ensure consistency of lasso, provided the number of variables is less than the sample size, or the number of variables increases as a function of the number of observations in high-dimensional settings.

1.3.2 Bias in lasso

Beside the advantages of lasso, the method suffers a non-removable bias that is a direct result of the bias-variance trade off. In other words, lasso adds some bias to the entire coefficient space at the price of reducing variance that by itself leads to *inconsistent* estimation of the parameters. For the example in (1.3) the bias is

$$(\beta - \beta_{\min}) = \begin{cases} \frac{\lambda}{2}\,\text{sign}\,(y) & |y| > \frac{\lambda}{2} \\ y & |y| \leq \frac{\lambda}{2} \end{cases} . \tag{1.5}$$

Then, a line of research has focused on an extension of model to decrease this bias. To this end, [Zou, 2006] proposes a weighted version of lasso and introduces a new class of estimators called *adaptive-lasso*,

$$\begin{aligned} \hat{\beta}(\lambda) &= \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda \sum_{i=1}^{r} w_i |\beta_i| \\ &= \arg\min_{\beta} \|y - X\beta\|_2^2 + \sum_{i=1}^{r} \lambda_i |\beta_i|, \end{aligned}$$

with weights different for each parameter. Adaptive lasso in particular has two advantages over ordinary lasso: (i) for small coefficients, adaptive lasso solution is also small and (ii) when coefficients are relatively large, then little shrinkage is imposed to make estimator has less bias. [Zou, 2006] also, proposes a simple but efficient computation algorithm for adaptive-lasso. [Huang, 2008] proves the oracle property of this regularization in high-dimensional setting; and [Qian, 2013] studies the effect of weights on variable selection performance of adaptive-lasso.

1.3.3 Essential theorems and proofs

In Chapter § 2 and § 3 of this work, we frequently refer to [Knight and Fu, 2000] for the asymptotic properties of l_1 regularized estimations in the linear framework. In this section we concisely review this paper to provide a theoretical foundation for the rest of the current work.

Consider again the linear model in (1.1),

$$y_i = \beta_1 x_{1i} + \ldots + \beta_r x_{ri} + e_i = x_i \beta + e_i, \tag{1.6}$$

where we assume all covariates are normalized to have zero means and unit variance, and the response to have zero mean. Then lasso estimation of the parameters is the solution to the minimization problem in (1.2). To find the limit distribution of the estimations, [Fan and Li, 2001] assume the following regularity conditions:

1. $\Sigma_T = \frac{1}{T}\sum_{i=1}^{T} x_i' x_i \to \Sigma$ where Σ is the true covariance matrix of the variables and is assumed to be non-singular and positive semi-definite
2. $\frac{1}{T}\max_{1\le i\le r} x_i x_i' \to 0$.

The following theorem determines the limit distribution of the estimations.

Theorem 1.1. Under the regularity conditions (1,2) for model in (1.6), and given $u \in \mathbb{R}^r$ and $\lambda_T/\sqrt{T} \to \lambda_\circ \ge 0$ then

$$\sqrt{T}(\hat{\beta} - \beta) \to \arg\min_u(Q),$$

where $\hat{\beta}$ is the solution to the minimization problem in (1.2) and

$$Q(u) = -2u'N(0, \sigma^2\Sigma) + u'\Sigma u + \lambda_\circ \Sigma_{j=1}^{r}\{u_j \operatorname{sign}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)\},$$

with N denoting a random variable with multivariate normal distribution.

Proof. [Fan and Li, 2001, Theorem 2] □

We should stress that Theorem (1.1) for $\lambda_\circ = 0$ results in $\arg\min_u Q = -2u'N + u'\Sigma u$ that leads to

$$u = \Sigma^{-1}N \sim N(0, \sigma^2\Sigma^{-1}),$$

and $\sqrt{T}(\hat{\beta} - \beta) \sim N(0, \sigma^2\Sigma^{-1})$, which is in line with the classical results.

Using Theorem (1.1), one can explain the asymptotic bias of estimating non-zero coefficients under the l_1 regularization as it was previously shown in equation (1.5) for a simple case. To show the bias we consider the example where all the coefficients are positive. Then,

$$\frac{\partial Q(u)}{du} = -2N + 2u'\Sigma + \lambda_\circ \mathbf{1} = 0,$$

where $\mathbf{1}$ is a vector of ones. Solving the final equation with respect to u leads to $u' = \Sigma^{-1}(N - \frac{\lambda_\circ \mathbf{1}}{2}) \sim \Sigma^{-1}N(-\frac{\lambda_\circ \mathbf{1}}{2}, \sigma^2\Sigma)$ that has non-zero mean. Thus, imposing an l_1 constrain on the non-zero coefficients results in a non-removable bias in the estimations, provided $\lambda_\circ > 0$.

Further, if some of the coefficients are zero, one can show that the limiting distribution in (1.1) puts non-zero probability at 0. To show this fact, without loss of generality we assume that the first m coefficients are non-zero and the rest are zero and define the following notations,

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \qquad N = \begin{pmatrix} N_1 \\ N_2 \end{pmatrix}, \qquad u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

where for example Σ_{11} is the $m \times m$ block matrix of Σ corresponding to non-zero coefficients. If $Q(u)$ is minimized at $u_2 = 0$ then,

$$\Sigma_{11}u_1 - N_1 = -\frac{\lambda_\circ}{2}\,\mathrm{sign}\,(\beta),$$

and

$$u_1 = \Sigma_{11}^{-1}\left(N_1 - \frac{\lambda_\circ}{2}\,\mathrm{sign}\,(\beta)\right).$$

On the other hand,

$$-\frac{\lambda_\circ}{2}\mathbf{1} \le \Sigma_{21}u_1 - N_2 \le \frac{\lambda_\circ}{2}\mathbf{1},$$

and

$$-\frac{\lambda_\circ}{2}\mathbf{1} \le \Sigma_{21}\Sigma_{11}^{-1}\left(N_1 - \frac{\lambda_\circ}{2}\,\mathrm{sign}\,(\beta)\right) - N_2 \le \frac{\lambda_\circ}{2}\mathbf{1}.$$

For the special case of $\beta = 0$ the final equation results in,

$$-\frac{\lambda_\circ}{2}\mathbf{1} \le N_2 \le \frac{\lambda_\circ}{2}\mathbf{1},$$

that puts non-zero probabilities at zero for the zero estimations.

1.4 Implementation of lasso

In contrast to standard quadratic programming, the lasso solution needs to be computed over the entire path of the tuning parameter λ. Then standard convex optimizers such as interior point [Chen et al., 1998] may not be efficient for lasso. Fortunately, one can show that the optimal solution path in lasso is piecewise linear, meaning, $\frac{\partial\beta(\lambda)}{d\lambda}$ is piecewise constant [Rosset and Zhu, 2007]. In other words, for sufficiently close values of λ, say λ_1 and λ_2, one can show that for any $\alpha \in (0,1)$, $\hat{\beta}_\lambda = \alpha\hat{\beta}_{\lambda_1} + (1-\alpha)\hat{\beta}_{\lambda_2}$, that is the regularization path between λ_1 and λ_2 is linear. Using this fact, the entire solution path of lasso can be computed in a finite number of steps. Taking this property into account, several algorithms have been proposed in the literature. For instance, pathwise coordinate optimization [Friedman et al., 2007], Grafting algorithm [Perkins et al., 2003], homotopy algorithm [Michael R. Osborne et al., 1999,

Turlach, 2005], iterative shrinkage-thresholding algorithms [Daubechies et al., 2004, Beck and Teboulle, 2009], shooting algorithm [Fu, 1998], and Least Angle Regression (LARS) [Efron et al., 2004]. Amongst these algorithms, we briefly review LARS as it is well-studied in the literature and extensively used in applications, see e.g., [Augugliaro et al., 2013, James et al., 2009, Hesterberg et al., 2008].

LARS is a multi-step algorithm that starts by finding the most correlated variable with the response. Then it moves in the direction of this predictor until another predictor has as much correlation with the current residuals. Next the algorithm moves ahead in a direction equiangular between the two predictors until a third feature vector becomes equally correlated with the residuals. This procedure continues until all variables are included in model.
The popularity of LARS is mainly due to three remarkable advantages:

- Finding the entire solution path in LARS is at the cost of solving OLS for $r < n$.
- With slight modifications, LARS can be used in stagewise regression, see [Rish and Grabarnik, 2014, p.85], [Hastie et al., 2013, p.60] and [An et al., 2008].
- Generally, the termination of the algorithm can be optimally determined by a closed form equation.

1.5 Bayesian variable selection

The regularized models described above were developed in a frequentist framework. Equivalent approaches can be developed in a Bayesian context. Here the idea is to recreate the penalised likelihood by imposing a proper prior on the parameters. The advantage of Bayesian approaches is that the whole posterior distribution of parameters would be returned, from which confidence/credible intervals can be readily obtained.

In detail, let the general form of the posterior be

$$p(\beta|y,X) \propto L(y,X|\beta) \times p(\beta|\lambda),$$

where $L(y,X|\beta)$ is the likelihood and $p(\beta|\lambda)$ is the prior on the parameters. Assuming independent Laplace densities for the prior, $p(\beta_i|\lambda) = \frac{\lambda}{2}e^{-\lambda|\beta_i|}, i = 1,2,\ldots,r$, and taking the log of the posterior leads to,

$$\log p(\beta|y,X) \propto \log\, L(y,X|\beta) + \lambda \sum_i^r |\beta_i|, \tag{1.7}$$

which is equivalent to the l_1 penalized likelihood. Similarly, a Gaussian prior acts like an l_2 norm penalty,

$$\log p(\beta|y,X) \propto \log\, L(y,X|\beta) + \frac{\lambda}{2} \sum_i^r \beta_i^2. \tag{1.8}$$

It is also convenient to assume a prior on the tuning parameter λ, called hyper-prior. Note that (1.7) and (1.8) differ from lasso and ridge in the sense that the Bayesian approach aims to estimate the entire

posterior and is not a point estimation procedure. One can connect the two techniques by estimating the parameters using a chosen statistic from the posterior distribution e.g. mode, median, mean and so on.

Having the general form of the posterior in (1.7) and (1.8), there are some techniques to make inference about the posterior from data and priors. One approach is directly sampling from the posterior, which commonly uses a Metropolis–Hastings (MH) algorithm [Hastings, 1970]. Alternatively Gibbs samplers uses the conditional posterior over the parameters. For example, [Park and Casella, 2008] suggest a Gaussian-Exponential and [Hoerl and Kennard, 1970] propose a Gaussian-InverseGamma for lasso and ridge penalties respectively. We refer to [Fahrmeir et al., 2013] for a comprehensive discussion about Bayesian regularization in generalized linear models; also [Kyung et al., 2010a] for a discussion about alternative approaches in Bayesian model selection including Spike-Slab [Ishwaran and Rao, 2005], Kuo and Mallick [Kuo and Mallick, 1998], Gibbs Variable Selection (GVS) [Dellaportas et al., 2000], Stochastic Search Variable Selection (SSVS) [George and McCulloch, 1993] and Reversible jump MCMC (RJMCMC) [Green, 1995]. Amongst these methods we focus on MH since we use it in Chapter § 4 of this thesis.

Metropolis–Hastings sampler was originally developed by [Metropolis et al., 1953] and then reformulated and extended by [Hastings, 1970]. Given a set of random variables $x = (x_1, \ldots, x_r)$ under the likelihood $L(x|\theta)$ and prior $p(\theta)$ where θ is the set of parameters, one can summarize the MH algorithm for estimating posterior $p(\theta|x) = L(x|\theta)p(\theta)$ by repeating the following four steps for certain iterations.

1. Set a proposal distribution $g(.)$ on the full set of parameters θ.
2. Draw a random sample from the proposal distribution, e.g. π_k at iteration k.
3. Evaluate the acceptance probability

$$\alpha = \min\left(1, \frac{L(x|\pi_k)p(\pi_k)g(\pi_{k-1}|\pi_k)}{L(x|\pi_{k-1})p(\pi_{k-1})g(\pi_k|\pi_{k-1})}\right).$$

4. Accept the proposal π_k with probability of α.

Repeating the four steps above leads to an estimation for the posterior, provided the proposal distribution is carefully chosen similar to the true distribution of the parameters.

1.6 Regression for counts

The previous sections have covered the case of the standard regression model, as defined in equation (1.1). In this section we briefly review regression models for a discrete response, as we will refer to these in Chapter § 4. Some examples of discrete data are the number of visits to a specialist [Machado and Santos Silva, 2005], the number of cycles a machine break down [Nagakawa and Osaki, 1975] and in criminology to count the number of offends [Parker, 2004, Osgood, 2000, Sampson and Laub, 1996, Paternoster and Brame, 1997].

Translating the regression problem from a continuous response to a discrete response results in the general family of Generalized Linear Models (GLM) [Cameron and Trivedi, 2013, Nelder and Wedderburn, 1972]

that aims to estimate the conditional distribution of a discrete response variable given some covariates. A typical example of GLM for discrete response is Poisson regression that models the conditional mean of the counts as a linear function of covariates via a logarithmic link function. That is, for a set of covariates, $x = (1, x_1, \ldots, x_r)$ and a discrete response y, then $y|x \sim Poisson(\lambda)$ under the link function,

$$\log \lambda = \beta_0 + \beta_1 x_1 + \ldots + \beta_r x_r = x\beta.$$

The Poisson model has an obvious appeal, as it is relatively simple to interpret because the right hand side of the log transformation is a linear combination of covariates and when exponentiated, the regression coefficients are interpreted as multipliers [Berk and MacDonald, 2008].

We should stress that mean and variance in Poisson distribution are the same. This is commonly referred to as the equi-dispersion property and it results in limited applicability of Poisson regression, as real data usually have different mean and variance. Negative Binomial (NB) regression relaxes the assumption of equi-dispersion and is often considered as the default choice for "over dispersed" data. Essentially over-dispersion points to the fact that there is more variation in the data than allowed by the Poisson model. In contrary, "under dispersion" is evident if there is less variation in data than captured by the Poisson model.

Although NB regression is the default choice for many applications, it scarcely applies to power-law data with long tails and highly skewed data with an excessive number of zeros. A general treatment for the zero excessive data is applying zero-inflated [Lambert, 1992] and hurdle models [Hu et al., 2011]. Moreover, NB model is not capable of handling under-dispersion in data. Therefore, alternative models such as generalised Poisson regression model [Efron, 1986], COMPoisson regression [Sellers and Shmueli, 2010] and hyper-Poisson [Sáez-Castillo and Conde-Sánchez, 2013] are developed in the literature to cope with under dispersion.

Recently [Kalktawi et al., 2016] have proposed a Discrete Weibull (DW) [Khan et al., 1989] regression model. In particular, they propose a double log transformation to link the covariates to the distribution parameters. Precisely, the conditional probability mass function of a DW random variable y given covariates x is defined over all non-negative integers by the following function,

$$f(y|q(x), \beta) = \begin{cases} q(x)^{y^\beta} - q(x)^{(y+1)^\beta} & y = 0, 1, 2, 3, \ldots \\ 0 & o.w. \end{cases},$$

where $\beta > 0$, $0 < q < 1$ and the proposed link function is $\log(-\log(q)) = x\phi$ for unknown parameter ϕ. [Kalktawi et al., 2016] show the successful application of DW regression for capturing power-law behaviour, under-dispersion, excessive zeros or high skewness in the underlying conditional distributions without the need for an additional mixture component.

1.7 Thesis outline and contribution

The outline of the thesis is as follows. In Chapter § 2, we develop a novel regularized regression model for time-dependent data. This is beyond the independent framework which is common to many of the developed regularized regression models. We propose to account for the time dependency in the

data by explicitly adding to the model autoregressive terms to the response variable together with an autoregressive process for the residuals. We derive the asymptotic properties of the estimators and assess the performance of the model on simulations and real data application.

In Chapter § 3, we develop a new penalty function. Despite all the advantages of the l_1 penalty, this penalty is not differentiable at zero, and neither are the alternatives that are proposed in the literature. The only exception is the ridge penalty, which does not lead to variable selection. Motivated by this gap, and noting the advantages that a differentiable penalty can give, such as increased computational efficiency in some cases and the derivation of more accurate model selection criteria, we develop a new penalty function based on the error function. We study the theoretical properties of this function and of the estimators obtained in a regularized regression context. Finally, we perform a simulation study and use the new penalty to analyse a diabetes and prostate cancer dataset.

In Chapter § 4, we address the novel problem of variable selection in regression for counts when the response variable follows a discrete Weibull distribution. In this chapter we introduce discrete Weibull regression under two link functions, which connect the response distribution to the covariates. We propose a Bayesian approach for estimating the parameters and for variable selection, followed by several simulations and real data illustrations.

In Chapter § 5, we summarize and draw conclusions of the work conducted. This chapter also discusses some suggestions for future work.

Chapter 2

Penalised inference for dynamic regression in the presence of autocorrelated residuals

2.1 Lasso and correlated framework

Traditional lasso approach relies on the assumption that samples are mutually independent. However, this assumption is violated when there exists a structure in the variables, such as a dependency over time. In recent years, a lot of efforts has been dedicated to lasso-like models in time dependent frameworks. For instance, [Wang et al., 2007] show the successful application of lasso in the context of linear regression with autocorrelated residuals (REGAR), given a fixed autoregressive order. They propose a model of the form

$$y_t = \sum_{i=1}^{r} x_{ti}\beta_i + \epsilon_t, \qquad \epsilon_t = \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + e_t,$$

where ϵ_ts are residuals from the regression term. [Wu and Wang, 2012] extend this model by assuming an autoregressive-moving averages (ARMA) process for the residuals and call the resulting model REGARMA. More precisely, their model is given by

$$y_t = \sum_{i=1}^{r} x_{ti}\beta_i + \epsilon_t, \qquad \epsilon_t = \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + e_t - \sum_{k=1}^{s} \theta_k e_{t-k},$$

where e_t are i.i.d Gaussian errors. [Y. Nardi, 2011] studies the lasso applications in autoregressive models when the order of AR increases with the number of data points, T. [Suo and Tibshirani, 2015] study regularized regression approaches when lags of covariates, $x_{(t-j)i}, j = 1, \ldots, k, i = 1, \ldots, r$, are involved. [Song and Bickel, 2011] study the estimation of vector AR (VAR) models. [Medeiros, 2012] studies the asymptotic property of adaptive-lasso in high-dimensional time series when the number of variables increases as a function of the number of observations and concludes that adaptive-lasso successfully selects the relevant variables in high-dimensional settings, even when the errors do not follow a Gaussian

distribution. The paper also discusses the advantages of adaptive-lasso in the situation where errors are conditionally *heteroskedastic*. We refer to [Fan et al., 2011] for a review and recent developments in high-dimensional time dependent penalized likelihood approaches.

In this chapter, we extend the idea of REGAR in [Wang et al., 2007] to include lags of response and call the resulting model DREGAR, for dynamic regression and autocorrelated residuals. In the next section we formulate the model as well as stating the necessary assumptions and notations. In Section § 2.3 we compare the proposed model with existing ones in the literature followed by introducing likelihood, l_1 and l_2 regularized likelihood in Section § 2.4, § 2.5 and § 2.6 respectively. In Section § 2.7, we focus on the special case of $DREGAR(p,0)$ and discuss the asymptotic properties of this model. An algorithm for implementing DREGAR(p,q) is proposed in Section § 2.9. A simulation study, given in Section § 2.10, will accompany the theoretical results. In Section § 2.11, we consider two applications of the model. In the first one, we consider the pollution and climate data of [Wu and Wang, 2012] and compare our results with theirs. In the second one, we consider stock market data. Finally, a discussion and pointers to future work are given in Section § 2.12.

2.2 Introduction to DREGAR

The general form of DREGAR consists of a lagged response, covariates and autocorrelated residuals. In particular, we define the model by:

$$y_t = \sum_{j=1}^{p} \phi_j y_{t-j} + x_t'\beta + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + e_t. \tag{2.1}$$

where ϵ_ts are residuals at time t, $x_t' = (x_{t1}, \ldots, x_{tr})$ is the t^{th} row of a $T \times r$ design matrix X.

Before introducing the assumptions of the model, we formally define a stationary and ergodic process as well as the backward shift operator.

Definition 2.1 (Stationary process). A process $\{w_t\}$ is strictly *stationary* if for any set of indices $\{t_1, t_2, \ldots, t_n\}$, the distribution of $(w_{t_1}, w_{t_2}, \ldots, w_{t_n})$ and $(w_{t_1+s}, w_{t_2+s}, \ldots, w_{t_n+s})$ do not depend on the time shift s. In other words,

$$f(w_{t_1}, w_{t_2}, \ldots, w_{t_n}) \stackrel{d}{=} f(w_{t_1+s}, w_{t_2+s}, \ldots, w_{t_n+s}) \qquad \forall s \in \mathbb{Z}.$$

Remark 2.1 (Weakly stationary). A process $\{w_t\}$ is weakly stationary if $\mathbb{E}(w_t) < \infty$ and $\mathbb{V}ar(w_t) < \infty$ and they do not depend on t.

Definition 2.2 (Ergodic process). A stationary process is called *ergodic* if any two variables positioned far apart in the sequence are almost independently distributed. Then, $\{w_t\}$ is ergodic if $\lim_{j\to\infty} \mathbb{C}ov(w_t, w_{t+j}) \to 0$.

Remark 2.2 (Stationarity and ergodicity for Gaussian process). A Gaussian covariance stationary process is ergodic if its covariances satisfy

$$\sum_{j=0}^{\infty} |\mathbb{C}ov(w_t, w_{t+j})| < \infty.$$

Following the literature, we define the backward shift operator L by $L(t) = t-1$ that is one-step backward acting on the time index.

We summarize the necessary assumptions for DREGAR(p,q) as follows:

(a) The response variable is assumed to be stationary and ergodic with finite second order moment. Further, we assume that the two polynomials $1 - \sum_{i=1}^{p} \phi_i L^i = 0$ and $1 - \sum_{i=1}^{q} \theta_i L^i$ have all the roots *unequal* and outside the unit circle.

(b) The covariates are assumed to be mutually independent of each other and of the error term. Following REGARMA [Wu and Wang, 2012] and REGAR [Wang et al., 2007], we assume that covariates $x_{.s}, s = 1, \ldots, r$ are generated from stationary and ergodic processes with finite second-order moment.

(c) e_ts are i.i.d Gaussian random variables with finite fourth moments.

(d) $\frac{1}{n} X'X \xrightarrow{a.s} \mathbb{E}(X'X) < \infty$ and $\max_{1 \le i \le r} x_i x_i' < \infty$.

The first three assumptions guarantee that the mean and variance of the entire system remain unchanged over time. The last assumption guarantees the existence and convergence of the sample moments.

The assumption of normality for the errors may not hold in some applications. Then taking the log [Hamilton, 1994, p. 126] or Cox-Box transformation [Box, 1964],

$$y_t^{(\Lambda)} = \begin{cases} \frac{y_t^{\Lambda} - 1}{\Lambda} & \Lambda \neq 0 \\ \log(y_t) & \Lambda = 0 \end{cases},$$

is useful. The value of $\Lambda \in \mathbb{R}$ is chosen so that it maximizes the likelihood under the assumption that $y_t^{(\Lambda)}$ is a Gaussian process. For the data that include negative values, a shift towards the positive side of the axes prior to applying the transformations is necessary.

2.2.1 Notation

In this section, we collect the necessary notations and conventions that will remain unchanged throughout this chapter.

$$
\begin{aligned}
x_t' = (x_{t1}, x_{t2}, x_{t3}, \dots x_{tr}) \quad &, \quad \text{vector of } r \text{ independent covariates } (1 \times r) \text{ at time } t \\
\beta = (\beta_1, \beta_2, \beta_3, \dots, \beta_r)' \quad &, \quad \text{vector of regression coefficients } (r \times 1) \\
\phi = (\phi_1, \phi_2, \phi_3, \dots, \phi_p)' \quad &, \quad \text{vector of dynamic coefficients } (p \times 1) \\
\theta = (\theta_1, \theta_2, \theta_3, \dots, \theta_q)' \quad &, \quad \text{vector of auroregressive coefficients } (q \times 1) \\
e \overset{iid}{\sim} N(O, \sigma^2) \quad &, \quad \text{vector of independent Gaussina errors } (T \times 1).
\end{aligned}
$$

To remove the constant from the model, we follow [Knight and Fu, 2000, Tibshirani, 1996, Huang, 2008] and normalize the covariates to zero-means and unit variance. In addition, we standardize the response, y to zero mean and divide it by a known σ_y or its consistent estimator ($\hat{\sigma}_y \xrightarrow{p} \sigma_y$) where p denotes convergence in probability. Finally, we define the full set of parameters by $\Theta = (\beta, \phi, \theta)'$.

2.3 Link with existing methods

In order to compare DREGAR with the closest methods in literature, namely REGARMA [Wu and Wang, 2012] and REGAR [Wang et al., 2007], we rewrite the three models using the backward shift operator,

$$
\begin{aligned}
DREGAR : L(\theta)L(\phi)y_t &= L(\theta)x_t'\beta + e_t, \\
REGARMA : L(\theta)y_t &= L(\theta)x_t'\beta + L(\phi)e_t, \\
REGAR : L(\theta)y_t &= L(\theta)x_t'\beta + e_t,
\end{aligned}
$$

where $L(.)$ represents a stationary polynomial of L and $\mathrm{L}(\theta)\mathrm{L}(\phi)$ represents a special case of an $AR(p+q)$ process. From these equations, one can see how REGAR and REGARMA impose the same autoregressive structure on both response and covariates, whereas DREGAR assumes different structures on each of them. We found this aspect to be particularly advantageous on a number of analyses of real datasets, which we report at the end of this chapter, where DREGAR fits the data better than the two competitive models. In contrast to REGAR and DREGAR, REGARMA contains a moving average process on the errors. The MA component, however, induces a higher level of complexity in the parameter estimation and in the proofs of the theoretical results.

Despite their differences, all three models belong to the general framework of ARMAX [Ljung, 1998, Nelles, 2013], which is common in the system identification and signal processing literature [Zhu, 2001, Nelles, 2013, Dos Santos, 2012, Keesman, 2011, Pintelon and Schoukens, 2004]. A general ARMAX model is defined by

$$
L(\theta)y_t = L(\gamma)x_t' + L(\phi)e_t,
$$

where $L(\theta)$, $L(\gamma)$ and $L(\phi)$ represent different structures on the corresponding parameters. Figure (2.1) provides a schematic view of ARMAX, DREGAR, REGAR and REGARMA where A, C, D and S are polynomials in the backward shift operator and B contains the regression coefficients. This figure shows

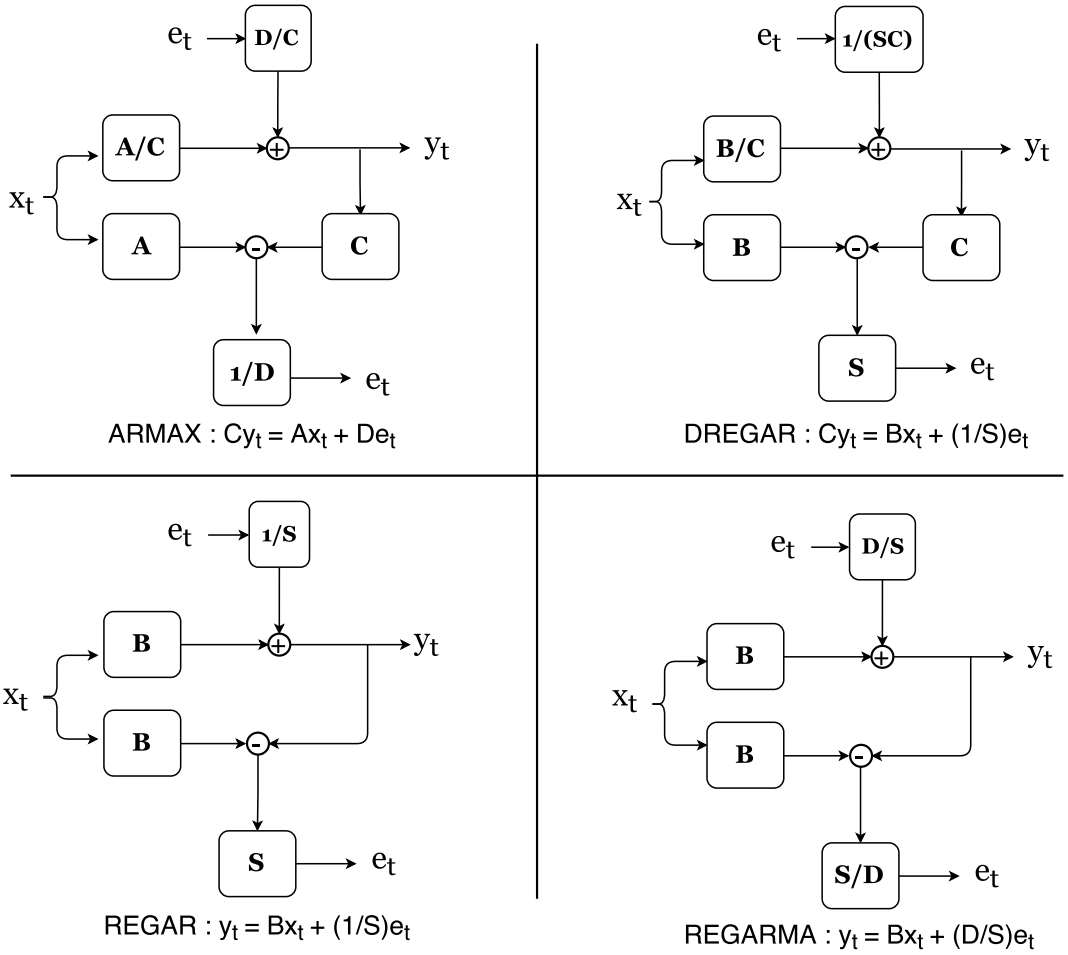

FIGURE 2.1: Schematic illustration of ARMAX, DREGAR, REGAR and REGARMA.

that DREGAR and ARMAX impose an extra filter on y_t whereas REGARMA and REGAR do not.

The focus of this chapter is on DREGAR and we consider in particular the high-dimensional case where maximum likelihood estimation fails. We therefore devise a penalised likelihood approach for parameter estimation, in the same spirit as in the REGAR and REGARMA contributions. Looking at the literature on ARMAX, we found a small number of contributions to parameter estimation in high-dimensional cases. In particular, [Chiuso and Pillonetto, 2012] and [Bańbura et al., 2010] discuss Bayesian approaches to non-parametric identification and regularization for high-dimensional dynamical networks. [Pillonetto and Chiuso, 2015], [Pillonetto et al., 2015] and [Pillonetto and Aravkin, 2014] discuss kernel-based regularization in linear system identification via stable spline kernels [Aravkin et al., 2013].

2.4 Likelihood estimation for DREGAR

We now consider parameter estimation for a DREGAR model, starting from the traditional likelihood estimation method. The conditional likelihood of the parameters given the prior information up to time

$t-1$ is given by,

$$f(y,x|\Theta,\mathcal{F}) = \prod_{t=T_0+1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}\left(y_t - \mathbb{E}(y_t|\mathcal{F}_{t-1})\right)^2}, \qquad T_\circ < T$$

where $T_\circ = p+q$, $\mathcal{F}_{t-1}$ denotes the σ-field consists of the information on x and y up to time $t-1$ and

$$\begin{aligned}\mathbb{E}(y_t|\mathcal{F}_{t-1}) &= x_t'\beta + \sum_{i=1}^{p}\phi_i y_{t-i} + \sum_{j=1}^{q}\theta_j\epsilon_{t-j},\\ \mathbb{V}ar(y_t|\mathcal{F}_{t-1}) &= \sigma^2.\end{aligned}$$

Maximizing the log-likelihood is equivalent to minimizing,

$$L_T(\Theta) = \sum_{t=T_0+1}^{T}(y_t - x_t'\beta - \sum_{i=1}^{p}\phi_i y_{t-i} - \sum_{j=1}^{q}\theta_j\epsilon_{t-j})^2, \tag{2.2}$$

where $n = T - T_\circ$ is the total number of observations in the likelihood. In matrix notation we define $H'_{(total)} = (H_{(p)}, H_{(q)}, X)$ as a $n \times (p+q+r)$ matrix including dynamic lags ($H_{(p)}$), residuals lags ($H_{(q)}$) and design matrix. Then, the general form of the model is,

$$y = H_{(total)}\Theta + e.$$

Fixing the first $(p+q)$ observations and assuming $T \geq (r+p+q)$, OLS estimation of the parameters is given by

$$\hat{\Theta} = (H_{(total)}H'_{(total)})^{-1}H_{(total)}y, \tag{2.3}$$

provided $H_{(total)}H'_{(total)}$ is positive-definite.

2.4.1 Consistency of OLS estimations

In this section we focus on the limiting distribution of the estimators in equation (2.3) and show that the OLS estimation of the parameters suffers a bias that is a direct result of autocorrelated residuals.

Precisely we show that

$$\sqrt{n}(\hat{\Theta} - \Theta) \xrightarrow{d} N(bias, \sigma^2 Q^{-1}),$$

where

$$Q = \begin{pmatrix} I_{r\times r} & \mathbb{E}(x_{s_3 t}(\frac{L^{s_4}}{A}x_t'\beta)|s_3,s_4,t) & O_{r\times q} \\ & \mathbb{E}\left((\frac{L^{s_7}}{A}x_t'\beta)(\frac{L^{s_8}}{A}x_t'\beta) + (\frac{L^{s_7}}{AB}e_t\frac{L^{s_8}}{AB}e_t)|s_7,s_8,t\right) & \mathbb{E}\left((\frac{L^{s_1}}{A}\frac{1}{B}e_t)(\frac{L^{s_2}}{B}e_t)|s_1,s_2,t\right) \\ & & \mathbb{E}\left(\frac{L^{s_9}}{B}e_t\frac{L^{s_{10}}}{B}e_t|s_9,s_{10},t\right) \end{pmatrix},$$

with $(s_1,s_4,s_7,s_8) \in \{1,2,\ldots,p\}$, $(s_2,s_9,s_{10}) \in \{1,2,\ldots,q\}$, $s_3 \in \{1,2,\ldots,r\}$, $A = (1 - \sum_{i=1}^{p} L^i\phi_i)$, $B = (1 - \sum_{i=1}^{q} L^i\theta_i)$, $t = T_\circ + 1,\ldots,T$ and O is the null matrix.

To show the limit distribution we start with

$$\begin{aligned}\hat{\Theta} &= \Theta + (H_{(total)}H'_{(total)})^{-1}H_{(total)}e \\ &= \Theta + \left((X', H_{(p)}, H_{(q)})'(X', H_{(p)}, H_{(q)})\right)^{-1}(X', H_{(p)}, H_{(q)})'e, \end{aligned} \tag{2.4}$$

where the second term in the right hand side (RHS) of (2.4) is a stochastic process constituting of $\{y, X, e\}$ and follows a certain distribution. Considering the asymptotic form of the bias $\sqrt{n}(\mathbb{E}\hat{\Theta} - \Theta)$,

$$\begin{aligned}\sqrt{n}(\mathbb{E}\hat{\Theta} - \Theta) &= \left(\frac{1}{n}(X', H_{(p)}, H_{(q)})'(X', H_{(p)}, H_{(q)})\right)^{-1} \\ &\qquad \times \frac{1}{\sqrt{n}}(X', H_{(p)}, H_{(q)})'e. \\ &= (H_1)^{-1}H_2'e, \end{aligned} \tag{2.5}$$

where $H_1 = \frac{1}{n}(X', H_{(p)}, H_{(q)})'(X', H_{(p)}, H_{(q)})$, $H_2 = \frac{1}{\sqrt{n}}(X', H_{(p)}, H_{(q)})$ and $n = T - T_\circ$. All we do in the next paragraphs is simplifying H_1 and $H_2'e$ and discussing their asymptotic distribution.

Starting with H_1. Let us rewrite H_1 as

$$H_1 = \frac{1}{n}\begin{pmatrix} XX' & XH_{(p)} & XH_{(q)} \\ H'_{(p)}X' & H'_{(p)}H_{(p)} & H'_{(p)}H_{(q)} \\ H'_{(q)}X' & H'_{(q)}H_{(p)} & H'_{(q)}H_{(q)} \end{pmatrix}.$$

Then for the first block of this matrix we get,

$$\frac{1}{n}XX' = \frac{1}{n}\sum_{t=1}^{n} x_t x_t' = \frac{1}{n}\begin{pmatrix} x_1x_1' & x_1x_2' & \dots & x_1x_r' \\ x_2x_1' & x_2x_2' & \dots & x_2x_r' \\ \vdots & \vdots & \ddots & \vdots \\ x_rx_1' & x_rx_2' & \dots & x_rx_r' \end{pmatrix} = \Sigma \to I,$$

where I is identity matrix. This convergence is guaranteed by the assumption (d). Following a similar approach, we expand other elements in H_1. To keep this section simple, we report only the result and refer to Appendix § A.0.1 for a complete proof of each block-matrix in H_1. In particular we show that

$$H_1 \to Q = \begin{pmatrix} I_{r\times r} & \mathbb{E}(x_{s_3 t}(\frac{L^{s_4}}{A}x_t'\beta)|s_3, s_4, t) & O_{r\times q} \\ & \mathbb{E}\left((\frac{L^{s_7}}{A}x_t'\beta)(\frac{L^{s_8}}{A}x_t'\beta) + (\frac{L^{s_7}}{AB}e_t\frac{L^{s_8}}{AB}e_t)|s_7, s_8, t\right) & \mathbb{E}\left((\frac{L^{s_1}}{A}\frac{1}{B}e_t)(\frac{L^{s_2}}{B}e_t)|s_1, s_2, t\right) \\ & & \mathbb{E}\left(\frac{L^{s_9}}{B}e_t\frac{L^{s_{10}}}{B}e_t|s_9, s_{10}, t\right) \end{pmatrix},$$

with $(s_1, s_4, s_7, s_8) \in \{1, 2, \dots, p\}$, $(s_2, s_9, s_{10}) \in \{1, 2, \dots, q\}$, $s_3 \in \{1, 2, \dots, r\}$, $A = (1 - \sum_{i=1}^{p} L^i\phi_i)$ and $B = (1 - \sum_{i=1}^{q} L^i\theta_i)$.

For the second term in (2.5), $H_2'e$, under assumptions [a-d], it is possible to derive an asymptotic distribution.

Theorem 2.1 (Convergence of a stationary and ergodic process). Let S_t be a stationary process with finite moments given by $\mathbb{E}(S_t) = \mu$ and $\mathbb{E}(S_t - \mu)(S_{t-j} - \mu) = \gamma_j$ for all t and absolutely summable

autocorrelations $\sum_{j=0}^{\infty} |\gamma_j| < \infty$. Then,

$$\bar{S}_n = \frac{1}{n}\sum_{t=1}^{n} S_t \xrightarrow{a.s} \mu,$$
$$\lim_{n\to\infty}\{n \times \mathbb{E}(\bar{S}_n - \mu)^2\} = \sum_{j=-\infty}^{\infty} \gamma_j.$$

Proof. [Hamilton, 1994, proposition 7.5, p 188]. □

Theorem 2.2 (Convergence of inverse matrices). The matrix inverse function is continuous at every point that represents a non-singular matrix. Then, for example, if $\frac{X'X}{n} \underset{w.r.t\, n}{\to} M$, a finite non-singular matrix, then $(\frac{X'X}{n})^{-1} \underset{w.r.t\, n}{\to} M^{-1}$.

Proof. [White, 2001, p 16]. □

Considering Theorem (2.1) and (2.2), and assuming that $x_i, i = 1, 2, \ldots, r$ and y are stationary and ergodic with finite second moments, Theorem (2.1) guarantees that asymptotic means of all elements in H_1 exist, $\mathbb{E}(H_{(total)}H'_{(total)}) \to Q$. If H_1 and Q are positive definite, Theorem (2.2) leads to the fact that Q^{-1} is non-singular and, $H_1^{-1} \to Q^{-1}$. Moreover, defining $\mathcal{F}_{t-1} = \{x_{(t-1)r}, y_{t-1} | r = 1, 2, \ldots, r\}$ as a σ-field including the information up to time $t-1$, then $H_2'e$ is a martingale difference sequence. *If* $e_t \overset{iid}{\sim} N(0, \sigma^2)$ *and* $\mathbb{E}(e_t^4) < \infty$, martingales central limit theorem results in,

$$\mathbb{V}ar(H_2'e) = \mathbb{E}(H_2 e' e H_2') = \sigma^2 Q$$
$$H_2'e \xrightarrow{d} N\Big(bias, \sigma^2 Q\Big),$$

and

$$\sqrt{n}(\hat{\Theta} - \Theta) \xrightarrow{d} N(bias, \sigma^2 Q^{-1}). \tag{2.6}$$

The bias in equation (2.6) is a direct result of estimating ϵs from a primary step precisely from $\hat{\epsilon} = y - H_{(p)}\hat{\theta} - X'\hat{\beta}$ using OLS. We should stress that ϵ and θ in DREGAR are both *unknown*. As a result, an extra step is needed for estimating the ϵ. In Appendix § A.0.1 we show that applying OLS to estimating the parameters in $y = H_{(p)}\theta - X'\beta + \epsilon$ leads to a bias in the estimations. In the next section, we show the bias theoretically and practically in DREGAR(1,1) model.

2.4.1.1 Case study : DREGAR(1,1)

In this section we make use of an illustrative example to show the bias in OLS estimation of DREGAR parameters previously shown in equation (2.6). Moreover, we show that the model may have some identifiability issues under some circumstances. To this end, we simulate 500 observations from the following DREGAR(1,1) model,

$$\begin{aligned} y_t &= x_t - 0.61 y_{t-1} + \epsilon_t \\ \epsilon_t &= 0.36\epsilon_{t-1} + e_t, \end{aligned} \tag{2.7}$$

where x_t is a single covariate generated from an underlying Gaussian $AR(1)$ process, with $\phi = 0.5$ to ensure stationarity. The parameters are estimated using OLS for an overall of 5000 repetitions. The number of repetitions (5000) is intentionally chosen so that the variation in estimations can be clearly observed from the histogram. The left panel of Figure (2.2) shows the OLS estimations and the bias, whereas the right panel shows the corresponding histogram. From both plots a bias of $(\phi - \hat{\phi}) = 0.22$ is observed.

Due to the simplicity of DREGAR(1,1) model, we study this case in more details.

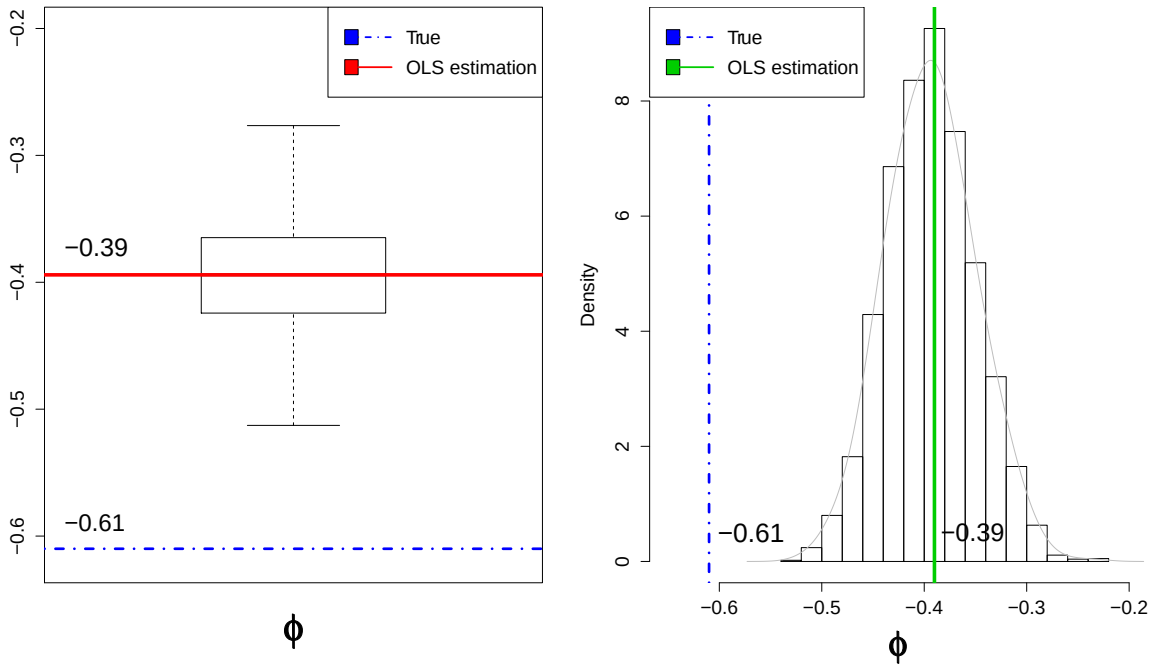

FIGURE 2.2: Simulation result for DREGAR(1,1) with one explanatory variable. (Left) OLS estimation of ϕ where dotted line denotes the true value of the parameter and solid line shows the median of the estimations. (Right) Corresponding histogram where solid vertical line represents the mode of the distribution.

Let the true underling model be

$$y_t = x_t'\beta + \phi y_{t-1} + \epsilon_t$$
$$\epsilon_t = \theta\epsilon_{t-1} + e_t,$$

where ϕ and θ are time-dependent parameters and β is a single static parameter. Ordinary least squares estimation of ϕ results in,

$$\mathbb{E}(\hat{\phi}) = \frac{\sum_t y_t y_{t-1}}{\sum_t y_{t-1}^2} \overset{T\to\infty}{=} \phi + \frac{\mathbb{C}ov(y_{t-1}, \epsilon_t) + \mathbb{C}ov(y_{t-1}, x_t'\beta)}{\mathbb{V}ar(y_{t-1})},$$

where the denominator equals to 1 as the response is standardized prior to analysis. On the other hand, $\mathbb{C}ov(y_{t-1}, \epsilon_t)$ and $\mathbb{C}ov(y_{t-1}, x_t'\beta)$ are not necessarily zero, as

$$\mathbb{C}ov(y_{t-1}, x_t'\beta) \propto \beta^2 \sum_{k=1}^{\infty} \gamma_k, \tag{2.8}$$

where γ_k denotes the k^{th} autocorrelations between x_t and x_{t-k} that is not necessarily zero. Moreover,

$$\begin{aligned}\mathbb{C}ov(y_{t-1}, \epsilon_t) &= \mathbb{E}(y_{t-1}\epsilon_t) \\ &= \phi\theta\mathbb{E}(\epsilon_{t-1}y_{t-2}) + \theta\mathbb{E}(\epsilon_{t-1}^2).\end{aligned}$$

But, $\mathbb{E}(\epsilon_{t-1}^2) = \mathbb{E}(\epsilon_t^2)$ and $\mathbb{E}(\epsilon_{t-1}y_{t-2}) = \mathbb{E}(\epsilon_t y_{t-1})$. Then,

$$\mathbb{C}ov(y_{t-1}, \epsilon_t) = \frac{\theta\mathbb{V}ar(\epsilon_t)}{1-\theta\phi}. \tag{2.9}$$

The last equality shows that the correlation between y_{t-1} and ϵ_t is zero iff $\theta = 0$, that is when there is no serial correlation among residuals.

Consequently, there is always a non-zero bias in the OLS estimation of the time dependent parameter ϕ, provided the residuals are autocorrelated. In line with the theoretical results, estimating the bias using equation (2.8) and (2.9) for the model in (2.7) results in $(\phi - \hat{\phi}) = 0.23$ that is comparable to the empirical result.

Moreover, to show the correlation between estimators, in particular between $\hat{\theta}$ and $\hat{\phi}$ we assume that there is no covariate in a DREGAR(1,1) model. Thus, the model reduces to,

$$\begin{aligned}y_t &= \phi y_{t-1} + \epsilon_t \\ \epsilon_t &= \theta\epsilon_{t-1} + e_t.\end{aligned}$$

Given T is sufficiently large, (2.9) results in

$$\mathbb{E}(\hat{\phi}) = \phi + \frac{\theta\sigma_\epsilon^2}{1-\theta\phi},$$

and estimating ϵ_t by $\hat{\epsilon} = y_t - \hat{\phi}y_{t-1}$ leads to,

$$\hat{\epsilon}_t \overset{T\to\infty}{=} \epsilon_t + \frac{\theta\sigma_{\epsilon_t}^2}{1-\theta\phi}y_{t-1}.$$

Defining $H = (y_{t-1}, \hat{\epsilon}_{t-1})$ as the matrix of the first lags and using LS methods for estimating the parameters $\Theta = (\phi, \theta)$ leads to

$$\hat{\Theta} = (H'H)^{-1}H'y,$$

so that,

$$\sqrt{n}\mathbb{V}ar(\hat{\Theta}) \overset{n\to\infty}{\to} \left(\frac{H'H}{n}\right)^{-1}\sigma^2,$$

where $n = T - (p+q) = T-2$. In what follows, we find the asymptotic covariance matrix of the estimators,

$$\left(\frac{H'H}{n}\right)^{-1} = \left[\frac{1}{n}\begin{pmatrix} \sum_t y_{t-1}^2 & \sum_t y_{t-1}\hat{\epsilon}_{t-1} \\ \sum_t y_{t-1}\hat{\epsilon}_{t-1} & \sum \hat{\epsilon}_{t-1}^2 \end{pmatrix}\right]^{-1},$$

where $n = T - T_\circ$ and the summations define over $T_\circ = p+q = 2$ and T.

For n sufficiently large we have,

$$\frac{H'H}{n} \overset{n\to\infty}{\to} \begin{pmatrix} \sigma_y^2 & \frac{1}{n}\sum_t y_{t-1}\hat{\epsilon}_{t-1} \\ \frac{1}{n}\sum_t y_{t-1}\hat{\epsilon}_{t-1} & \sigma_{\hat{\epsilon}}^2 \end{pmatrix},$$

where the first element of the matrix $\sigma_y^2 = 1$ by the assumption. Further,

$$\begin{aligned} \frac{1}{n}\sum_t y_{t-1}\hat{\epsilon}_{t-1} &= \frac{1}{n}\sum_t y_{t-1}\left(\epsilon_{t-1} + \frac{\theta\sigma_\epsilon^2}{1-\theta\phi}y_{t-2}\right) \\ &= \frac{1}{n}\sum_t y_{t-1}\epsilon_{t-1} + \frac{\theta\sigma_\epsilon^2}{(1-\theta\phi)}\frac{1}{n}\sum_t y_{t-1}y_{t-2}. \end{aligned} \tag{2.10}$$

From equation (A.4) in the appendix, the first summation on the RHS of (2.10) tends to $\frac{\sigma_\epsilon^2}{1-\phi\theta}$ and from (2.9) the second one tends to $\phi + \frac{\theta\sigma_\epsilon^2}{1-\theta\phi}$. Consequently,

$$\begin{aligned} \frac{1}{n}\sum_t y_{t-1}\hat{\epsilon}_{t-1} &\overset{n\to\infty}{=} \frac{\sigma_\epsilon^2}{1-\theta\phi} + \frac{\theta\sigma_\epsilon^2}{1-\theta\phi}\left(\phi + \frac{\theta\sigma_\epsilon^2}{1-\theta\phi}\right) \\ &= \frac{\sigma_\epsilon^2}{1-\theta\phi}\left(1 + \theta(\phi + \frac{\theta\sigma_\epsilon^2}{1-\theta\phi})\right), \end{aligned}$$

and,

$$\left(\frac{H'H}{n}\right)^{-1} \overset{n\to\infty}{=} \frac{1}{\sigma_{\hat{\epsilon}}^2 - \left[\frac{\sigma_\epsilon^2}{1-\theta\phi}\left(1+\theta(\phi+\frac{\theta\sigma_\epsilon^2}{1-\theta\phi})\right)\right]^2} \begin{pmatrix} \sigma_{\hat{\epsilon}}^2 & -\frac{\sigma_\epsilon^2}{1-\theta\phi}\left(1+\theta(\phi+\frac{\theta\sigma_\epsilon^2}{1-\theta\phi})\right) \\ -\frac{\sigma_\epsilon^2}{1-\theta\phi}\left(1+\theta(\phi+\frac{\theta\sigma_\epsilon^2}{1-\theta\phi})\right) & 1 \end{pmatrix}.$$

From the last equality, the correlation between the estimations can be unstable, given $\theta\phi \to 1$ that is both parameters are close enough to one on the same sign. Moreover, the determinant of the matrix tends to zero, given the parameters are the roots of the second order polynomial in the denominator. That is, there is the identifiability problem on some combination of the parameters. For a simple case where $\sigma_\epsilon^2 = 1$ it is $\theta = 2\phi/(2\phi^2 - 1)$.

As it is pointed out, the source of the bias is the initial OLS that is used for estimating ϵ, see equation (2.9). In other words, removing the autoregressive process over ϵ results in unbiased estimations for the parameter ϕ. This motivates us to consider the case of DREGAR(p,0) in the theoretical sections. However, we show that using iterative OLS results in improving estimations and decreasing the bias in application. In the next two sections we introduce the l_1 and l_2 regularized likelihoods of DREGAR followed by the theoretical properties of DREGAR(p,0) in Section § 2.7.

2.5 L_1 penalized likelihood for DREGAR

The estimation approach described in the previous section does not work when $T < (r+p+q)$. In addition, it does not perform variable selection, i.e. the estimates of the regression coefficients are not necessarily zero. In the spirit of lasso methods, we impose an l_1 penalty on the regression coefficients. Thus, we propose minimizing

$$\begin{aligned} Q_n(\Theta) = &\sum_{t=T_0+1}^{T} \left(y_t - x_t'\beta - \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \right)^2 \\ &+ \sum_{i=1}^{r} \lambda_n |\beta_i| + \sum_{j=1}^{p} \gamma_n |\phi_j| + \sum_{k=1}^{q} \tau_n |\theta_k|, \end{aligned} \tag{2.11}$$

where $n = T - T_\circ$ and $\lambda_n, \gamma_n, \tau_n$ are tuning parameters. Moreover, considering the superiority of adaptive penalties [Zou, 2006], we propose an adaptive form of the likelihood as

$$\begin{aligned} Q_n^*(\Theta) = &\sum_{t=T_0+1}^{T} \left(y_t - x_t'\beta - \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \right)^2 \\ &+ n \sum_{i=1}^{r} \lambda_i^* |\beta_i| + n \sum_{j=1}^{p} \gamma_j^* |\phi_j| + n \sum_{k=1}^{q} \tau_k^* |\theta_k|, \end{aligned}$$

where $\lambda_i^*, \gamma_j^*, \tau_k^*, i = 1, \ldots, r; j = 1, \ldots, p; k = 1, \ldots, q$ are tuning parameters.

In matrix form, we have

$$Q_n(\Theta) = L_n(\Theta) + n\lambda'|\beta| + n\gamma'|\phi| + n\tau'|\theta|,$$

where

$$\begin{aligned} &L_n(\Theta) = |y - H_{(total)}\Theta|_2^2, \\ &\lambda' = \{\lambda\}_{1\times r}, \qquad \gamma' = \{\gamma\}_{1\times p}, \qquad \tau' = \{\tau\}_{1\times q} \\ &\beta = (\beta_1, \beta_2, \beta_3, \ldots, \beta_r)', \qquad \phi = (\phi_1, \phi_2, \phi_3, \ldots, \phi_p)', \qquad \theta = (\theta_1, \theta_2, \theta_3, \ldots, \theta_q)'. \end{aligned}$$

Similarly, the adaptive form of the regularized likelihood results in

$$Q_n^*(\Theta) = L_n(\Theta) + n\lambda'^*|\beta| + n\gamma'^*|\phi| + n\tau'^*|\theta|,$$

where

$$\begin{aligned} \lambda^{*'} &= (\lambda_i^*, i = 1, \ldots, r)_{1\times r} &&= \{\lambda_1^*, \lambda_2^*, \ldots, \lambda_r^*\}, \\ \gamma^{*'} &= (\gamma_j^*, j = 1, \ldots, p)_{1\times p} &&= \{\gamma_1^*, \gamma_2^*, \ldots, \gamma_p^*\}, \\ \tau^{*'} &= (\tau_k^*, k = 1, \ldots, q)_{1\times q} &&= \{\tau_1^*, \tau_2^*, \ldots, \tau_q^*\}, \end{aligned}$$

and $\lambda^*, \gamma^*, \tau^*$ are tuning parameters.

2.6 L_2-penalized solution to DREGAR

The l_1 penalty discussed in the previous section has advantages and disadvantages. Although there is no analytic solution to the DREGAR optimization problem in this case, the l_1 penalty results in a sparse

solution, thus it naturally leads to variable selection. However, the penalty indirectly penalises more the coefficients with lower values. This fact may be important in some applications. For example, if we take an AR(1) model, $y_t = \phi y_{t-1} + e_t$, the stationarity condition requires that the root of the polynomial, $1 - L\phi$, lies outside the unit circle, that is $|\phi| < 1$ i.e. the coefficient lies in the $(-1,1)$ interval. It can be shown that for any stationary AR(p) process, all the coefficients must be in the $(-1,1)$ interval. This limitation of the lasso approach can be addressed by considering an l_2 penalty instead. In this case, there is less penalty on low coefficients, to the expense of a non-sparse solution. In this section, we consider the l_2 penalty for the DREGAR model, an approach that goes under the name of ridge regression in the case of linear regression.

The solution to minimizing the l_2 regularized (log)likelihood

$$Q_{n,l_2}(\Theta) = L_n(\Theta) + n\lambda'^{*}|\beta|_2^2 + n\gamma'^{*}|\phi|_2^2 + n\tau'^{*}|\theta|_2^2$$

is given by ,

$$\hat{\Theta} = (H_{(total)}H'_{(total)} + n\Lambda I)^{-1}H_{(total)}y,$$

where $0 < \Lambda = (\lambda_i, \gamma_j, \tau_k)'; i = 1,2,3,\ldots,r,\ j = 1,2,3,\ldots,p\ ,k = 1,2,3,\ldots,q$ and $I = I_{(r+p+q)\times(r+p+q)}$. Obviously $\hat{\Theta}$ is biased due to the existence of non-vanishing Λ as well as the estimation procedure.

As variable selection is desirable in many contexts, the chapter focuses mainly on the l_1-penalised method. In the next section, we consider the theoretical properties of the estimators that are derived from this approach.

2.7 Theoretical properties of l_1 penalized DREGAR(p,0)

In this section we focus on theoretical properties of l_1 penalized DREGAR(p,0) including asymptotic properties of the estimators. As it is evident from equation (2.6), the general form of a DREGAR(p,q) model suffers from being biased that is due to using OLS for the estimation of the autoregressive coefficients when the disturbances are autocorrelated. Thus, we concentrate on the theoretical properties of DREGAR(p,0) as there is asymptotically no bias in this model. This model differs from REGAR(p) [Wang et al., 2007] as it imposes an autoregressive process on the response whereas REGAR(p) considers the case of autocorrelated residuals (i.e. DREGAR(0,q)). In the upcoming subsection, we collect the necessary notations and in the next subsection we discuss the asymptotic properties of DREGAR(p,0) and adaptive-DREGAR(p,0).

2.7.1 Notations and Definitions

To study the theoretical properties of both, DREGAR(p,0) and adaptive-DREGAR(p,0), we make the following assumptions:

1. There is a correct model with coefficients $\Theta^\circ = (\beta^\circ, \phi^\circ)'$,

$$\Theta^\circ = (\beta_1^\circ, \beta_2^\circ, \beta_3^\circ, \ldots, \beta_r^\circ, \phi_1^\circ, \phi_2^\circ, \phi_3^\circ, \ldots, \phi_p^\circ)'_{1\times(r+p)}.$$

2. There are $r_\circ < r$, and $p_\circ < p$ non-zero parameters.

3. We define

   i. $s_1 = \{i \in \mathbb{N}, 1 \leq i \leq r; \quad \beta_i^\circ \neq 0\}$ Indices for non-zero REG coefficients.

   ii. $s_2 = \{j \in \mathbb{N}, 1 \leq j \leq p; \quad \phi_j^\circ \neq 0\}$ Indices for non-zero DA coefficients.

s_1^c and s_2^c are complementary sets and containing zero indices. We also define $\beta_{s_1}^\circ, \phi_{s_2}^\circ$ and their corresponding (DREGAR(p,0)) estimations by $\hat{\beta}_{s_1}, \hat{\phi}_{s_2}$. Similarly, adaptive-DREGAR(p,0) estimations by $\hat{\beta}_{s_1}^*, \hat{\phi}_{s_2}^*$. Finally, different combinations of parameters when needed e.g. $\beta_{s_1^c}^\circ, \beta_{s_1^c}, \hat{\beta}_{s_1^c}, \beta_{s_1^c}^*, \hat{\beta}_{s_1^c}^*$. Further we define,

I. $\Theta_1^\circ = \{\beta_{s_1}^{\circ\prime}, \phi_{s_2}^{\circ\prime}\}$, $\Theta_2^\circ = \{\beta_{s_1^c}^{\circ\prime}, \phi_{s_2^c}^{\circ\prime}\}$ True non-zero (significant) and zero (insignificant) parameters respectively.

II. $\hat{\Theta}_1 = \{\hat{\beta}_{s_1}', \hat{\phi}_{s_2}'\}$, $\hat{\Theta}_2 = \{\hat{\beta}_{s_1^c}', \hat{\phi}_{s_2^c}'\}$ DREGAR(p,0) non-zero and zero parameters respectively.

III. $\hat{\Theta}_1^* = \{\hat{\beta}_{s_1}^{*\prime}, \hat{\phi}_{s_2}^{*\prime}\}$, $\hat{\Theta}_2^* = \{\hat{\beta}_{s_1^c}^{*\prime}, \hat{\phi}_{s_2^c}^{*\prime}\}$ Adaptive-DREGAR(0,p) non-zero and zero parameters respectively.

2.7.2 Asymptotic properties of DREGAR(p,0)

Theorem 2.3 (Limit distribution of estimations). Assuming $\lambda_n\sqrt{n} \to \lambda_\circ$, $\gamma_n\sqrt{n} \to \gamma_\circ$, and $\lambda_\circ, \gamma_\circ \geq 0$. Then under assumptions [a-d], it follows that $\sqrt{n}(\hat{\Theta} - \Theta^\circ) \xrightarrow{d} \arg\min k(\delta)$ where

$$\begin{aligned} k(\delta) = & -2\delta' W + \delta' U_B \delta \\ & + \lambda_\circ \sum_{i=1}^{r} \{(u_i \operatorname{sign}(\beta_i^\circ) I(\beta_i^\circ \neq 0)) + |u_i| I(\beta_i^\circ = 0)\} \\ & + \gamma_\circ \sum_{j=1}^{p} \{(v_j \operatorname{sign}(\phi_j^\circ) I(\phi_j^\circ \neq 0)) + |v_i| I(\phi_j^\circ = 0)\}, \end{aligned}$$

and $\delta = (u, v)$ is a vector of parameters in $\mathbb{R}^{(r+p)}$, $W \sim \text{MVN}(O, \sigma^2 U_B)$ and $U_B = \mathbb{C}ov(X, H_{(p)})$.

Proof. Assuming $\lambda_n\sqrt{n} \to \lambda_\circ$, $\gamma_n\sqrt{n} \to \gamma_\circ$, and $\delta = (u', v')$. Define

$$k_n(\delta) = Q_n(\Theta^\circ + n^{-(1/2)}\delta) - Q_n(\Theta^\circ). \tag{2.12}$$

We should stress that k_n reaches the minimum at $\sqrt{n}(\hat{\Theta} - \Theta^\circ)$. Using (2.2) and (2.11),

$$\begin{aligned} k_n(\delta) = & \\ & \left(L_n(\Theta^\circ + \frac{\delta}{\sqrt{n}}) - L_n(\Theta^\circ) \right) \end{aligned} \tag{2.13a}$$

$$+ (n\lambda_n'|\beta^\circ + \frac{u}{\sqrt{n}}| - n\lambda_n'|\beta^\circ|) \tag{2.13b}$$

$$+ (n\gamma_n'|\phi^\circ + \frac{v}{\sqrt{n}}| - n\gamma_n'|\phi^\circ|), \tag{2.13c}$$

where the last two terms are

$$\begin{aligned}(n\lambda_n'|\beta^\circ + \frac{u}{\sqrt{n}}| - n\lambda_n'|\beta^\circ|) &= \left(\sqrt{n}u\lambda_n' \frac{|\beta^\circ + u/\sqrt{n}| - |\beta^\circ|}{u/\sqrt{n}}\right) \\ &\underset{n\to\infty}{=} \lambda_\circ \sum_{i=1}^{r}\{(u_i \operatorname{sign}(\beta_i^\circ) I(\beta_i^\circ \neq 0)) + |u_i| I(\beta_i^\circ = 0)\}. \\ (2.13c) &\underset{n\to\infty}{=} \gamma_\circ \sum_{j=1}^{p}\{(v_j \operatorname{sign}(\phi_j^\circ) I(\phi_j^\circ \neq 0)) + |v_j| I(\phi_j^\circ = 0)\}.\end{aligned}$$

(2.13*a*) is equal to:

$$\begin{aligned}(2.13a) = &- e'e + \\ &\left((y - H_{(p)}\phi^\circ - X'\beta^\circ) - (X', H_{(p)})\frac{\delta}{\sqrt{n}})\right)' \times \\ &\left((y - H_{(p)}\phi^\circ - X'\beta^\circ) - (X', H_{(p)})\frac{\delta}{\sqrt{n}})\right).\end{aligned}$$

Setting $A = (X', H_{(p)})$ and $e = y - H_{(p)}\phi^\circ - X'\beta^\circ$,

$$\begin{aligned}&Q_n(\Theta^\circ + \frac{\delta}{\sqrt{n}}) - Q_n(\Theta^\circ) \\ &\quad = (e' - \frac{\delta'}{\sqrt{n}}A')(e - A\frac{\delta}{\sqrt{n}}) - e'e + (2.13b) + (2.13c),\end{aligned}$$

which is equivalent to

$$(\frac{\delta' A'}{\sqrt{n}})(\frac{A\delta}{\sqrt{n}}) - (\frac{\delta' A'}{\sqrt{n}})e - e'(\frac{A\delta}{\sqrt{n}}) + (2.13b) + (2.13c). \tag{2.14}$$

From left to right, we prove that the first term in (2.14) is bounded and the next two terms follow (asymptotically) normal distributions:

$$(\frac{\delta' A'}{\sqrt{n}})(\frac{A\delta}{\sqrt{n}}) = O_p(1) \tag{2.15}$$

$$e'(\frac{A\delta}{\sqrt{n}}) = (\frac{\delta' A'}{\sqrt{n}})e = S_n, \tag{2.16}$$

where S_n is a random variable that follows a normal distribution. Similar calculations to Section § 2.4.1 show that (2.15) tends to $\delta' U_B \delta$ where U_B is the covariance matrix of $(X', H_{(p)})$ which is bounded, $O_p(1)$.

Recalling S_n from (2.16) as a function of n,

$$S_n = (\frac{\delta' A'}{\sqrt{n}})e = \frac{1}{\sqrt{n}}(u'X + v'H_{(p)}')e,$$

and using assumptions [a-d] and central limit theorem for martingales result in

$$S_n \overset{a.s}{\to} \delta' W,$$

where $\delta = (u', v')$ and $W \sim \text{MVN}(O, \sigma^2 U_B)$. Then

$$-(2.16) \overset{n\to\infty}{\to} -2\delta' W.$$

Substituting all results in equation (2.12),

$$k_n(\delta) \overset{n\to\infty}{\to} -2\delta' N(O, \sigma^2 U_B) + \delta' U_B \delta + \lambda_\circ \sum_{i=1}^{r} \{(u_i \,\text{sign}\,(\beta_i^\circ) I(\beta_i^\circ \neq 0)) + |u_i| I(\beta_i^\circ = 0)\}.$$
$$+ \gamma_\circ \sum_{j=1}^{p} \{(v_j \,\text{sign}\,(\phi_j^\circ) I(\phi_j^\circ \neq 0)) + |v_i| I(\phi_j^\circ = 0)\}.$$

Note that U_B is similar to Q in Appendix § A by removing the corresponding terms to $H_{(q)}$. Up to now, we have proved $k_n(\delta) \overset{n\to\infty}{\to} k(\delta)$. To show that $\arg\min k_n(\delta) = \sqrt{n}(\hat{\Theta} - \Theta^\circ) \overset{d}{\to} \arg\min k(\delta)$ is enough to prove that $\arg\min\{k_n(\delta)\} = O_p(1)$ [Kim and Pollard, 1990, Knight and Fu, 2000]. In order to do this, note that

$$\begin{aligned}
k_n(\delta) &= (\frac{\delta' A'}{\sqrt{n}})(\frac{A\delta}{\sqrt{n}}) - (\frac{\delta' A'}{\sqrt{n}})e - e'(\frac{A\delta}{\sqrt{n}}) + \\
&(n\lambda_n'|\beta^\circ + \frac{u}{\sqrt{n}}| - n\lambda_n'|\beta^\circ|) + (n\gamma_n'|\phi^\circ + \frac{v}{\sqrt{n}}| - n\gamma_n'|\phi^\circ|) \\
&\geq (\frac{\delta' A'}{\sqrt{n}})(\frac{A\delta}{\sqrt{n}}) - (\frac{\delta' A'}{\sqrt{n}})e - e'(\frac{A\delta}{\sqrt{n}}) - (n\lambda_n'|un^{-1/2}| - (n\gamma_n'|vn^{-1/2}|) \\
&\geq (\frac{\delta' A'}{\sqrt{n}})(\frac{A\delta}{\sqrt{n}}) - (\frac{\delta' A'}{\sqrt{n}})e - e'(\frac{A\delta}{\sqrt{n}}) - (\lambda_\circ' + \epsilon_\circ)|u| - (\gamma_\circ' + \epsilon_\circ)|v| \\
&= k_n^*(\delta),
\end{aligned}$$

where $\epsilon_\circ > 0$ is a vector of positive constants. The fourth term in $k_n^*(\delta)$ for example, comes from the fact that $\forall \epsilon_\circ > 0, \exists N, \text{if } n \geq N, |\lambda^\circ - \sqrt{n}\lambda_n| < \epsilon_\circ$. Then, $\sqrt{n}\lambda_n < \lambda^\circ + \epsilon_\circ$. In addition, $k_n(0) = k_n^*(0)$ and $f_n(\delta) = o_p(1)$. As a result $\arg\min\{k_n^*(\delta)\} = O_p(1)$ and $\arg\min\{k_n(\delta)\} = O_p(1)$.
The proof of the theorem is completed.

□

This theorem shows that the DREGAR estimator has the [Knight and Fu, 2000] asymptotic property and it implies that the tuning parameters in $Q_n(\Theta)$ do not shrink to zero at the speed faster than $n^{-1/2}$. In the proof of theorem (2.3), the errors must be independent and identically distributed and we do not make a specific assumption on the type of distribution. In other words, the central limit theorem for martingale guarantees the convergence to the normal distribution. This implies that by increasing the number of data points ($n \to \infty$), the errors can be weakly normally distributed.

As we showed in Chapter § 1 (Section § 1.3), lasso approaches to linear regression return biased estimates of the non-zero parameters [Knight and Fu, 2000]. In the following remark, we show this also in the context of the DREGAR model.

Remark 2.3 (Asymptotic bias in estimations). We consider a special case, where $\beta_i^\circ > 0, \quad 1 \leq (\forall i \in \mathbb{N}) \leq r$ and $\phi_{i_2}^\circ = 0$ for $1 \leq j_1 \leq q, 1 \leq j_2 \leq p, j_1, j_2 \in \mathbb{N}$ and we assume that there are enough observations and that the minimizer $k(\delta)$ correctly identifies coefficients. That is, $u \neq 0$ and $v = 0$. Then, $k(\delta)$ must satisfy

$$\begin{aligned}
\frac{\partial k(\delta)}{\partial u} &= \frac{\partial k(u,0)}{\partial u} \\
&= \frac{\partial}{\partial u}\Big(-2(u',0)W + (u',0)'U_B(u',0) + (2.13b) + (2.13c)\Big) \\
&= -2W_{1:r} + 2u'U_{B_{1:r}} + \lambda_\circ 1_{r\times 1} = 0 \\
&\to u' = \frac{1}{2}(2W_{1:r} - \lambda_\circ 1_{r\times 1})U_{B_{1:r}}^{-1} \\
&\text{Using Theorem (2.3): } \sqrt{n}(\hat{\beta} - \beta^\circ) \xrightarrow{d} \arg\min k(\delta = u') = \text{MVN}\left(\mathbb{E}(u') \neq 0,\, U_{B_{1:r}}^{-1}\right),
\end{aligned}$$

where $U_{B_{1:r}}$ is the first r rows of U_B corresponded to r covariates. From the final equation, DREGAR(p,0) suffers an asymptomatic bias, provided the tuning parameter is positive. In other words, lasso regularization of DREGAR(p,0) is not asymptotically consistent. In the next section we discuss the adaptive-DREGAR(p,0) where a fixed level penalty term is replaced by a weighted (adaptive) one. We show that under certain conditions adaptive-DREGAR(p,0) is consistent and enjoys the oracle property.

2.7.2.1 Adaptive DREGAR(p,0) model

Recall from Section § 2.5 that parameter estimation in adaptive-DREGAR(p,q) involves the minimization of

$$Q_n^*(\Theta) = \sum_{t=T_0+1}^{T}\left((y_t - x_t'\beta) - \sum_{i=1}^{p}\phi_i y_{t-i} - \sum_{j=1}^{q}\theta_j\epsilon_{t-j}\right)^2 + n\sum_{i=1}^{r}\lambda_i^*|\beta_i| + n\sum_{j=1}^{p}\gamma_j^*|\phi_j| + n\sum_{k=1}^{q}\tau_k^*|\theta_k|,$$

where $\lambda_i^*, \gamma_j^*, \tau_k^*$ are tuning parameters and $\Theta = (\beta, \phi, \theta)'$ is parameter space.

To prove the asymptotic property of adaptive-DREGAR(p,0) we follow [Wang et al., 2007, Knight and Fu, 2000] and define,

$$\begin{aligned}
a_n &= \max(\lambda_{i_1}^*, \gamma_{i_2}^*; \quad i_1 \in s_1, i_2 \in s_2) \\
b_n &= \min(\lambda_{i_1^c}^*, \gamma_{i_2^c}^*; \quad i_1^c \in s_1^c, i_2^c \in s_2^c),
\end{aligned}$$

where a_n and b_n are maximum and minimum penalties for non-zero and zero coefficients respectively.

Theorem 2.4 (Existence of the minimizer). Let $a_n = o(1)$ as $n \to \infty$. Then under assumptions [a-d] there is a local minimiser $\hat{\Theta}^*$ of $Q_n^*(\Theta)$ so that

$$(\hat{\Theta}^* - \Theta^\circ) = O_p(n^{-1/2} + a_n).$$

Proof. Let $\alpha_n = n^{-1/2} + a_n$, and $\{\Theta^\circ + \alpha_n\delta : ||\delta|| \leq d, \delta = (u,v)'\}$ be a ball around Θ°. Then for $||\delta|| = d$ we have

$$\begin{aligned} R_n(\delta) &= Q_n^*(\Theta^\circ + \alpha_n\delta) - Q^*(\Theta^\circ) \\ &\geq L_n(\Theta^\circ + \alpha_n\delta) - L_n(\Theta^\circ) + K_1 \\ &\geq L_n(\Theta^\circ + \alpha_n\delta) - L_n(\Theta^\circ) + K_2 \\ &\geq L_n(\Theta^\circ + \alpha_n\delta) - L_n(\Theta^\circ) + K_3 \end{aligned}$$

where

$$K_1 = n\sum_{i\in s_1}\lambda_i^*(|\beta_i^\circ + \alpha_n u_i| - |\beta_i^\circ|) \quad + n\sum_{j\in s_2}\gamma_j^*(|\phi_j^\circ + \alpha_n v_j| - |\phi_j^\circ|),$$

$$\begin{aligned} &\text{(Using triangular inequality)} : K_2 = -n\alpha_n\sum_{i\in s_1}\lambda_i^*|u_i| - n\alpha_n\sum_{j\in s_2}\gamma_j^*|v_j|, \\ &\text{(Penalties} \leq \alpha_n \text{ by definition)} : K_3 = -n\alpha_n^2(r_\circ + p_\circ)d. \end{aligned} \tag{2.17}$$

Last equation holds because of the decreasing speed of α_n. On the other hand, similar calculations to Theorem (2.3) results in

$$L_n(\Theta^\circ + \alpha_n\delta) - L_n(\Theta^\circ) \overset{n\to\infty}{=} n\alpha_n^2\{\delta' U_B\delta + o_p(1)\}. \tag{2.18}$$

Because (2.18) dominates (2.17), then for any gives $\eta > 0$, there is a large enough constant d so that

$$Pr[\inf_{||\delta||=d}\{Q_n^*(\Theta^\circ + \alpha_n\delta)\} > Q_n^*(\Theta^\circ)] \geq 1 - \eta.$$

This result shows that with probability at least $1-\eta$, there is a local minimiser in the ball $\{\Theta^\circ + \alpha_n\delta : ||\delta|| \leq d\}$ and as a result a minimiser $Q_n^*(\Theta)$, such that $||\hat{\Theta}^* - \Theta^\circ|| = O_p(\alpha_n)$. (See [Wang et al., 2007, Lemma 1], [Fan and Li, 2001])
The proof is completed. □

Theorem (2.4) implies that there exist a $\sqrt{n}-$*consistent* local minimiser $Q_n^*(\Theta)$, when tuning parameters (for non-zero variables) in DREGAR(p,0) converge to zero at the speed *faster* than $n^{-1/2}$.

In the next step we prove that under the case where the tuning parameter associated with zero variables in DREGAR(p,0) shrink to zero at a speed *slower* than $n^{-1/2}$, then their associate coefficients will be estimated exactly equal to zero with probability tending to 1. Further, in the next theorem we show that with increasing the penalties on the zero parameters at a certain speed, the probability of these coefficients to be estimated exactly zero tends to one.

Theorem 2.5 (Penalty weights for zero parameters). Let $b_n\sqrt{n} \to \infty$ and $||\hat{\Theta}^* - \Theta^\circ|| = O_p(n^{-1/2})$ then

$$Pr(\hat{\beta}^*_{s_1^c} = 0) \to 1, \qquad Pr(\hat{\phi}^*_{s_2^c} = 0) \quad \to 1.$$

Proof. This proof follows from the fact that the $Q_n^*(\hat{\Theta}^*)$ must satisfy

$$\frac{\partial Q_n^*(\Theta)}{\partial\beta_i}\bigg|_{\hat{\Theta}^*} = \frac{\partial L_n(\hat{\Theta}^*)}{\partial\beta_i} - n\lambda_i^*\,\text{sign}\,(\hat{\beta}_i^*)$$

$$= \frac{\partial L_n(\Theta^\circ)}{\partial \beta_i} + nU_i(\hat{\Theta}^* - \Theta^\circ)\{1 + o_p(1)\} - n\lambda_i^* \operatorname{sign}(\hat{\beta}_i^*) \tag{2.19}$$

where U_i is th i^{th} row of U_B and $i \in s_1^c$. The second term in (2.19) is a direct result of adding a $\pm X'\beta, \pm H_{(p)}\phi$ to $L_n(\hat{\Theta}^*)$. By using the central limit theorem, the first term in equation (2.19), $\sum_t e_t x'_{ti}$, is of order $O_p(n^{1/2})$ and the second term is $O_p(n^{1/2})$. Furthermore, both terms are dominated by $n\lambda_i^*$ since $b_n\sqrt{n} \to \infty$ (Expansion of [Wang et al., 2007], [Knight and Fu, 2000]). Then sign of $\frac{\partial Q_n^*(\hat{\Theta}^*)}{\partial \beta_i}$ is dominated by the sign of $\hat{\beta}_i^*$. Then $\hat{\beta}_i^* = 0$ in probability. Analogously, we can show that $Pr(\hat{\phi}_{s_2^c}^*) \xrightarrow{p} 1$. The proof is completed. □

Theorem (2.5) shows that adaptive-DREGAR(p,0) is capable of producing sparse solutions. Theorem (2.4) and (2.5) indicate that a $\sqrt{n} - consistent$ estimator $\hat{\Theta}^*$ must satisfy $Pr(\hat{\Theta}_2^* = 0) \to 1$. Then, adaptive-DREGAR(p,0) is a sparse model.

Theorem 2.6 (Consistency of adaptive-DREGAR(p,0)). Assume $a_n\sqrt{n} \to 0$ and $b_n\sqrt{n} \to \infty$. Then, under assumptions [a-d] we have

$$\sqrt{n}(\hat{\Theta}_1^* - \Theta_1^\circ) \xrightarrow{p} \text{MVN}(O, \sigma^2 U_0^{-1}),$$

where U_0 is the sub-matrix U_B corresponding to Θ_1°, and $\hat{\Theta}_1^*$ corresponds to non-zero elements of $\hat{\Theta}^*$.

Proof. It is concluded from Theorem (2.4) and (2.5) that $Pr(\hat{\Theta}_2^* = 0) \xrightarrow{p} 1$. Thus, the minimiser $Q_n^*(\Theta) \xrightarrow{with \quad pr \to 1} Q_n^*(\Theta_1)$. So it implies that the lasso estimator $\hat{\Theta}_1^*$ satisfies the following equation

$$\frac{\partial Q_n^*(\Theta_1)}{\partial \Theta_1}|_{\Theta_1 = \hat{\Theta}_1^*} = 0.$$

From Theorem (2.4), $\hat{\Theta}_1^*$ is a $\sqrt{n} - consistent$ estimator. Thus a Taylor expansion of the above equation yields

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}} \frac{\partial L_n(\hat{\Theta}_1^*)}{\partial \Theta_1} + F(\hat{\Theta}_1^*)\sqrt{n} \\ &= \frac{1}{\sqrt{n}} \frac{\partial L_n(\hat{\Theta}_1^\circ)}{\partial \Theta_1} + F(\hat{\Theta}_1^\circ)\sqrt{n} + U_0\sqrt{n}(\hat{\Theta}_1^* - \Theta_1^\circ) + o_p(1), \end{aligned}$$

where F is the first-order derivation of the tuning function

$$\sum_{i \in s_1} \lambda_i |\beta_i| + \sum_{j \in s_2} \gamma_j |\phi_j|,$$

and for n sufficiently large, $F(\hat{\Theta}_1^*) = F(\Theta_1^\circ)$. Thus,

$$\begin{aligned} (\Theta_1^\circ - \hat{\Theta}_1^*)\sqrt{n} &= \frac{U_0^{-1}}{\sqrt{n}} \frac{\partial L_n(\Theta_1^\circ)}{\partial \Theta_1} + o_p(1) \\ &\xrightarrow{d} N(0, \sigma^2 U_0^{-1}). \end{aligned}$$

The proof is completed. □

Theorem (2.6) implies that, adaptive DREGAR(p,0) is asymptotically an oracle estimator provided a_n tends to zero at the speed faster than $\sqrt{n}$ (or $a_n\sqrt{n} \to 0$) and simultaneously b_n increase at the speed slower than $\sqrt{n}$ (or $b_n\sqrt{n} \to \infty$).

2.8 Estimating the conditional variance of y_t

In Section § 2.2.1 we assumed that y_t has *known* conditional variance and established all the results using this assumption. In this section we consider the estimation of this variance.

Recall the DREGAR model from (2.1),

$$y_t = \sum_{i=1}^{r} x'_{ti}\beta_i + \sum_{j=1}^{p} \phi_j y_{t-j} + \sum_{l=1}^{q} \epsilon_{t-l}\theta_l + e_t$$
$$(y_t - \sum_{j=1}^{p} \phi_j y_{t-j}) = \sum_{i=1}^{r} x'_{ti}\beta_i + \sum_{l=1}^{q} \epsilon_{t-l}\theta_l + e_t.$$

Using the backward shift operator,

$$(1 - \sum_{j=1}^{p} \phi_j L^j) y_t = \sum_{i=1}^{r} x'_{ti}\beta_i + \sum_{l=1}^{q} (\theta_l L^l)\epsilon_t + e_t, \tag{2.20}$$

where

$$\epsilon_t = (1 - \sum_{j=1}^{p} \phi_j L^j) y_t - \sum_{i=1}^{r} x'_{ti}\beta_i. \tag{2.21}$$

Substituting (2.21) in (2.20),

$$\left((1 - \sum_{l=1}^{q} \theta_l L^l)(1 - \sum_{j=1}^{p} \phi_j L^j) \right) y_t = \sum_{i=1}^{r} (1 - \sum_{l=1}^{q} \theta_l L^l) x'_{ti}\beta_i + e_t.$$

Converting this equation to an infinite moving average results in

$$y_t = \sum_{i=1}^{r} \frac{1}{L(\phi)} x_{ti}\beta_i + \frac{1}{L(\phi)L(\theta)} e_t,$$

where

$$L(\phi) = (1 - \sum_{l=1}^{q} \phi_l L^l), \qquad L(\theta) = (1 - \sum_{j=1}^{p} \theta_j L^j).$$

Let $a_1, a_2, a_3, \ldots, a_p$ and $b_1, b_2, b_3, \ldots, b_q$ be the roots of $L(\phi)$ and $L(\theta)$ respectively. Therefore, it is possible to rewrite $L(\phi)$ and $L(\theta)$ as

$$L(\phi) = (1 - \sum_{l=1}^{p} \phi_l L^l) = (a_1 - L)(a_2 - L)(a_3 - L)\ldots(a_p - L),$$

$$L(\theta) = (1 - \sum_{j=1}^{q} \theta_j L^j) = (b_1 - L)(b_2 - L)(b_3 - L)\dots(b_q - L),$$

and using $\frac{1}{a-x} = \sum_{i=0}^{\infty} \frac{1}{a}(\frac{x}{a})^i$,

$$L(\phi)^{-1} = \left(1 - \sum_{l=1}^{p} \phi_l L^l\right)^{-1} = \frac{1}{(a_1 - L)(a_2 - L)(a_3 - L)\dots(a_p - L)} = \prod_{l=1}^{p}(\sum_{k=0}^{\infty} \frac{1}{a_l}(\frac{L}{a_l})^k)$$
$$L(\theta)^{-1} = \left(1 - \sum_{j=1}^{q} \theta_j L^j\right)^{-1} = \frac{1}{(b_1 - L)(b_2 - L)(b_3 - L)\dots(b_q - L)} = \prod_{j=1}^{q}(\sum_{k=0}^{\infty} \frac{1}{b_j}(\frac{L}{b_j})^k).$$

Finally, the DREGAR model can be written as

$$y_t = \sum_{i=1}^{r} \left(\prod_{j=1}^{p}(\sum_{k=0}^{\infty} \frac{1}{a_j}(\frac{L}{a_j})^i)\right) x_{ti}\beta_i + \left(\prod_{j=1}^{q}(\sum_{i=0}^{\infty} \frac{1}{b_j}(\frac{L}{b_j})^i)\right)\left(\prod_{l=1}^{p}(\sum_{k=0}^{\infty} \frac{1}{a_l}(\frac{L}{a_l})^k)\right) e_t.$$

From this, the variance of y_t is given by

$$\mathbb{V}ar(y_t|x) = \mathbb{V}ar\left[\left(\prod_{j=1}^{q}(\sum_{i=0}^{\infty} \frac{1}{b_j}(\frac{L}{b_j})^i)\right)\left(\prod_{l=1}^{p}(\sum_{k=0}^{\infty} \frac{1}{a_l}(\frac{L}{a_l})^k)\right) e_t\right]. \tag{2.22}$$

This can be shown by a single geometric series

$$\mathbb{V}ar(y_t|x) = \sum_{i=0}^{\infty} \Omega^i \sigma^2,$$

where Ωs can be computed based on coefficients as in (2.22). In the special case where $e_t = 0$ for all $t \leq 0$ we get

$$\mathbb{V}ar(y_t|x) = \sum_{i=0}^{t} \Omega^i \sigma^2,$$

that is a function of σ and the coefficients. σ is assumed to be known prior to the analysis and parameters are estimated from the model. Then, an estimator for σ_y^2 is given by

$$\widehat{\mathbb{V}}ar(y_t|x) = \sum_{i=0}^{t} \hat{\Omega}^i \sigma^2.$$

2.9 Implementation

The most trivial implementation of DREGAR can be performed by assuming a grid of three values for λ, γ and τ and solving the penalized likelihood within the grid. However, in this section we propose two algorithms for estimating the parameters in DREGAR and adaptive-DREGAR that are computationally less complex than the naive way of grid search. To this end, we use LARS [Efron et al., 2004], which has a good performance in the correlated frameworks [Hebiri and Lederer, 2013].

We should stress that ϵ is *unknown* in DREGAR and must be estimated from an auxiliary step. For $k = 0, 1, \dots$, we propose the following 6-step algorithm for DREGAR:

Step 1. Estimate $\phi_{(k)}(\gamma_{(k)})$ by minimizing $||y - H_{(p)}\phi||_2^2 + \gamma|\phi|_1$ where the tuning parameter is selected using BIC, AIC, GCV, CV etc. We assume that the model selection criteria is the same in all steps below.

Step 2. Estimate $\beta_{(k)}(\lambda_{(k)})$ by minimizing $||(y - H_{(p)}\hat{\phi}_{(k)}) - X'\beta||_2^2 + \lambda|\beta|_1$.

Step 3. Estimate $\theta_{(k)}(\tau_{(k)})$ by minimizing $||(y - H_{(p)}\hat{\phi}_{(k)} - X'\hat{\beta}_{(k)}) - \hat{H}_{(q)}\theta||_2^2 + \tau|\theta|_1$.

Step 4. Update $\beta_{(k)} \to \beta_{(k+1)}$ and $\lambda_{(k)} \to \lambda_{(k+1)}$ by minimizing $||(y - \hat{H}_{(q)}\hat{\theta}_{(k)}) - X'\beta||_2^2 + \lambda|\beta|_1$.

Step 5. Update $\phi_{(k)} \to \phi_{(k+1)}$ and $\gamma_{(k)} \to \gamma_{(k+1)}$ by minimizing $||(y - \hat{H}_{(q)}\hat{\theta}_{(k)} - X'\hat{\beta}_{(k+1)}) - H_{(p)}\phi)||_2^2 + \gamma|\phi|_1$.

Step 6. Update $\theta_{(k)} \to \theta_{(k+1)}$ and $\tau_{(k)} \to \tau_{(k+1)}$ by minimizing $||(y - H_{(p)}\hat{\phi}_{(k+1)} - X'\hat{\beta}_{(k+1)}) - \hat{H}_{(q)}\theta)||_2^2 + \tau|\theta|_1$.

Step 7. Return to `Step 4` provided the algorithm does not meet the stopping criteria.

The same algorithm can be used for adaptive-DREGAR. However, we propose a two-step algorithm based on adaptive-lasso [Zou, 2006] for adaptive-DREGAR that is computationally less complex and involves fewer steps ($2 < 6$) compared to the non-adaptive one. The algorithm assumes the same tuning parameter for the entire parameter space, but with different weights. The first step provides an estimation for ϵ from an auxiliary adaptive-DREGAR(p,0) whereas the second step leads to an estimation for the entire parameter space, Θ. We propose that iterating these two steps refines the estimation of ϵ in the first step and improves the estimation of the parameters in the final stage. The algorithm can be summarized as following:

Step 1. For $k = 0, 1, \ldots$, estimate ϵ from the DREGAR(p,0) model by solving iteratively for k

$$(\hat{\beta}_{(k+1)}, \hat{\phi}_{(k+1)}) = \arg\min_{\beta,\phi} ||y - X'\hat{\beta} - H_{(p)}\hat{\phi}||_2^2 + \lambda^*_{(k)}|\beta|_1 + \gamma^*_{(k)}|\phi|_1,$$

where $\lambda^*_{(k)} = \omega_1/|\beta_{(k)}|$, $\gamma^*_{(k)} = \omega_1/|\phi_{(k)}|$. $\beta_{(0)}$ and $\phi_{(0)}$ are initial estimations from OLS or lasso and we assume the same ω_1 for both terms to simplify the problem to the ordinary adaptive-lasso problem. The procedure of estimating/re-estimating in this step is continued till a stopping criterion e.g., minimum AIC, BIC, GCV or CV, is met.

Step 2. Estimate ϵ from $\hat{\epsilon} = y - X'\hat{\beta} - H_{(p)}\hat{\phi}$ using the estimations provided from the first step, and substituting in the full model,

$$y = X'\beta + H_{(p)}\phi + \hat{H}_{(q)}\theta + e,$$

and re-estimating all parameters by:

$$\hat{\Theta} = (\hat{\beta}_{(k+1)}, \hat{\phi}_{(k+1)}, \hat{\theta}_{(k+1)}) = \arg\min_{\beta,\phi,\theta} ||y - X'\beta - H_{(p)}\phi - \hat{H}_{(q)}\theta||_2^2 + \lambda^*|\beta|_1 + \gamma^*|\phi|_1 + \tau^*|\theta|_1,$$

where $\lambda^* = \omega_2/|\beta_{(k)}|$, $\gamma^* = \omega_2/|\phi_{(k)}|$ and $\tau^* = \omega_2/|\theta_{(k)}|$ for $k = 0, 1, \ldots$. Similar to the first step, the parameters are estimated using an estimate/re-estimate procedure.

More formally we minimize the following *penalized likelihoods* with respect to the parameters:

$$\textbf{Step 1. } Q^*_{S_1}(\Theta) = \sum_t \left((y_t - x_t'\beta) - \sum_{i=1}^{p} \phi_i y_{t-i} \right)^2 + \sum_{i=1}^{r} \lambda^*_{i,1}|\beta_i| + \sum_{j=1}^{p} \gamma^*_{j,1}|\phi_j|$$

$$\textbf{Step 2. } Q^*_{S_2}(\Theta) = \sum_t \left((y_t - x_t'\beta) - \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{j=1}^{q} \theta_j \hat{\epsilon}_{t-j} \right)^2 + \sum_{i=1}^{r} \lambda^*_i|\beta_i| + \sum_{j=1}^{p} \gamma^*_j|\phi_j| + \sum_{k=1}^{q} \tau^*_k|\theta_k| \quad ,$$

or equivalently in matrix form,

$$\begin{cases} Q^*_{S_1}(\Theta) = (y - X'\beta - H_{(p)}\phi)'(y - X'\beta - H_{(p)}\phi) \\ \qquad\qquad + \lambda^*_1|\beta| + \gamma^*_1|\phi| \\ Q^*_{S_2}(\Theta) = (y - X'\beta - H_{(p)}\phi - \hat{H}_{(q)}\theta)'(y - X'\beta - H_{(p)}\phi - \hat{H}_{(q)}\theta) \\ \qquad\qquad + \lambda^*|\beta| + \gamma^*|\phi| + \tau^*|\theta| \end{cases} .$$

The first step provides an initial guess for ϵ. Replacing the estimations from Step 2 in 1 and repeating the steps iteratively provides a solution to adaptive-DREGAR.
In both algorithms we define the stopping criteria by either setting a tolerance on the difference of consecutive estimations; or by fixing a maximum number of iterations and then taking the estimates that achieve the minimum BIC or AIC,

$$BIC = -2\text{loglik} + p\log(T)$$
$$AIC = -2\text{loglik} + 2p$$

where loglik, p, T are estimated (non-penalized) log-likelihood using the parameters that are estimated from both algorithms, the number of non-zero parameters and total observations respectively.

2.9.1 Choosing the tuning parameters

All regularization methods rely heavily on the choice of the tuning parameters, as these control the amount of regularization and sparsity in the estimations. Consequently, choosing a proper value for any tuning parameter is crucial. A number of methods have been proposed in the literature to select the tuning parameters and weights in adaptive-lasso. One can utilize Cross Validation (CV) or Generalized Cross Validation (GCV), see e.g. [Arlot et al., 2010, Usai et al., 2009, Tibshirani, 1996] and citations therein. Although these techniques are recommended and discussed in the original paper of [Tibshirani, 1996], using cross validation for model selection in time-dependent frameworks is criticized by some authors e.g., [Medeiros, 2012] and [Shao, 1993]. BIC and AIC are also extensively studied in the literature, e.g. [Wang, 2007]. In particular, [Zhang, 2010] recommends using BIC and proves that it enjoys the oracle property in sparse model selection. The paper proposes also a Generalized Information Criterion (GIC) that encompasses both AIC and BIC. Finally, [Hirose et al., 2011] propose Mallows's C_p criteria for selecting tuning parameters.

All AIC, BIC, GCV and C_p are implemented in the R package https://cran.r-project.org/web/packages/DREGAR/index.html associated with the adaptive-DREGAR method, but we will consider closely CV and BIC in our simulation and real data studies.

2.9.2 Choosing model orders p and q

Some notes are required for selecting the autoregressive orders p and q. We propose two general approaches:

1. Setting a grid of P and Q values and choosing the model with minimal BIC or AIC within the grid.
2. Setting an upper limit for P and Q and letting the model choose the optimal orders.

Although the two approaches above look similar, the main difference is that the second approach needs to remove the first $P + Q$ observations a priori. That may cause problems in high-dimensional cases where (usually) the number of observations is rather small. Then a rule of thumb is to choose the first approach for small datasets and the second one for large datasets.

2.9.3 R package

An implementation of (adaptive) DREGAR using (two) six-step algorithm and Mallows's C_p, AIC, BIC and GCV for model selection is provided in the complementary R package that accompanies this chapter. This is performed by the function `dregar2`. Moreover, the six-step algorithm for DREGAR and adaptive-DREGAR is implemented in the function `dregar6`. Both functions allow different combination of orders for dynamic and AR orders as well as several options for standardizing the data and setting the number of iterations prior to the analysis. This R package encompasses two more functions to simulating data from an arbitrary DREGAR model and generating stationary autoregressive coefficients. We refer to the R package manual (https://cran.r-project.org/web/packages/DREGAR/DREGAR.pdf) for a detailed description of the package.

2.10 Simulation study

In this section we follow the general outline proposed in [Ulgen, 1994, Tibshirani, 1996, Lozano, 2013, Y. Nardi, 2011] to simulate data from our models. In particular, simulations in this section are divided into two groups. In the first group, data are generated from a DREGAR(p,q) model and the tuning parameters are selected by minimizing the 10-fold cross validation error, while in the second one we choose the tuning parameters corresponding to the minimum BIC.

Under assumptions [a-d], we design the simulation study with varying number of parameters and models as following:

- The coefficients are sampled from a uniform distribution in $(-1, 1)$, where time-dependent parameters are chosen so that the stationary polynomials have all roots unequal and outside the unit circle.

- 90%, 70% and 10% percent of REG coefficients are set to zero.
- The covariates are generated independently from a random AR(1) Gaussian process.
- $e \overset{iid}{\sim} \sigma \times N(0,1)$ with varying levels of noise $\sigma \in \{0.5, 1, 1.5\}$.
- Data are simulated for a range of values of T and r as well as $\{p,q\} \in \{1,2\}$.
- Each combination of parameters is repeated 25 times.
- Cross validation and BIC are used for choosing the optimal tuning parameters.

For all datasets, we fit lasso (L), adaptive-lasso (AD), DREGAR (D-L), and adaptive-DREGAR (D-AD). The models are compared in terms of Mean Squared Error, $MSE = \frac{1}{25}\sum_{i=1}^{25}(\hat{\Theta} - \Theta^{\circ})^2$, and BIC.

2.10.1 Simulation results

Figure (2.3) shows the comparisons between lasso (adaptive-lasso) versus DREGAR (adaptive-DREGAR) for varying number of covariates $r = (20, 120, 540)$ (top label) and observations $T = (20, 40, 60)$ (middle label) and for different time-dependent parameters $p \in \{1,2\}$, $q \in \{1,2\}$(bottom labels). We propose the ratio of MSE amongst models namely adaptive-lasso and lasso versus adaptive-DREGAR and DREGAR respectively. Obviously, if this ratio takes a value greater than one, then the MSE in the denominator is less than the nominator and consequently the model in the denominator outbids the other one.

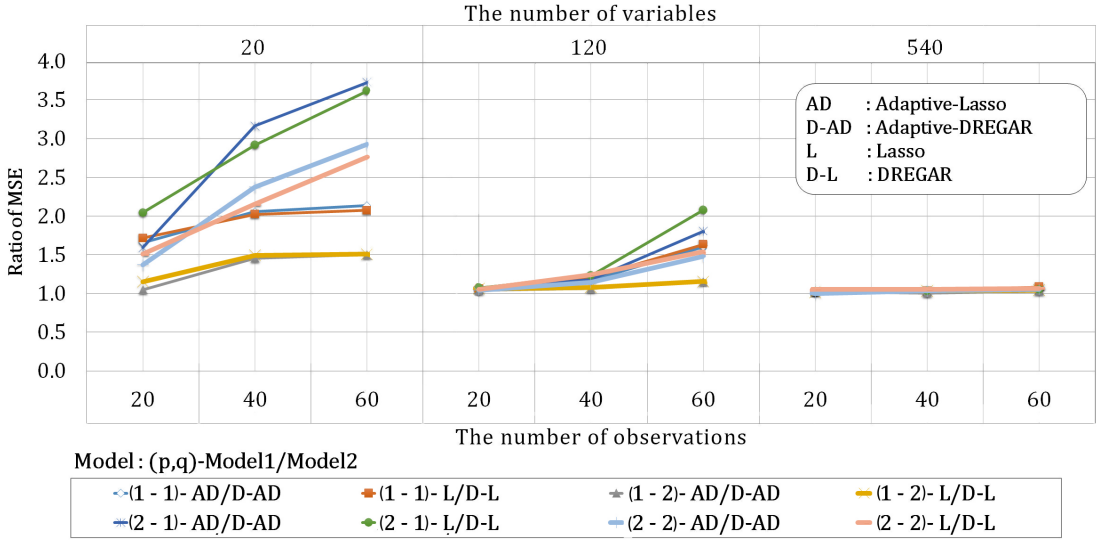

FIGURE 2.3: Comparison of adaptive-lasso (lasso) and adaptive-DREGAR (DREGAR) with respect to MSE ratio for varying number of covariates, observations, p and q. Tuning parameters for all models are chosen by CV.

As it is evident from Figure (2.3), adaptive-DREGAR (D-AD) and DREGAR (D-L) perform better than ordinary adaptive-lasso (AD) and lasso (L), respectively, in terms of MSE ratio and for all combinations of p and q as well as for different values of r and T. With regards to the number of covariates, as expected, the figures shows that an increase in r compared to $p+q$ results in a decrease in the effect of

DA and AR components in the model, as noted also by [Hibbs Jr, 1973]. In these cases, D-AD and R-L tend to ordinal AD and L and their corresponding MSE ratios tend to one. The figure also shows that DREGAR outperforms the opponent for $T \ll r$. However, DREGAR shows considerably better results than AL and L when r/T decreases, as shown in Figure (2.4).

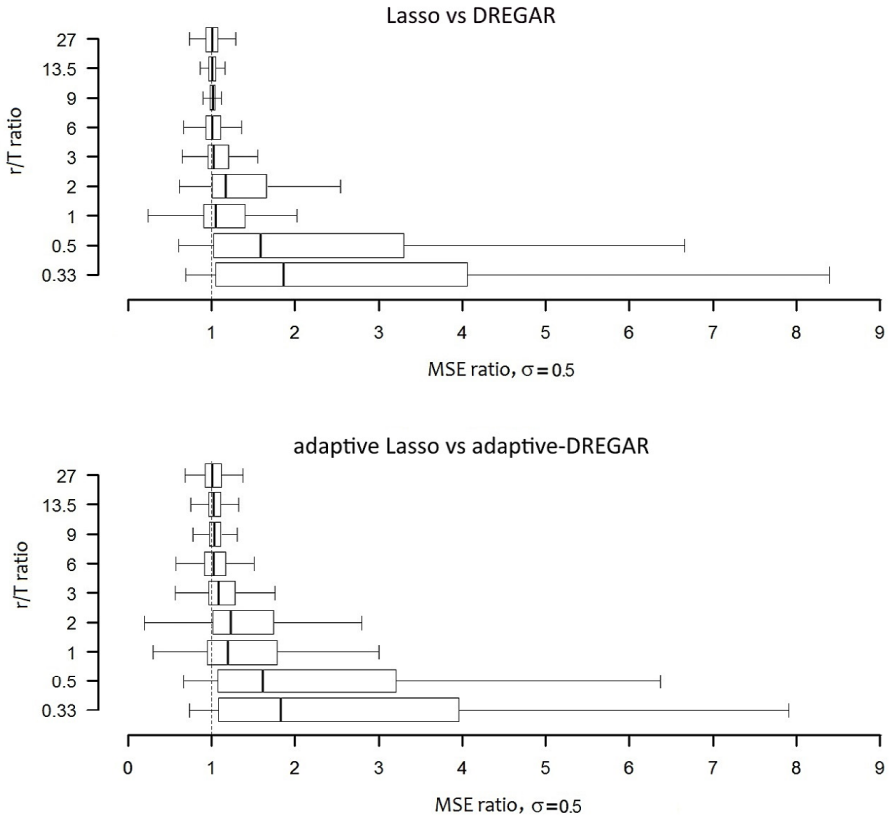

FIGURE 2.4: Comparison of DREGAR and Lasso (top) as well as adaptive-Lasso versus adaptive-DREGAR (bottom) with respect to MSE ratio of estimations under $\sigma = 0.5$ and sliding (r/T).

Figure (2.5) and (2.6) show similar comparisons for the case when the tuning parameters are selected by BIC. In particular, Figure (2.5) compares adaptive-DREGAR and adaptive-lasso in terms of BIC for $T = 50, 100, 150, 200, 250$ and $r = 25, 75, 200, 300, 400$. The results show that, increasing the number of data points, T, results in a significant improvement in BIC for adaptive-DREGAR compared to adaptive-lasso. However, adaptive-DREGAR shows a slightly better performance than adaptive-lasso if $T \ll r$. Figure (2.6) shows the comparison between the models in terms of the mean squared error of the regression parameter estimates. As it is evident from this figure, adaptive-DREGAR estimates the coefficients with a lower level of bias compared to adaptive-lasso for all combinations of T and r.

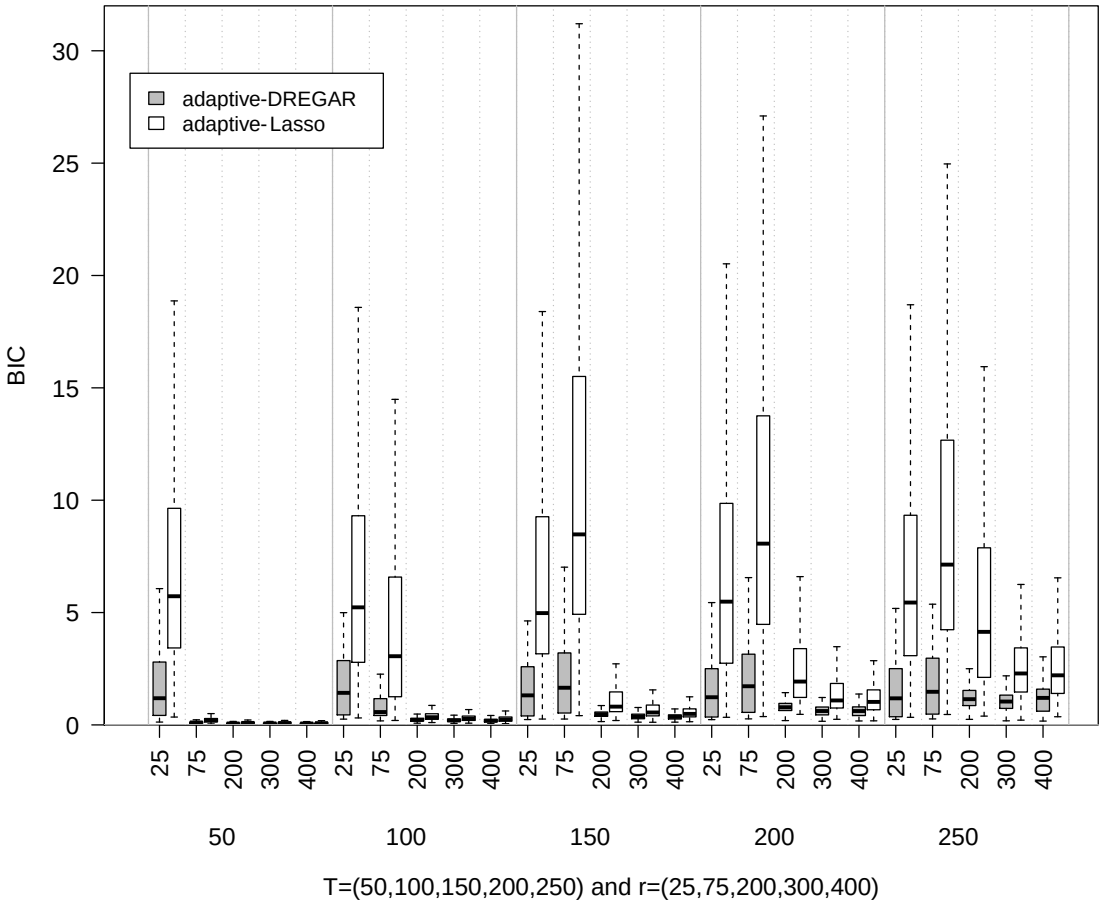

FIGURE 2.5: Comparison of adaptive-lasso and adaptive-DREGAR in terms of BIC under different values for r and T. The tuning parameters are chosen by BIC.

2.11 Real data illustration

2.11.1 Analysis of air pollution data

In this section, we show the performance of the model on the National Mortality, Morbidity and Air Pollution Study (NMMAPS) dataset. This dataset is publicly available from http://www.ihapss.jhsph.edu/data/NMMAPS/ and contains daily mortality, air pollution, and weather data for 108 cities in the US from January 1, 1987 to December 31, 2000. The variables include six indicators for mortality (total non-accidental, cardiovascular disease, respiratory, pneumonia, chronic obstructive pulmonary disease, accidental), six indicators of air pollution (repairable particulates (PM10)/(PM25), carbon monoxide (CO), ozone (O_3), sulphur dioxide (SO_2), nitrogen dioxide (NO_2)) as well as three indicators of weather (temperature (T), dew point temperature (D), relative humidity (H)). Similar to [Wu and Wang, 2012] we study the relationship between ground level of ozone and indicators of air pollution and weather conditions in Chicago in 1995. Differently to [Wu and Wang, 2012], we take the effect of carbon monoxide (CO) into account. The covariates in the model consist of NO_2, SO_2, CO, PM10, temperature and relative humidity as well as all two-ways interactions. We show the interactions by initials, for instance NS represents the interaction between NO_2 and SO_2. A total number of 365 observations and 21 covariates are included in the analysis. All covariates and response are normalized to zero mean and unit

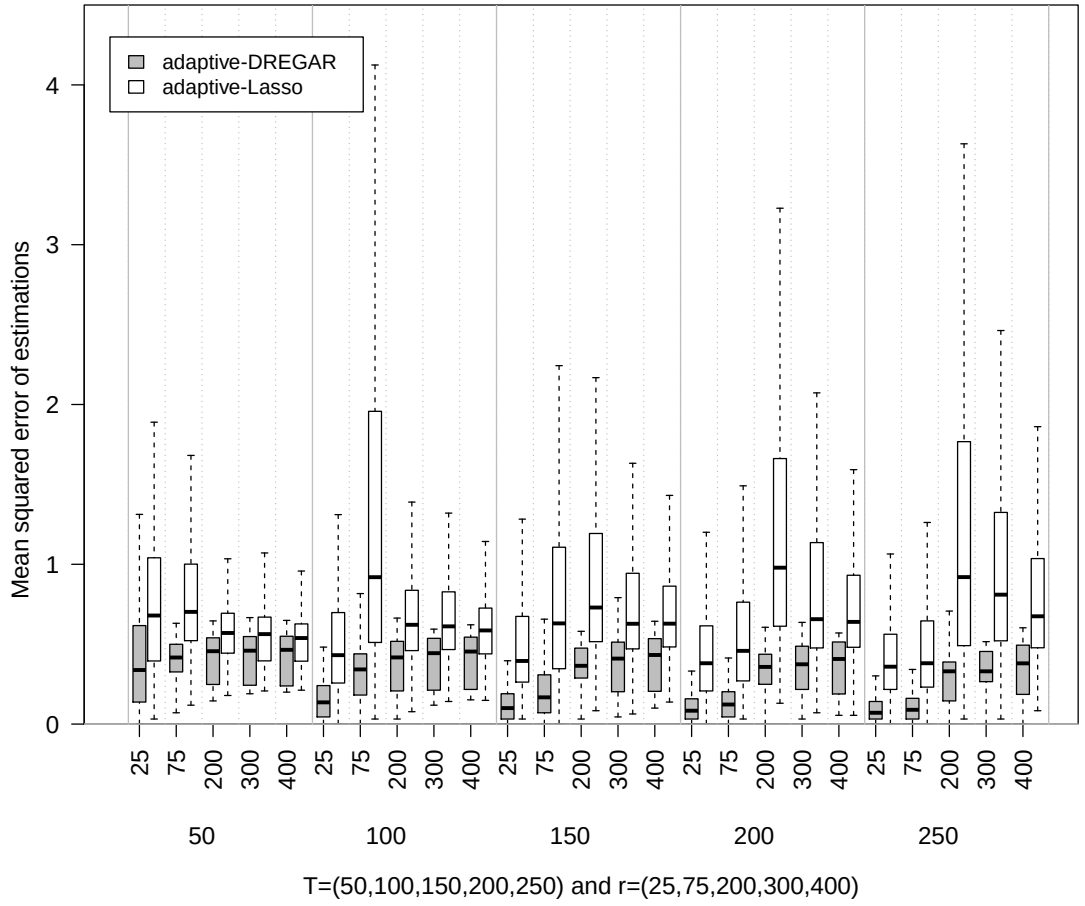

FIGURE 2.6: Comparison of adaptive lasso and adaptive-DREGAR in terms of mean squared error of $\hat{\beta}$ under varying values for r and T. The tuning parameters are chosen by BIC.

variance. We compare the DREGAR model with lasso, REGARMA(3,1) ([Wu and Wang, 2012] optimal model), DREGAR(p+q,0), DREGAR(p,0), DREGAR(0,q+p) and DREGAR(0,q) on the basis of a number of commonly used criteria: BIC, AIC, Quasi-likelihood Information Criteria (QIC) [Pan, 2001] and Consistent AIC (CAIC) [Bozdogan, 1987].

Following the second approach in Section § 2.9.2, we propose $P = 5$ and $Q = 5$ for the autoregressive orders. The parameters are estimated using the algorithm in Section § 2.9, setting a maximum of 15 for the iterations and selecting the tuning parameters by CV. Table (2.1) provides a detailed illustration of the parameter estimations as well as information for comparison of the models. *Non-zero* time series coefficients in the middle-bottom of the table propose an order of four and three for DREGAR as well as DREGAR(1,0) and DREGAR(0,3) for the other models. DREGAR(4,3) shows better results than REGARMA, DREGAR(p,0) and DREGAR(0,q) with respect to model performance as shown in the top panel of table (2.1). In line with [Wu and Wang, 2012], our results show several significant interactions, especially those between sulphur dioxide-temperature (ST) and humidity (SH), as well as between particulates and temperature (PT). However, there are also some differences: for example DREGAR(5,5) assigns a zero or significantly low weight to PH, SP and NT where REGARMA does not. The additional variable CO shows a significant effect on ground level of O_3 and non-zero effect for the interaction with weather indicators, CT and CH. We further report the Ljung-Box test [Box and Pierce, 1970] statistics in the bottom of the table (2.1). With the exception of lasso and DREGAR(10,0), all

Model comparison

Model	BIC	AIC	CAIC	QIC
Lasso	15467.69	15385.79	15488.69	42.39
DREGAR(5,5)	10302.58*	10182.55*	10333.58*	29.11*
DREGAR(10,0)	11419.65	11299.61	11450.65	32.25
DREGAR(0,10)	10860.55	10740.51	10891.55	30.68
DREGAR(0,5)	10854.76	10753.72	10880.76	30.20
DREGAR(5,0)	11414.39	11313.36	11440.39	31.75

Parameter estimation

Variables	Lasso	REGARMA(3,1)	DREGAR(5,5)	DREGAR(10,0)	DREGAR(0,10)	DREGAR(0,5)	DREGAR(5,0)
Temperature	5.29	4.14	4.67	3.66	5.33	5.36	3.56
PM10(P)	0	0	0	0	0	0	0
SO2(S)	-10.53	-9.74	-11.49	-8.69	-12.46	-11.86	-8.49
NO2(N)	-2.87	-1.81	-1.94	-1.37	-1.69	-2.10	-1.52
Humidity(H)	-1.10	-1.94	-1.34	-1.18	-1.32	-1.23	-1.12
CO(C)	-0.18	-	-0.44	-1.55	-0.93	-0.32	-1.27
NS	0	0.89	0.42	0.55	0.06	0.22	0.62
NP	-0.41	-1.26	-0.47	-0.98	-0.43	-0.56	-1.15
NT	0	-1.30	0	0	0	0	0
NH	1.08	-0.90	0.34	0.22	0.01	0.38	0.42
SP	0	0.77	0	0.42	0	0	0.59
ST	6.29	4.40	6.96	4.81	7.69	7.15	4.29
SH	5.34	6.55	5.63	4.59	5.91	5.89	4.95
PT	3.60	4.89	2.88	2.57	3.14	3.22	2.56
PH	0	-2.40	-0.08	0	0	0	0
TH	0	0	0	0	0.19	0.04	0
CN	0	-	0	0	0.25	0.25	0
CS	0	-	0	0	0	0	0
CP	0	-	0	0	0	0	0
CT	-0.47	-	-1.30	-0.41	-1.34	-1.13	-0.17
CH	-1.61	-	-0.57	0	-0.35	-1.22	-0.57

Time series coefficients

	Lasso	REGARMA(3,1)	DREGAR(5,5)	DREGAR(10,0)	DREGAR(0,10)	DREGAR(0,5)	DREGAR(5,0)
	-	$\phi_1 = 1.27$	$\phi_1 = 0.13$	$\phi_1 = 0.36$	$\theta_1 = 0.46$	$\theta_1 = 0.47$	$\phi_1 = 0.36$
	-	$\phi_2 = -0.28$	$\phi_2 = 0$	$\phi_2 = 0$	$\theta_2 = 0$	$\theta_2 = 0$	$\phi_2 = 0$
	-	$\phi_3 = 0$	$\phi_3 = 0$	$\phi_3 = 0$	$\theta_3 = 0.07$	$\theta_3 = 0.09$	$\phi_3 = 0$
	-	$\theta_1 = -0.88$	$\phi_4 = 0.036$	$\phi_4 = 0$	$\theta_4 = 0$	$\theta_4 = 0$	$\phi_4 = 0$
	-	-	$\phi_5 = 0$	$\phi_5 = 0$	$\theta_5 = 0$	$\theta_5 = 0$	$\phi_5 = 0$
	-	-	$\theta_1 = 0.31$	$\phi_6 = 0$	$\theta_6 = 0$	-	-
	-	-	$\theta_2 = 0.03$	$\phi_7 = 0$	$\theta_7 = 0$	-	-
	-	-	$\theta_3 = 0.14$	$\phi_8 = 0$	$\theta_8 = 0$	-	-
	-	-	$\theta_4 = 0$	$\phi_9 = 0$	$\theta_9 = 0$	-	-
	-	-	$\theta_5 = 0$	$\phi_{10} = 0$	$\theta_{10} = 0$	-	-

Ljung-Box statistic

	Lasso	REGARMA(3,1)	DREGAR(5,5)	DREGAR(10,0)	DREGAR(0,10)	DREGAR(0,5)	DREGAR(5,0)
P-value	0	0.635	0.82	0.08	0.813	0.704	0.09

TABLE 2.1: (Top) comparing lasso, DREGAR and REGARMA with respect to BIC, AIC, CAIC and QIC where the asterisk denotes the minimum value. (Middle-top) parameter estimation for regression terms. (Middle-bottom) Corresponding estimation for time-dependent coefficients. (Bottom) Ljung-Box p-value for the null hypothesis of residuals following white noise.

models show good fitting, i.e. no evidence against the white noise assumption. Figure (2.7) displays the scatter plot of lasso and DREGAR(5,5) fitted versus observed response, the residuals from the DREGAR(4,3) model and the corresponding sample ACF and PACF. The small curvature in the scatter plot, mentioned also by [Wu and Wang, 2012], can be an indication of a particular weather condition that results in an interaction between primary pollutants. The sample ACF and PACF plot suggest that the residuals are indeed white noise as confirmed also by the p-value of the Ljung-Box test (0.82).

Finally, we have also compared the fit of the best DREGAR model, DREGAR(4,3), with a DREGAR(0,7) model (the same as a REGAR(7) model), in order to assess the benefit in having different autoregressive structures for the response and the predictors, a unique feature of the model that we propose in this chapter. Without penalising the coefficients, the maximum likelihood for DREGAR(4,3) is -1106.884 and that of DREGAR(0,7) is -1110.832, suggesting an improved fit for the DREGAR(4,3) model.

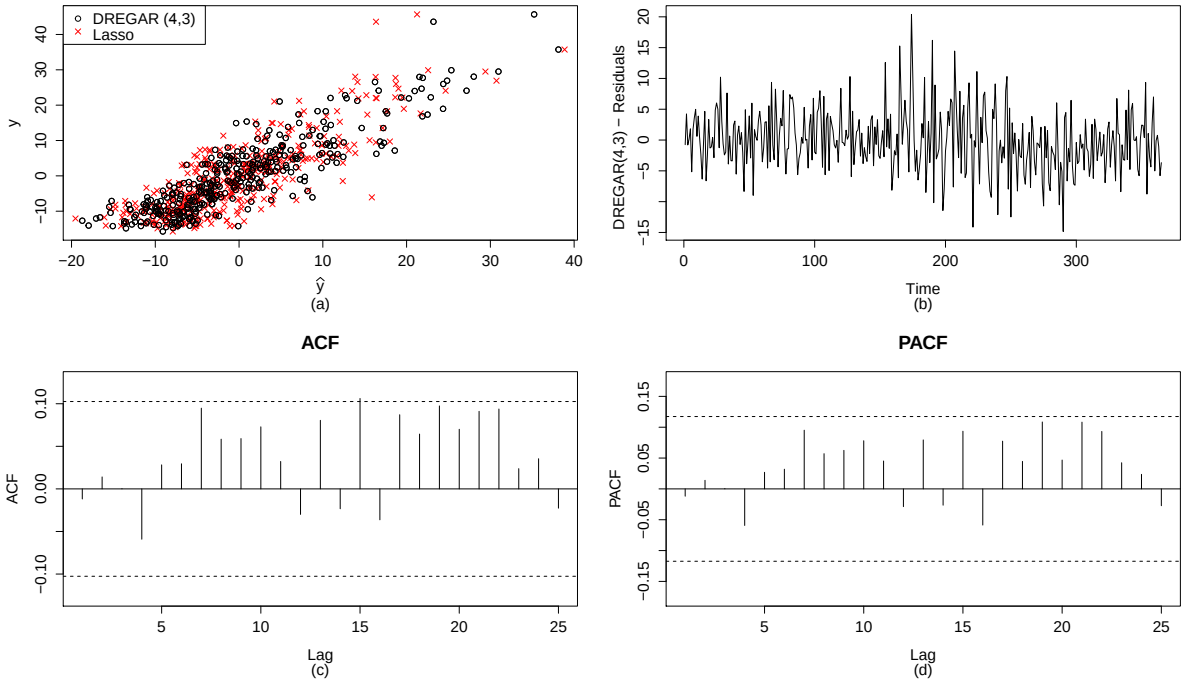

FIGURE 2.7: (a) scatter plot of DREGAR(4,3) and lasso fitted versus observed y, (b) DREGAR(4,3) residuals, (c) sample ACF and PACF for DREGAR(4,3) residuals,(d) sample PACF for DREGAR(4,3) residuals.

2.11.2 Analysis of stock market data

For the second real application we take an example from the stock market. To this end we apply DREGAR(p,q), DREGAR(p,0) and DREGAR(0,q) to DowJones30 daily returns from 2015. Data are collected from yahoo finance (`https://finance.yahoo.com`) and contain 251 closing prices for 30 indices in the DowJones market. We take the IBM index as the response and the remaining 29 indices as the covariates and study the correlations via the DREGAR family of models. The variables are listed as follows: 3M (MMM), American Express (AXP), Alcoa (AA), AT&T (T), Bank of America (BAC), Boeing (BA), Caterpillar (CAT), Chevron (CVX), Cisco Systems (C), Coca-Cola (KO), DuPont (DD), ExxonMobil (XOM), General Electric (GE), Hewlett-Packard (HPQ), The Home Depot (HD), Intel (INTC), IBM (IBM), Johnson & Johnson (JNJ), JPMorgan Chase (JPM), Kraft (KRFT), McDonald's (MCD), Merck (MRK), Microsoft (MSFT), Pfizer (PFE), Procter & Gamble (PG), General Motors (GM), United Technologies (UTX), Verizon (VZ), Wal-Mart (WMT), Walt Disney (DIS).

We apply first differences of the log-prices to get stationary returns [Kwiatkowski et al., 1992]. DREGAR(5,5), DREGAR(10,0), DREGAR(0,10), DREGAR(5,0) and DREGAR(0,5) are applied to the data and the tuning parameter is selected using CV. The models are compared on the basis of BIC, AIC, CAIC, QIC, Ljung-Box statistic and sparsity. The results are shown in Table (2.2).

This tables shows that DREGAR(5,5) is the winner amongst other methods with respect to BIC, AIC and CAIC as well as sparsity. Fitting DREGAR(5,5) to data results in an order of 3 for the dynamic term and an order of 4 for the residuals. So the final selected model is DREGAR(3,4). Among the most

significant (non-zero) variables, the model selects: MSFT (coefficient 0.3), HPQ (0.23), VZ (0.20), MMM (0.14), MRK (0.13) and CVX (0.10). Figure (2.8) top shows observed y versus fitted values for lasso and DREGAR(3,4). From this figure, DREGAR(3,4) has a better fit compared to lasso in terms of the correlation between the observed and fitted values ($\rho_{y,\hat{y}_{DREGAR(3,4)}} = 0.831$, $\rho_{y,\hat{y}_{Lasso}} = 0.819$). Finally, the sample ACF and PACF at the bottom of figure (2.8) confirm the results from the Ljung-Box statistic, showing that the residuals from DREGAR(3,4) are white noise.

Following the same steps as the previous section, we compare the fit of the best DREGAR(3,4) with DREGAR(0,7) model. Without penalising the coefficients, the maximum likelihood for DREGAR(3,4) is -243.98 and that of DREGAR(0,7) is -251.41, suggesting an improved fit for the DREGAR(4,3) model.

Model	BIC	AIC	CAIC	QIC	Ljung-Box p-value	#Non-zero
Lasso	600.74	547.84	615.72	2.4	0.52	15
DREGAR(5,5)	575.10*	528.91*	598.43*	2.3	0.83	13
DREGAR(10,0)	585.94	542.61	610.10	2.4	0.56	14
DREGAR(0,10)	589.70	536.82	604.70	2.3	0.54	15
DREGAR(5,0)	590.46	537.58	605.47	2.4	0.58	15
DREGAR(0,5)	596.08	543.19	611.10	2.3	0.65	15

TABLE 2.2: Comparison of lasso and DREGAR for the DowJones30 dataset on the basis of BIC, AIC, CAIC, QIC, sparsity and Ljung-Box statistic. For the information criteria, the asterisk denotes the minimum.

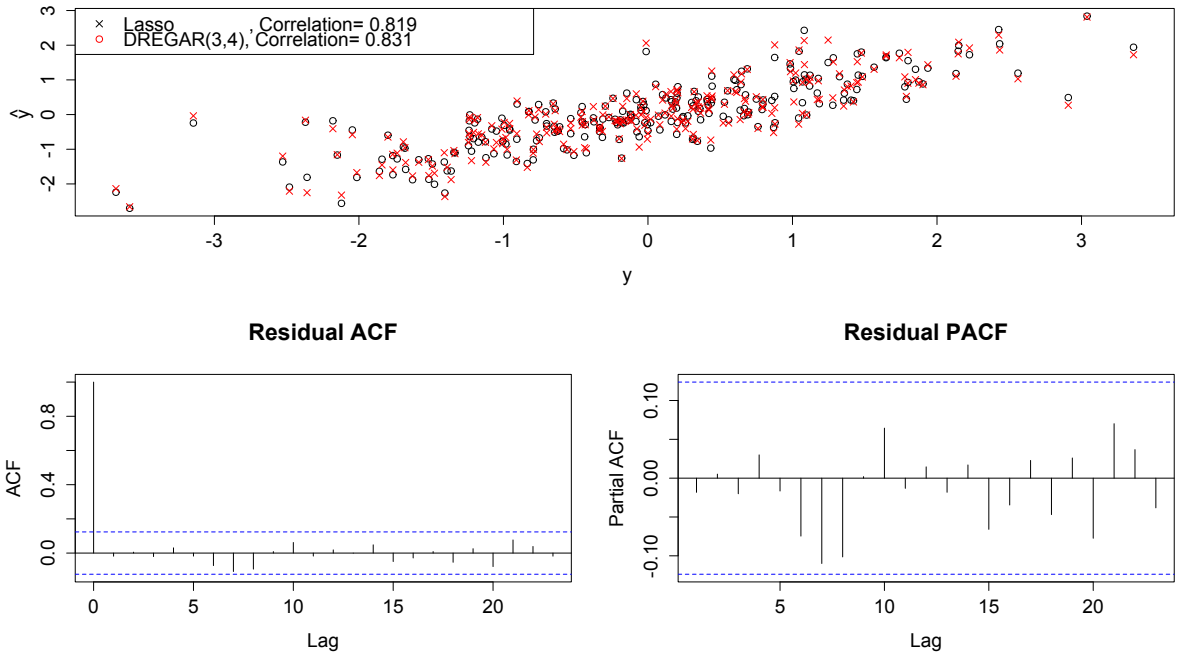

FIGURE 2.8: (Top) Scatter plot of DREGAR(3,4) and Lasso fitted versus observed y, (bottom) Sample ACF and PACF for DREGAR(3,4) residuals.

2.12 Conclusion remarks

This chapter addressed the problem of dynamic regression in the presence of autocorrelated residuals by proposing an extension of the regression model of [Wang et al., 2007]. The extension allows lags of the response. We showed that adding this dynamic term results in a structure more similar to a general ARMAX than REGAR and REGARMA, and with fewer difficulties in parameter estimations than REGARMA. Further, we proposed an l_1 penalized likelihood approach for variable selection for both regression and time-dependent coefficients. Additionally, we discussed the theoretical properties of the regularized estimators of a special case of the model, the one without the autocorrelated residuals, as the general form of the model suffers an OLS bias. We proposed a two-step iterative algorithm for parameter estimation and provided an R package for the implementation and simulation of data from the model. Finally, we show the applicability of the model and comparison with existing approaches by means of two simulation studies as well as two real data applications.

2.12.1 Future study

For future work, we plan to extend the methods presented in this chapter by estimating DREGAR coefficients using penalties that strike a trade-off between l_1 and l_2, such as elastic-net [Zou and Hastie, 2005]. We expect these methods to work well, as the l_2 penalty imposes less weight on small coefficients compared to the l_1 penalty. In addition, the covariance matrix of the regressors in DREGAR is not diagonal, thus violating the assumption of orthogonal predictors. In other words, there are always some correlations amongst the predictors. In this situation lasso algorithm, in particular LARS, chooses one of the regressors and ignores the other correlated ones. A weighted sum of l_1 and l_2 penalties can preserve the collinearity in time-dependent lags and is thus expected to lead to more accurate estimations.

Chapter 3

A differentiable alternative to l_1 lasso penalty

3.1 Main question

Let the underlying model be $y = X\beta + e$ where y, X , β and e are response, covariates, unknown vector of regression coefficients and i.i.d Gaussian error $\sim N(0, \sigma^2)$. In the previous chapter we imposed l_1 penalties on the (log)likelihood to derive a constrained estimation of the parameters. In other words, we imposed the constraint $\sum_i |\beta_i| \leq K$, $K \geq 0$ on the parameters, as there is a one-to-one correspondence between K in this definition and λ in the previous chapter. In this chapter we discuss a differentiable replacement for the l_1 penalty that is capable of producing similar results to lasso as well as Ridge and a range of smooth regularizations.

3.2 Introduction

We start with proposing a smooth alternative to the absolute value function that applies in the l_1 penalized likelihood, lasso. The main idea of a smooth penalty is introduced by some authors [Hebiri et al., 2011, Hebiri, 2008, Fan and Li, 2001, Zou and Hastie, 2005] and consists in adding a differentiable term to the likelihood or l_1 penalized likelihood, resulting in a trade off between maximum sparsity in lasso and increasing the number of selected covariates. It should be stressed that the maximum sparsity of lasso is $\min(r, T)$ where r is the number of covariates and T is total observations. Then, for $r \gg T$ lasso is limited to T variables that is a limitation for a variable selection method. Ridge [Hoerl and Kennard, 1970], elastic-net [Zou and Hastie, 2005], smoothly clipped absolute deviation (SCAD) [Fan and Li, 2001] and smooth lasso [Hebiri et al., 2011] are four well known examples of smooth models. For instance, the amount of smoothness in elastic-net is controlled by the l_2 term; or similarly in smooth-lasso it is controlled by the second norm over the consecutive difference of the coefficients. Amongst these models, ridge is differentiable at zero whereas SCAD, elastic-net and smooth-lasso are not. We should stress that non-smooth penalties can lead to some limitations in certain cases, such as computational efficiency for non-linear models [Schmidt et al., 2007a] or derivation of the degrees of freedom for model selection

criteria, such as the generalised information criterion [Konishi and Kitagawa, 1996], as pointed out by [Abbruzzo et al., 2014]. It is worth noting that in strictly smooth models like ridge sparsity suffers, whereas models like elastic-net and smooth-lasso require a separate tuning parameter to be optimized.

In this chapter, we propose a differentiable penalty that allows choosing from a nearly flat to a very sharp regularization. The use of a differentiable term results in reducing the optimization problem to an ordinary minimization problem that can be implemented by a broad range of algorithms in the literature. In other words, the proposed differentiable penalty removes the dependency of the method to specialized optimization algorithm, e.g. LARS, as well as providing more flexibility than l_1 penalized likelihood by covering l_0, l_1, l_2 and more norms.

3.3 Our proposal: dlasso

Looking at the literature for differentiable alternatives to the absolute value, a number of proposals have been made, such as

$$|x| \approx \sqrt{x^2+\epsilon}, \qquad \epsilon \in \mathbb{R}_+, \tag{3.1}$$

$$\frac{x^2}{\sqrt{x^2+u^2}} \leq |x| \leq \sqrt{x^2+u^2}, \quad u \in \mathbb{R}_+. \tag{3.2}$$

$$|x| \approx |x|_\alpha = \frac{1}{\alpha}[\log(1+e^{-\alpha x}) + \log(1+e^{\alpha x})], \alpha \in \mathbb{R}_+ \tag{3.3}$$

Equation (3.1) is studied in details by [Ramirez et al., 2014]. It is a special case of (3.2) and it is straightforward to show that the length of the interval in (3.2) is always less than u [Nesterov, 2005]. The approximation in (3.3) has been used by [Schmidt et al., 2007b] in a penalized likelihood context. This function is twice differentiable and $|x| = \lim_{\alpha\to\infty} |x|_\alpha$ with the maximum absolute deviance of $\big||x| - |x|_\alpha\big| \leq 2\frac{\log(2)}{\alpha}$, but it does not pass through zero.

In this chapter, we propose the following penalty function

$$f(x,s) = x\left(\frac{2}{\sqrt{\pi}}\int_0^{x/s} e^{-t^2}dt\right), \quad s \in \mathbb{R}_+, \tag{3.4}$$

which we call it **dlasso** for differentiable lasso. The second term in RHS of (3.4) is so called error function, erf(x/s), and can be considered as a probability distribution, see [Olver et al., 2010] for a comprehensive discussion about error function. The accuracy of approximating $|x|$ by this function increases as s tends to zero.

For brevity, in the rest of this chapter we call ϵ, u, α and s in (3.1-3.4) precision values. Figure (3.1) compares these functions for different values of precision. As is evident from this graph, for the same value of the precision, dlasso converges to $|x|$ at the speed faster than the other opponents. Moreover, function in (3.4) passes the origin regardless of the value of precision that is of interest for a loss function. All these properties motivate us to utilize this function in linear regularization problems as a replacement to l_1 norm. To this end, we borrow the definition of a smooth function in [Ramirez et al., 2014] and show

that all proposed functions in (3.1-3.4) are smooth approximations of absolute value function.

Definition 3.1 (Smooth approximation of $|x|$). A function $f : \mathbb{R} \to \mathbb{R}$ is a smooth approximation of $|x|$ if it is differentiable and the following limits hold,

$$\lim_{x\to\pm\infty} \frac{f(x)}{|x|} = 1, \quad \lim_{x\to\pm\infty} \frac{f'(x)}{\text{sign}\,(x)} = 1.$$

Having Definition (3.1) and using simple algebra,one can show that all functions in (3.1-3.4) are smooth.

This chapter is arranged as follows. In Section § 3.4 we prove some key properties of dlasso. Next, the application of dlasso in penalized likelihood is discussed in Section § 3.5. Theoretical properties of a linear model under this new penalty are studied in § 3.6. A discussion about computation complexity of the dlasso penalty as well as proposing a simple approximation for Gaussian CDF are provided in Section § 3.7. Algorithm and selection of the tuning parameter are discussed in Section § 3.8. Finally, simulations and real data illustrations in Section § 3.11 accompany the theoretical results.

3.4 Some key properties of dlasso

The proposed function in equation (3.4) is a special case of a general family of functions as follows,

$$f(x,s,\alpha,\gamma) = x\left[\frac{2}{\sqrt{\pi}}\int_0^{(\frac{x}{s})^\alpha} e^{-t^2}dt\right]^\gamma = x \times \left[\text{ erf }\left((\frac{x}{s})^\alpha\right)\right]^\gamma, \qquad \{\gamma,\alpha\} \in \mathbb{R}, s > 0, x \geq 0,$$

so that $\alpha = \gamma = 1$ results in the function in (3.4), dlasso. We focus entirely on this case and leave the general form of the function for the future studies. Figure (3.2) shows a three dimensional demonstration for the behaviour of dlasso for different values of s. As it is evident from this figure, the proposed function shows a variety of geometric shapes as the values of s changes. For instance, (a), (c) and (d) are corresponded to similar penalties to lasso, ridge and flat (OLS).

We should stress that dlasso is not convex. To show this fact, we find the second derivative of the function,

$$\begin{aligned}
\frac{d^2}{dx}f(x,s) &= \frac{d}{dx}\left(\text{ erf }(\frac{x}{s}) + 2(\frac{x}{s})\phi\left(\frac{x}{s},0,\frac{1}{\sqrt{2}}\right)\right) \\
&= \frac{2}{s}\phi(\frac{x}{s},0,\frac{1}{\sqrt{2}}) + \frac{2}{s}\phi(\frac{x}{s},0,\frac{1}{\sqrt{2}}) - \frac{4}{x}(\frac{x}{s})^3\phi(\frac{x}{s},0,\frac{1}{\sqrt{2}}) \\
&= 4\phi(\frac{x}{s},0,\frac{1}{\sqrt{2}})(\frac{1}{s})\left(1-(\frac{x}{s})^2\right),
\end{aligned}$$

and see that the function is not necessarily positive (or negative). For example the second derivative is positive if $\left(1-(\frac{x}{s})^2\right) > 0$ or equivalently $|x| < s$.

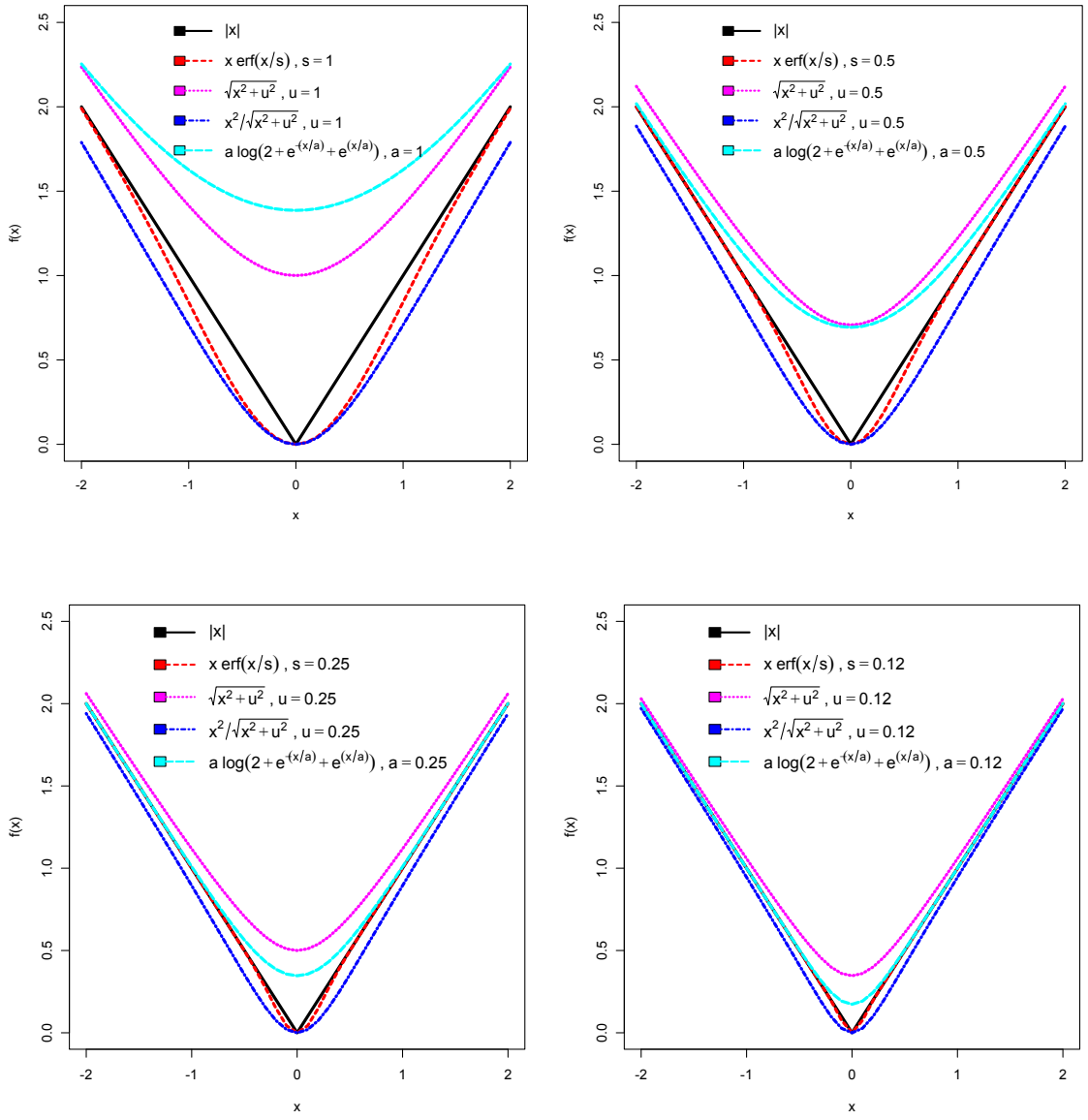

FIGURE 3.1: Comparison of the different alternatives to absolute value function. From up-left to the down-right the precision values decrease at the same rate.

Further, for a fixed value of $s = \frac{2}{\sqrt{\pi}}$ and small values of x dlasso behaves like x^2. This fact, as illustrated in Figure (3.3), shows that the function has a similar behaviour to x^2 for small $(x, s = \frac{2}{\sqrt{\pi}})$. This correspondence motivates the case of the second norm, ridge penalties.
Former can be shown by fixing s and focusing on the small values for x. Thus

$$\frac{2}{\sqrt{\pi}} x \int_0^{x/s} e^{-t^2} dt \overset{x \to 0}{=} \frac{2}{\sqrt{\pi}} \frac{x^2}{s} e^{-(\frac{x}{s})^2} = x^2 \frac{2}{s} \phi(\frac{x}{s}, \mu = 0, \sigma = \frac{1}{\sqrt{2}}),$$

where $\phi(.)$ is the density function for normal distribution. By setting $s = \frac{2}{\sqrt{\pi}}$, the limit becomes $\sqrt{\pi} x^2 \ \phi(\frac{\sqrt{\pi} x}{2}, 0, \frac{1}{\sqrt{2}})$. On the other hand, $\phi(\frac{\sqrt{\pi} x}{2}, 0, \frac{1}{\sqrt{2}}) \overset{x \to 0}{\approx} \frac{1}{\sqrt{\pi}}$ and $\sqrt{\pi} x^2 \phi(\frac{\sqrt{\pi} x}{2}, 0, \frac{1}{\sqrt{2}}) \approx x^2$, that is an approximation for x^2. We should stress that, to reduce computation time of $\frac{2}{\sqrt{\pi}} = 1.128379$, we set

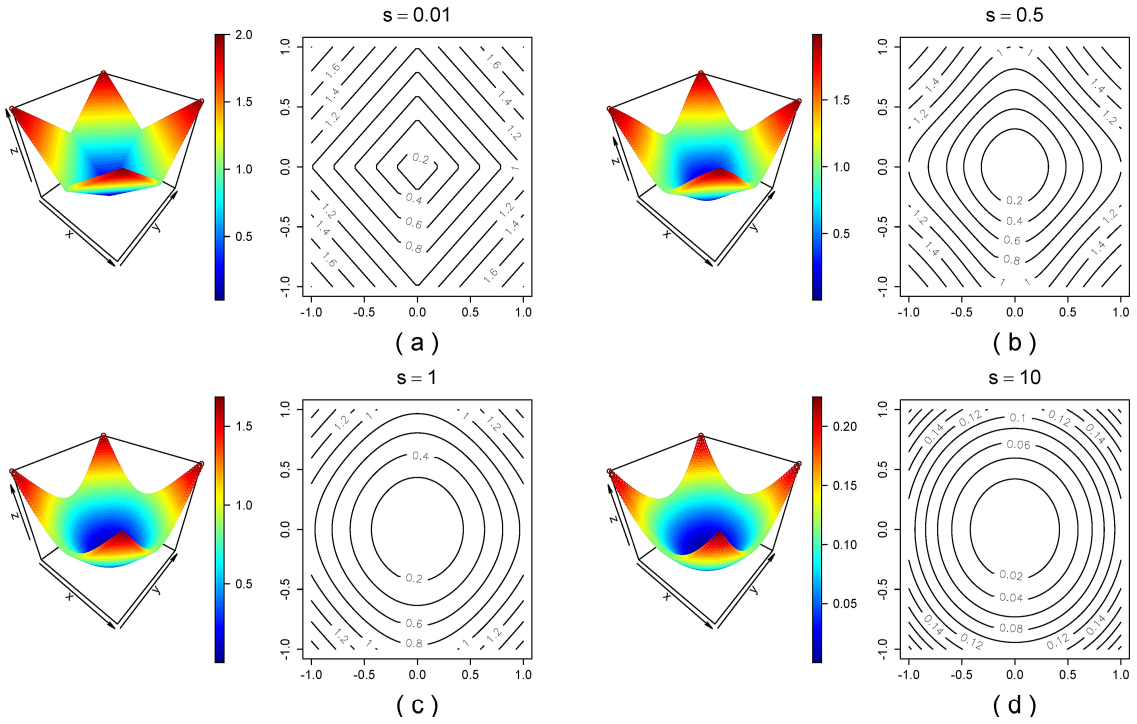

FIGURE 3.2: 3D demonstration and countor plot for the dlasso under $s = 0.01, 0.5, 1, 10$.

$s = 1$ in applications.

Moreover, one can derive an identical form of dlasso using the concept of normal density and distribution. To this end, we use the concept of half normal density $HN(z,0,\sigma) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} e^{\frac{-z^2}{2\sigma^2}}, z \geq 0$ in [Ahsanullah et al., 2014, p.18]. Denote the cumulative distribution function of this density by $CHN(z,0,\sigma)$. Then, $\frac{2}{\sqrt{\pi}} \int_0^{x/s} e^{-t^2} dt$ can be replaced by $CHN(\frac{x}{s}, 0, \frac{1}{\sqrt{2}})$. Further,

$$\begin{aligned} CHN(z,0,\sigma) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \int_0^z e^{\frac{-t^2}{2\sigma^2}} dt &= 2\int_0^z \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-t^2}{2\sigma^2}} dt \\ &= 2\Phi(z,0,\sigma) - 1, \end{aligned} \tag{3.5}$$

where $\Phi(z,0,\sigma)$ is the cumulative normal distribution with mean of zero and variance equals to σ^2. Consequently, $\text{erf}\,(\frac{x}{s}) = \frac{2}{\sqrt{\pi}} \int_0^{x/s} e^{-t^2} dt = CHN(\frac{x}{s}, 0, \frac{1}{\sqrt{2}}) = 2\Phi(\frac{x}{s}, 0, \frac{1}{\sqrt{2}}) - 1$.

Finally, we show that the maximum deviance of the function from $|x|$ decreases exponentially with s. Obviously for $x = 0$ the difference is zero. Then we prove $\big||x| - x(2\Phi(\frac{x}{s}, 0, \frac{1}{\sqrt{2}}) - 1)\big| \leq 2s\phi(\frac{x}{s}, 0, \frac{1}{\sqrt{2}})$ for

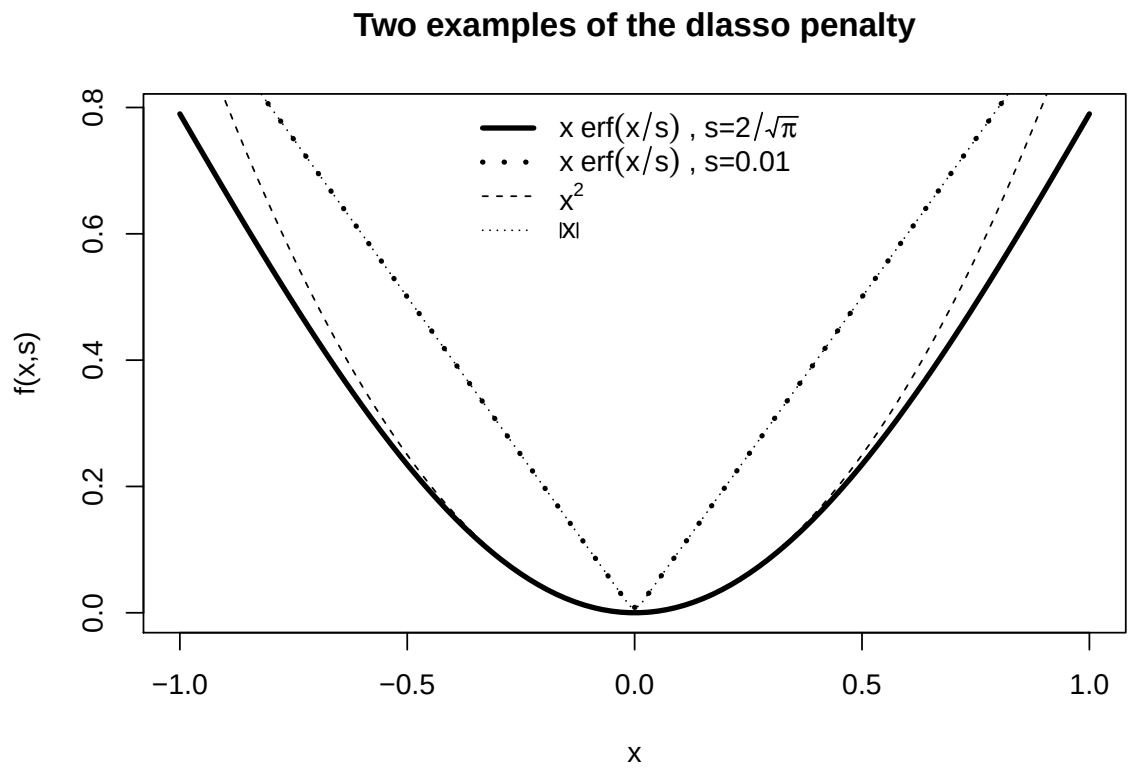

FIGURE 3.3: Comparison of x^2 and $|x|$ with limit behaviour of $x\,(2\Phi(x/s, 0, 1/\sqrt{2}) - 1)$ for $s = 0.01$ and $s = \frac{2}{\sqrt{\pi}}$ over the small values for x.

all $x \neq 0$ and $s > 0$. To this end, we have

$$\begin{aligned}\Phi^c(\frac{x}{s}, 0, \frac{1}{\sqrt{2}}) = 1 - \Phi(\frac{x}{s}, 0, \frac{1}{\sqrt{2}}) &= \int_{\frac{x}{s}}^{\infty} \frac{1}{\sqrt{\pi}} e^{-t^2} dt \\ [\text{ from } (\frac{t}{(\frac{x}{s})} > 1)\ ] &< \int_{\frac{x}{s}}^{\infty} (\frac{t}{(\frac{x}{s})}) \frac{1}{\sqrt{\pi}} e^{-t^2} dt \\ &= (\frac{1}{2(\frac{x}{s})}) \frac{1}{\sqrt{\pi}} e^{-(\frac{x}{s})^2}.\end{aligned}$$

Using the equation $g(t) = \Phi^c(\frac{x}{s}, 0, \frac{1}{\sqrt{2}}) - \frac{1}{\sqrt{\pi}} \frac{(\frac{x}{s})}{1+2(\frac{x}{s})^2} e^{-(\frac{x}{s})^2}$, one can prove that $\Phi^c(\frac{x}{s}, 0, \frac{1}{\sqrt{2}}) > \frac{1}{\sqrt{\pi}} \frac{(\frac{x}{s})}{1+2(\frac{x}{s})^2} e^{-(\frac{x}{s})^2}$ [Chang et al., 2011]. Then,

$$\frac{1}{\sqrt{\pi}} \frac{(\frac{x}{s})}{1+2(\frac{x}{s})^2} e^{-(\frac{x}{s})^2} < \Phi^c(\frac{x}{s}, 0, \frac{1}{\sqrt{2}}) < (\frac{1}{2(\frac{x}{s})}) \frac{1}{\sqrt{\pi}} e^{-(\frac{x}{s})^2}.$$

Referring to [Abramowitz and Stegun, 2012], a tighter bound is given by

$$\frac{e^{-(\frac{x}{s})^2}}{(\frac{x}{s}) + \sqrt{(\frac{x}{s})^2 + 2}} < \int_{\frac{x}{s}}^{\infty} e^{-t^2} dt \leq \frac{e^{-(\frac{x}{s})^2}}{(\frac{x}{s}) + \sqrt{(\frac{x}{s})^2 + \frac{4}{\pi}}}$$

$$\text{Or: } \frac{1}{\sqrt{\pi}} \frac{e^{-(\frac{x}{s})^2}}{(\frac{x}{s}) + \sqrt{(\frac{x}{s})^2 + 2}} < \Phi^c(\frac{x}{s}, 0, \frac{1}{\sqrt{2}}) \leq \frac{1}{\sqrt{\pi}} \frac{e^{-(\frac{x}{s})^2}}{(\frac{x}{s}) + \sqrt{(\frac{x}{s})^2 + \frac{4}{\pi}}}.$$

Using the inequalities above we get,

$$\begin{aligned} x>0 \rightarrow \left|x - x(2\Phi(\frac{x}{s},0,\frac{1}{\sqrt{2}})-1)\right| &= \left|2x(1-\Phi(\frac{x}{s},0,\frac{1}{\sqrt{2}}))\right| \\ &= 2x(1-\Phi(\frac{x}{s},0,\frac{1}{\sqrt{2}})) \\ &\leq \frac{2x}{\sqrt{\pi}}\frac{e^{-(\frac{x}{s})^2}}{(\frac{x}{s})+\sqrt{(\frac{x}{s})^2+\frac{4}{\pi}}} \\ &= \frac{2s}{\sqrt{\pi}}e^{-(\frac{x}{s})^2}\frac{1}{1+\sqrt{1+\frac{4s^2}{\pi x^2}}} \\ &= \phi(\frac{x}{s},0,\frac{1}{\sqrt{2}})\frac{2s}{1+\sqrt{1+\frac{4s^2}{\pi x^2}}} \\ &\leq \frac{2s}{\sqrt{\pi}}e^{-(\frac{x}{s})^2} = 2s\phi(\frac{x}{s},0,\frac{1}{\sqrt{2}}). \end{aligned}$$

Following a similar approach for $x<0$ leads to the same result. Consequently,

$$\begin{aligned} \left||x| - x(2\Phi(\frac{x}{s},0,\frac{1}{\sqrt{2}})-1)\right| &\leq 2\phi(\frac{x}{s},0,\frac{1}{\sqrt{2}})\frac{s}{1+\sqrt{1+\frac{4s^2}{\pi x^2}}} \\ &\leq 2s\phi(\frac{x}{s},0,\frac{1}{\sqrt{2}}). \end{aligned}$$

The final equation tends to zero as $s \rightarrow 0$ at an exponential speed.

3.5 Regularized regression based on dlasso

Let $X=(x_1,x_2,x_3,\ldots,x_r)$ be a known design matrix where $x_i, i=1,2,\ldots,r$ are independent column vectors of length T and the corresponding ratio r/T can be much greater than 1. We assume linearity for the link function $y=X\beta+e$ where $\beta=(\beta_1,\beta_2,\ldots,\beta_r)$ are regression coefficients and $e=\{e_i, i=1,2,\ldots,T\} \overset{i.i.d}{\sim} N(0,\sigma^2)$ are independent fixed-level noise. Then ML estimation of the parameters is equivalent to minimizing $\frac{1}{2}(y-X\beta)'(y-X\beta)$ with respect to β.

As discussed in Chapter § 1, given $T<r$, a standard approach to cope with high dimensionality is by imposing a penalty term on the likelihood that results in a constrained minimization problem $\frac{1}{2}(y-X\beta)'(y-X\beta)+\lambda\sum_{i=1}^{r}|\beta_i|$ where λ is the tuning parameter and controls the amount of sparsity in the solutions. Replacing the absolute value function with dlasso leads to

$$Q(\beta,s)=\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)+\lambda\frac{2}{\sqrt{\pi}}\sum_{i=1}^{r}\beta_i\int_0^{\beta_i/s}e^{-t^2}dt, \quad s>0,\lambda\geq 0.$$

Without loss of generality we assume that $\sigma^2=1$ in the entire chapter. Rewriting this equation using (3.5) leads to

$$Q(\beta,s)=\frac{1}{2}(y-X\beta)'(y-X\beta)+\lambda\sum_{j=1}^{r}\beta_i\left(2\Phi(\frac{\beta_i}{s},0,\frac{1}{\sqrt{2}})-1\right) \quad s>0,\lambda\geq 0. \tag{3.6}$$

Consequently, the problem reduces to minimizing (3.6) with respect to β. Without loss of generality we define $\lambda^* = 2\lambda$ and remove the multiplier from the first equation in the RHS of (3.6). Then,

$$L(\beta) = (y - X\beta)'(y - X\beta) + \lambda^* \sum_{j=1}^{r} \beta_i \left(2\Phi(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \right) \quad s > 0, \lambda^* \geq 0. \tag{3.7}$$

We should stress that the dlasso penalty together with the tuning parameter λ can be seen as a function of two tuning parameters,

$$P(\lambda, s) = \lambda\beta \left(2\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \right)$$

where there is some redundancy between those two tuning parameters, i.e.

$$\begin{cases} P(\lambda, s) \to 0 & \text{if} \quad \lambda \to 0 \\ P(\lambda, s) \to 0 & \text{if} \quad s \to \infty. \end{cases}$$

We should notice that fixing λ and sliding $s \to \infty$ compared to fixing s and sliding $\lambda \to 0$ has different effect on the shape of the final function, so that for the first case $s \to \infty$ would flatten the curvature of the function at zero whereas $\lambda \to 0$ flatten the gradient of the function, given s is close enough to zero. In this chapter we do not follow this idea and assume that s is a fixed quantity that controls the sharpness of the penalty function at zero.

It should be noticed that (3.6) is differentiable with respect to β. In particular, the first derivative is,

$$\frac{\partial L}{\partial \beta} = -X'(y - X\beta) + \lambda^* \left(I_{r\times r}(2\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) - \mathbf{1}) + 2 \left[\phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) \right]_{r\times r} \frac{\beta}{s} \right),$$

where $\left[\phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) \right]_{r\times r}$ is a diagonal matrix consists of derivatives. Given X is normalized so that $X'X/T \to I_{r\times r}$,

$$\frac{\partial L}{\partial \beta} = -X'y + T\beta + \lambda^* \left(2I_{r\times r}\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) - \mathbf{1} + 2 \left[\phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) \right]'_{r\times r} \frac{\beta}{s} \right),$$

which exists for any β.

For an illustrative example, take the linear function $y = x\beta$ and assume that $x = 1$ as well as the following constrained minimization problem,

$$L(\beta) = (y - \beta)^2 + \lambda^*\beta \left(2\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \right), \qquad s > 0, \lambda^* \geq 0.$$

Finding the derivative with respect to β results in

$$\frac{d}{d\beta} L(\beta) = \lambda^* \left(2\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) - 1 + 2(\frac{\beta}{s})\phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) \right) - (y - \beta) = 0,$$

for which there is no close form solution. Thus, we numerically find the root of this equation for different values of s and show the corresponding effect on the estimation. Results for $s \in \{0.01, 0.5, 1, 20\}$, $y \in (-1.5, 1.5)$ and $\lambda^* = 1$ are shown in Figure (3.4). As it is evident from this plot, (a), (c) and (d) show similar regularization to lasso, ridge and non-penalized linear regression respectively.

3.6 Theoretical properties of dlasso estimator

In this section we concentrate on the theoretical properties of the penalized likelihood under the dlasso penalty. For this purpose, we show that the estimators asymptotically reach the similar bias to lasso, given $s \to 0$. To this end we follow a similar approach to [Knight and Fu, 2000] also introduced in Section § 1.3.1 and define a loss function so that it reaches the minimum at the estimators, $\hat{\beta}$.

Theorem 3.1 (Similarity to lasso). For any $u \in \mathbb{R}^r$, $\lambda^* \geq 0$ and $s > 0$ define,

$$k(u, s) = L(\beta + u) - L(\beta),$$

where $L(\beta)$ is the penalized (log)likelihood in equation (3.7). Then,

$$\lim_{s \to 0} k(u, s) = u'X'Xu - 2u'N\Big(0, \sigma^2(X'X)\Big) + \lambda^* \sum_{i=1}^{r} \Big(|u_i| I(\beta_i = 0) + u_i \operatorname{sign}(\beta_i + u_i) \Big),$$

where N denotes a normally distributed random variable.

Proof. Recall from equation (3.7)

$$L(\beta) = (y - X\beta)'(y - X\beta) + \lambda^* \sum_{i=1}^{r} \beta_i \Big(2\Phi(\frac{\beta_i}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \Big) \quad s > 0, \lambda^* \geq 0.$$

Then,

$$\begin{aligned}
\lim_{s \to 0} k(u) &= \lim_{s \to 0} L(\beta + u) - \lim_{s \to 0} L(\beta) \\
&= (e - Xu)'(e - Xu) - e'e + \lim_{s \to 0} \Big\{ 2\lambda^* \sum_{i=1}^{r} \beta_i \int_{\frac{\beta_i}{s}}^{\frac{\beta_i + u_i}{s}} \frac{1}{\sqrt{\pi}} e^{-t^2} dt \\
&\qquad + \lambda^* \sum_{i=1}^{r} u_i \Big(2\Phi(\frac{\beta_i + u_i}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \Big) \Big\} \\
&= u'X'Xu - 2u'N\Big(0, \sigma^2(X'X)\Big) \\
&\qquad + \lim_{s \to 0} \lambda^* \sum_{i=1}^{r} \Big(2\beta_i \frac{u_i}{s} \phi(\frac{u_i}{s}, 0, \frac{1}{\sqrt{2}}) + \lim_{s \to 0} \Big\{ u_i \Big(2\Phi(\frac{\beta_i + u_i}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \Big) \Big\} \Big) \\
&= u'X'Xu - 2u'N\Big(0, \sigma^2(X'X)\Big) \\
&\qquad + \lim_{s \to \infty} \lambda^* \sum_{i=1}^{r} \begin{cases} u_i \Big(2\Phi(\frac{u_i}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \Big) & \beta_i = 0 \\ 2\beta_i \frac{u_i}{s} \phi(\frac{u_i}{s}, 0, \frac{1}{\sqrt{2}}) + u_i \Big(2\Phi(\frac{\beta_i + u_i}{s}, 0, \frac{1}{\sqrt{2}}) - 1 \Big) & \beta_i \neq 0 \end{cases} \\
&= u'X'Xu - 2u'N\Big(0, \sigma^2(X'X)\Big)
\end{aligned}$$

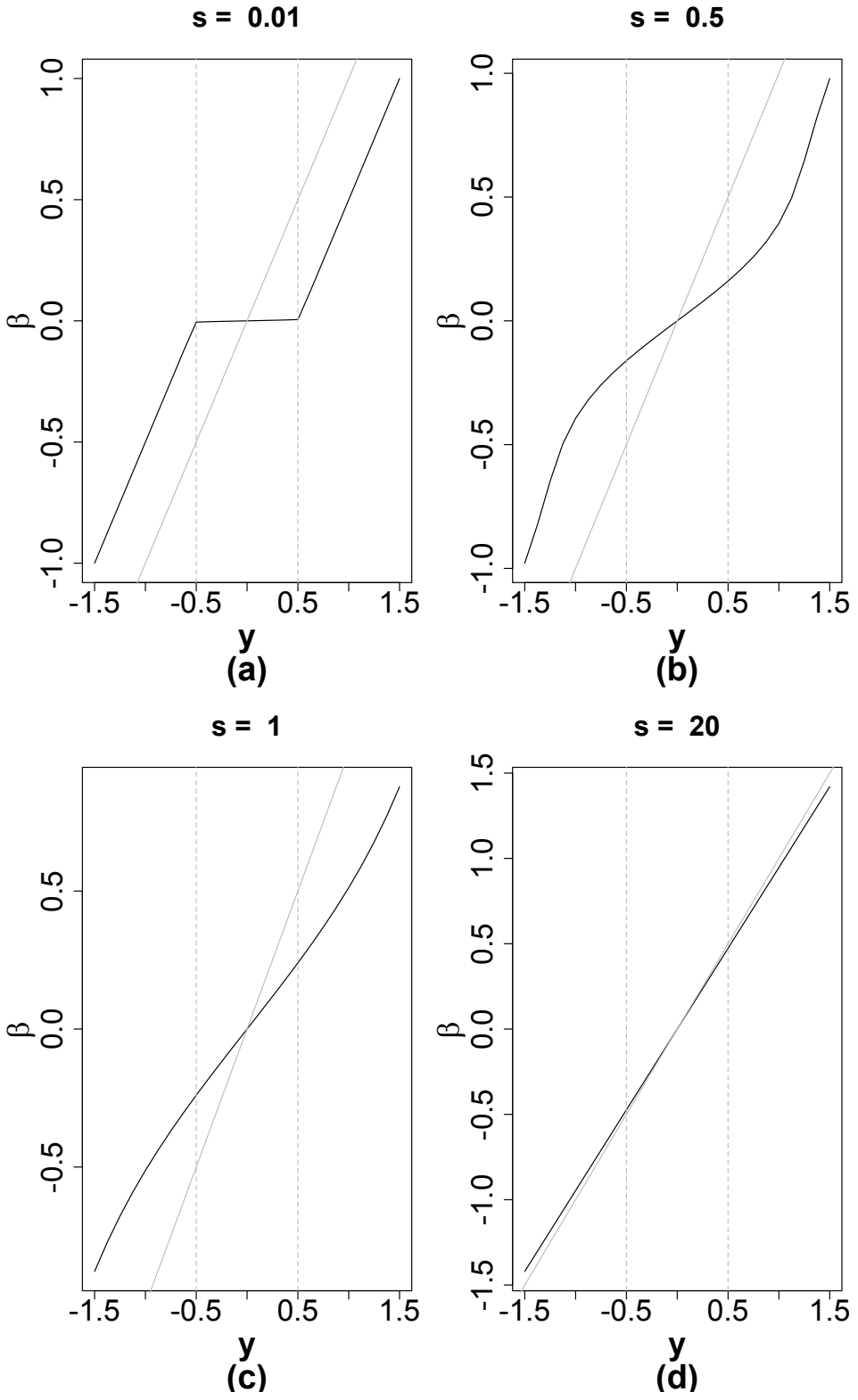

FIGURE 3.4: The estimation of β in linear function $y = x\beta$ for $x = 1$ results from imposing $\lambda^*\beta\big(2\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) - 1\big)$ constrain on the minimization problem $\min_\beta(y-\beta)^2$ under different values of s as well as fixed $\lambda^* = 1$. The gray solid line shows the $y = \beta$ line and dotted vertical lines show $\pm\frac{\lambda^*}{2}$.

$$+\lambda^* \sum_{i=1}^{r} \begin{cases} |u_i| & \beta_i = 0 \\ u_i & \beta_i + u_i > 0 \\ -u_i & \beta_i + u_i < 0 \end{cases},$$

and

$$k(u) = u'X'Xu - 2u'N\Big(0, \sigma^2(X'X)\Big) + \lambda^* \sum_{i=1}^{r} \Big(|u_i| I(\beta_i = 0) + u_i \operatorname{sign}(\beta_i + u_i)\Big).$$

Proof is completed. □

Theorem (3.1) shows that the limit distribution of estimator under the dlasso penalty is similar to lasso, provided s is close enough to zero. That is, the penalization is capable of producing sparse estimations. This theorem guarantees that the estimations are similar to lasso (provided s is close enough to zero) but does not provide any optimal value for s to ensure this convergence. Then, in the next theorem we show that the minimum speed of s that guarantees the convergence of estimations to lasso is $T^{-(1/2+\epsilon)}$ for any $\epsilon > 0$.

Theorem 3.2 (Optimal speed of s for reproducing lasso). Let β be a sparse set of coefficients, $u \in \mathbb{R}^r$, $s_T = s/(T^{1/2+\epsilon}) \to 0$, $\epsilon > 0$, $\lambda_T^*/\sqrt{T} \to \lambda_\circ \geq 0$, $\max_{1 \leq i \leq r} x_i x_i' < \infty$ and $X'X/T \xrightarrow{p} \Sigma$ where Σ is non-singular. Then, (a) $\sqrt{T}(\hat{\beta}_T - \beta) \xrightarrow{d} \arg\min_u k(u)$ where,

$$k(u) = -2u'N(O, \sigma^2\Sigma) + u'\Sigma u + \lambda_\circ \sum_{i=1}^{r} \{u_i \operatorname{sign}(\beta_i) I(\beta_i \neq 0) + |u_i| I(\beta_i = 0)\},$$

(b) this configuration guarantees obtaining the sparse estimation of the parameters.

Proof. We define,

$$k_T(u) = L(\beta + \frac{u}{\sqrt{T}}) - L(\beta).$$

Then

$$\begin{aligned} k_T(u) &= (e - X\frac{u}{\sqrt{T}})'(e - X\frac{u}{\sqrt{T}}) - e'e + 2\lambda_T^* \sum_{i=1}^{r} \beta_i \int_{\frac{\beta_i}{s}}^{\frac{\beta_i + \frac{u_i}{\sqrt{T}}}{s}} \frac{1}{\sqrt{\pi}} e^{-t^2} dt \\ &\quad + \lambda_T^* \sum_{i=1}^{r} \frac{u_i}{\sqrt{T}} \Big(2\Phi(\frac{\beta_i + \frac{u_i}{\sqrt{T}}}{s}, 0, \frac{1}{\sqrt{2}}) - 1\Big) \\ &\overset{T \to \infty}{=} u'\Sigma u - 2u'N(0, \sigma^2\Sigma) + \lim_{T \to \infty} 2\lambda_T^* \sum_{i=1}^{r} \beta_i \int_{\frac{\beta_i}{s}}^{\frac{\beta_i + \frac{u_i}{\sqrt{T}}}{s}} \frac{1}{\sqrt{\pi}} e^{-t^2} dt \\ &\quad + \lim_{T \to \infty} \lambda_T^* \sum_{i=1}^{r} \frac{u_i}{\sqrt{T}} \Big(2\Phi(\frac{\beta_i + \frac{u_i}{\sqrt{T}}}{s}, 0, \frac{1}{\sqrt{2}}) - 1\Big) \\ &= u'\Sigma u - 2u'N(0, \sigma^2\Sigma) + 2\lambda_\circ \sum_{i=1}^{r} \beta_i \frac{u_i}{s\sqrt{\pi}} \lim_{T \to \infty} e^{-(\frac{\beta_i + \frac{u_i}{\sqrt{T}}}{s})^2} \\ &\quad + \lambda_\circ \sum_{i=1}^{r} u_i \Big(\lim_{T \to \infty} 2\Phi(\frac{\beta_i + \frac{u_i}{\sqrt{T}}}{s}, 0, \frac{1}{\sqrt{2}}) - 1\Big) \end{aligned}$$

$$
\begin{aligned}
&= u'\Sigma u - 2u'N(0,\sigma^2\Sigma) \\
&+ \lim_{T\to 0}\begin{cases} \lambda_\circ \sum_{i=m+1}^{r} u_i\left(2\Phi(\frac{\frac{u_i}{\sqrt{T}}}{s},0,\frac{1}{\sqrt{2}})-1\right) & \beta_i \in S_\circ \\ 2\lambda_\circ \sum_{i=1}^{m}\left(\beta_i \frac{u_i}{s\sqrt{\pi}} e^{-(\frac{\beta_i+\frac{u_i}{\sqrt{T}}}{s})^2} + u_i\left(2\Phi(\frac{\beta_i+\frac{u_i}{\sqrt{T}}}{s},0,\frac{1}{\sqrt{2}})-1\right)\right) & \beta_i \in S_\circ^c \end{cases},
\end{aligned}
$$

where $S_\circ$ and $S_\circ^c$ are sets of zero and non-zero coefficients respectively. This completes the first part of the theorem. For the second part of theorem, we follow a similar approach to Section § 1.3.3 and assume that $m-$ first coefficients are non-zero whereas the rest $r-m$ coefficients are zero and define the following splits,

$$
\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad N = \begin{bmatrix} N_1 \\ N_2 \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix},
$$

where $\Sigma_{11}, \Sigma_{22}, \Sigma_{21} = \Sigma_{12}'$ are $m\times m$, $(r-m)\times(r-m)$ and $(r-m)\times m$ block matrices corresponding to non-zero and zero coefficients. Further, u_1 and N_1 are vectors of the length m corresponded to non-zero coefficients, respectively. Finding the derivative of $k_T(u)$ with respect to u leads to

$$
\begin{aligned}
\frac{\partial}{\partial u}k(u) &= \frac{\partial}{\partial u}\left(u'\Sigma u - 2u'N + 2\lambda_\circ \sum_{i=1}^{r}\beta_i \frac{u_i}{s\sqrt{\pi}} e^{-(\frac{\beta_i}{s})^2} dt \lambda_\circ \sum_{i=1}^{r} u_i\left(2\Phi(\frac{\beta_i}{s},0,\frac{1}{\sqrt{2}})-1\right)\right) \\
&= 2u'\Sigma - 2N(0,\sigma^2\Sigma) \\
&+ \frac{\partial}{\partial u}\begin{cases} \lambda_\circ \sum_{i=m+1}^{r} u_i\left(2\Phi(\frac{\frac{u_i}{\sqrt{T}}}{s},0,\frac{1}{\sqrt{2}})-1\right) & \beta \in S_\circ \\ 2\lambda_\circ \sum_{i=1}^{m}\beta_i \frac{u_i}{s\sqrt{\pi}} e^{-(\frac{\beta_i+\frac{u_i}{\sqrt{T}}}{s})^2} + \lambda_\circ \sum_{i=1}^{m} u_i\left(2\Phi(\frac{\beta_i+\frac{u_i}{\sqrt{T}}}{s},0,\frac{1}{\sqrt{2}})-1\right) & \beta \in S_\circ^c \end{cases}.
\end{aligned}
$$

Using the assumption that $s\to 0$ at the speed faster than $\sqrt{T}$, the above equation results in a sparse estimation of the parameter because,

$$
\begin{aligned}
\frac{\partial}{\partial u}k(u) &= 2u'\Sigma - 2N(0,\sigma^2\Sigma) \\
&+ \frac{\partial}{\partial u}\begin{cases} \lambda_\circ \sum_{i=m+1}^{r} u_i\left(2\Phi(\frac{\frac{u_i}{\sqrt{T}}}{s},0,\frac{1}{\sqrt{2}})-1\right) & \beta \in S_\circ \\ 2\lambda_\circ \sum_{i=1}^{m}\beta_i \frac{u_i}{s\sqrt{\pi}} e^{-(\frac{\beta_i+\frac{u_i}{\sqrt{T}}}{s})^2} + \lambda_\circ \sum_{i=1}^{m} u_i\left(2\Phi(\frac{\beta_i+\frac{u_i}{\sqrt{T}}}{s},0,\frac{1}{\sqrt{2}})-1\right) & \beta \in S_\circ^c \end{cases} \\
&= 2u'\Sigma - 2N(0,\sigma^2\Sigma) + \frac{\partial}{\partial u}\begin{cases} \lambda_\circ \sum_{i=m+1}^{r}|u_i| & \beta = 0 \\ \lambda_\circ \sum_{i=1}^{m} u_i & \beta > 0 \\ \lambda_\circ \sum_{i=1}^{m} -u_i & \beta < 0 \end{cases},
\end{aligned}
$$

and

$$
u_1'\Sigma_{11} - N_1 + \frac{\lambda_\circ}{2}\operatorname{sign}(\beta_{1:m}) = 0, \quad u_1'\Sigma_{12} - N_2 \pm \frac{\lambda_\circ}{2}1 = 0,
$$

that is similar to Theorem (1.1) (See further [Knight and Fu, 2000, Eq. 9 and 10]). Then solving the equations with respect to u_1 results in,

$$
-\frac{\lambda_\circ}{2}\mathbf{1} \le \Sigma_{21}\Sigma_{11}^{-1}(N_1 - \frac{\lambda_\circ}{2}\operatorname{sign}(\beta_{1:m})) - N_2 \le \frac{\lambda_\circ}{2}\mathbf{1},
$$

where **1** is a vector of 1's.
Proof is completed. □

For the special case of $\beta_1 = \beta_2 = \ldots = \beta_r = 0$ the final inequality results in $-\frac{\lambda_\circ}{2}\mathbf{1} \leq N_2 \leq \frac{\lambda_\circ}{2}\mathbf{1}$ that is non-zero probability at 0 for zero coefficients.

In the proof of Theorem (3.2), we assumed that $s_T\sqrt{T} \to 0$. This gives a way for determining the optimal value of s to get similar results to lasso. In other words, if one chooses any s less than $1/\sqrt{T}$ then Theorem (3.2) guarantees the similarity of results to the lasso in terms of the distribution of the estimations, provided that there are enough observations.

On the other hand, in the following corollary we discuss the asymptotic property of penalized likelihood under a vector of precisions s^* so that $s_i^*\sqrt{T} \to q_i \in [0,\infty), i = 1,2,\ldots,r$. This case can be considered as an adaptive form of penalization.

Corollary 3.1 (Arbitrary value for s). Under similar conditions to Theorem (3.2) but given a vector of precisions $s_{T,i} = s_i^*/\sqrt{T} \to q_i \geq 0, i = 1,2,\ldots,r$, and,

$$\frac{u_2'}{s^*}\phi(\frac{u_2}{s^*}, 0, \frac{1}{\sqrt{2}}) \overset{T\to\infty}{=} 0$$

and u_2 represents near-zero estimations, then, minimizing (3.7) results in less sparse estimation of the parameters than lasso.

Proof. Similar to Theorem (3.2) we define $k_T(u) = L(\beta + u/\sqrt{T}) - L(\beta)$. Then,

$$\begin{aligned}
k_T(u) &= (e - X\frac{u}{\sqrt{T}})'(e - X\frac{u}{\sqrt{T}}) - e'e + 2\lambda_n^* \sum_{i=1}^{r} \beta_i \int_{\frac{\beta_i}{\frac{s_i^*}{\sqrt{T}}}}^{\frac{\beta_i + \frac{u_i}{\sqrt{T}}}{\frac{s_i^*}{\sqrt{T}}}} \frac{1}{\sqrt{\pi}} e^{-t^2} dt \\
&\quad + \lambda_T^* \sum_{i=1}^{r} \frac{u_i}{\sqrt{T}} \left(2\Phi(\frac{\beta_i + \frac{u_i}{\sqrt{T}}}{\frac{s_i^*}{\sqrt{T}}}, 0, \frac{1}{\sqrt{2}}) - 1 \right) \\
&= u'\Sigma u - 2u'N(0,\sigma^2\Sigma) + 2\lambda_T^* \sum_{i=1}^{r} \beta_i \int_{\frac{\beta_i}{\frac{s_i^*}{\sqrt{T}}}}^{\frac{\beta_i + \frac{u_i}{\sqrt{T}}}{\frac{s_i^*}{\sqrt{T}}}} \frac{1}{\sqrt{\pi}} e^{-t^2} dt \\
&\quad + \lambda_T^* \sum_{i=1}^{r} \frac{u_i}{\sqrt{T}} \left(2\Phi(\frac{\beta_i + \frac{u_i}{\sqrt{T}}}{\frac{s_i^*}{\sqrt{T}}}, 0, \frac{1}{\sqrt{2}}) - 1 \right) \\
&= u'\Sigma u - 2u'N(0,\sigma^2\Sigma) + 2\lambda_\circ \sum_{i=1}^{r} \beta_i \frac{u_i}{\frac{s_i^*}{\sqrt{T}}\sqrt{\pi}} e^{-(\frac{\beta_i + \frac{u_i}{\sqrt{T}}}{\frac{s_i^*}{\sqrt{T}}})^2} \\
&\quad + \lambda_\circ \sum_{i=1}^{r} u_i \left(2\Phi(\frac{\beta_i + \frac{u_i}{\sqrt{T}}}{\frac{s_i^*}{\sqrt{T}}}, 0, \frac{1}{\sqrt{2}}) - 1 \right)
\end{aligned}$$

$$= u'\Sigma u - 2u'N(0,\sigma^2\Sigma) + \begin{cases} \lambda_\circ \sum_{i=m+1}^{r} u_i \left(2\Phi(\frac{u_i}{s_i^*},0,\frac{1}{\sqrt{2}}) - 1\right) & \beta = 0 \\ \lambda_\circ \sum_{i=1}^{m} u_i & \beta > 0 \\ \lambda_\circ \sum_{i=1}^{m} -u_i & \beta < 0 \end{cases}.$$

Then, $u_1' = \left(\frac{\lambda_\circ}{2}\,\mathrm{sign}\,(\beta) + N_1\right)\Sigma_{11}^{-1}$ and

$$u_1'\Sigma_{12} - N_2 = \lambda_\circ \frac{\partial}{\partial u_2} u_2' \left(\Phi(\frac{u_2'}{s_i^*},0,\frac{1}{\sqrt{2}}) - 0.5\right).$$

In addition,

$$\frac{\partial}{\partial u_2} u_2' \left(\Phi(\frac{u_2'}{s_i^*},0,\frac{1}{\sqrt{2}}) - 0.5\right) = \left(\Phi(\frac{u_2'}{s_i^*},0,\frac{1}{\sqrt{2}}) - 0.5\right) + \frac{u_2'}{s_i^*}\phi(\frac{u_2'}{s_i^*},0,\frac{1}{\sqrt{2}}),$$

where the first term in the RHS is always bounded to $(-0.5, 0.5)$. Then,

$$|u_1'\Sigma_{12} - N_2 - \lambda_\circ \frac{u_2'}{s_i^*}\phi(\frac{u_2'}{s_i^*},0,\frac{1}{\sqrt{2}})| < \frac{\lambda_\circ}{2}\mathbf{1},$$

that again can produce sparse estimations.
Proof is completed. □

3.7 Computational complexity

The only challenging term in the proposed penalty is the error function, $\mathrm{erf}\,(x) \propto \int_0^x e^{-t^2} dt$, as there is no closed form for it. Then, alternative techniques such as Taylor approximation provide high precision approximation of this function. For instance, two examples are

$$\begin{aligned} \mathrm{erf}\,(x) &= \int_0^x e^{-u^2} du \\ &= \frac{2x}{\sqrt{\pi}} \sum_{j=0}^{\infty} \frac{(-1)^j x^{2j}}{j!(2j+1)} \qquad (3.8) \\ \text{Or } &= \frac{2xe^{-x^2}}{\sqrt{\pi}} \sum_{j=0}^{\infty} \frac{2^j x^{2j}}{1\cdot 3\cdot 5\cdots(2j+1)}. \qquad (3.9) \end{aligned}$$

For small $|x|$, series in (3.8) is slightly faster than series (3.9) because there is no need to compute an exponential. However, series (3.9) is preferable to (3.8) for moderate $|x|$ because it involves no cancellation. For large $|x|$, neither series are satisfactory and in this case it is preferable to use the asymptotic expansion for complementary error function $\mathrm{erf}^c(x) = 1 - \mathrm{erf}\,(x)$,

$$\mathrm{erf}^c(x) \approx \frac{e^{-x^2}}{x\sqrt{\pi}} \sum_{j=0}^{k} (-1)^j \frac{(2j)!}{j!} (2x)^{-2j}. \qquad (3.10)$$

Beside these complicated algorithms, based on Taylor approximations, there are fast algorithms that approximate the error function with quite high precision, see for example [Vazquez Leal et al., 2012], [Olver

et al., 2010], [Chevillard and Revol, 2008] [Cody, 1969], [Press, 1992], [Lee, 1992], [Cody, 1990], [Borjesson et al., 1979] for a range of fast methods. To increase the speed of the algorithm, we focus on fast algorithms such as,

$$\text{erf}\,(x) \approx \tanh\left(\frac{39x}{2\sqrt{\pi}} - \frac{111}{2}\arctan(\frac{35x}{111\sqrt{\pi}})\right),$$

that provides reliable results. For fast and reliable results a combination of methods in equation (3.8), (3.9), (3.10) and fast methods can provide fast and precise enough results. In this chapter we propose the following approximation for standard normal distribution that is fast enough for our purpose,

$$\begin{aligned}\Phi(x) = &\left(\frac{1}{1.9\sqrt{\pi}}\left(\sin(\frac{\pi x}{10}) + \sin(x)\right) + .5\right) I_{(|x|\leq 1.513859)} + \\ &\left(1 - e^{-1.78} + \frac{x}{e^{x+10}}\right) I_{(x>1.513859)} + \\ &\left(e^{-1.78|x|} - \frac{|x|}{e^{|x|+10}}\right) I_{(x<-1.513859)},\end{aligned}$$

where Φ denotes the cumulative Gaussian density. The maximum absolute error of this function is 10^{-4} that is suitable in many cases, including ours.

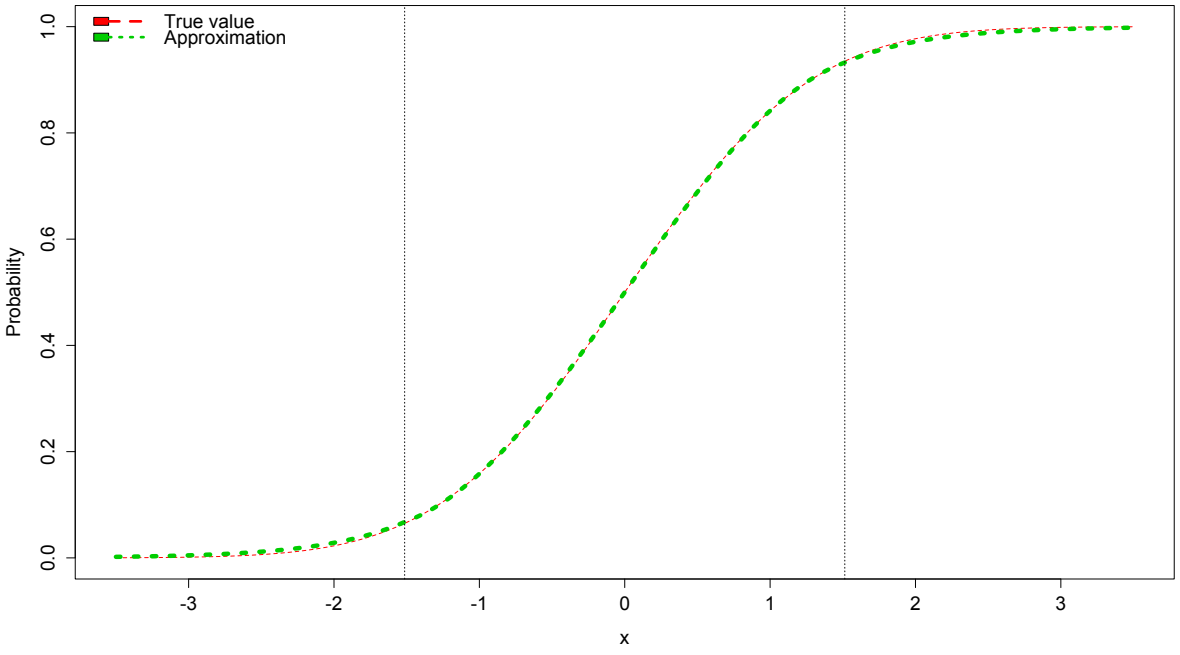

FIGURE 3.5: Visual illustration of the true value of $\Phi(x)$ (dashed line) versus the proposed approximation (dotted line) for a range of values for x in $(-3.5, 3.5)$ interval.

3.8 Algorithm

In this section we propose an algorithm for dlasso. To this end, we follow the literature in [Fan and Li, 2001] and define the iterative algorithm as,

$$\beta^{(k)} = \Big(X'X + \Sigma(\beta^{(k-1)}, \lambda^*, s)\Big)^{-1} X'y, \qquad k = 1, 2, \dots \tag{3.11}$$

where $\beta^{(0)}$ is an initial estimation for the parameters and $\Sigma(\beta^{(k-1)}, \lambda^*, s)$ is defined by,

$$\Sigma(\beta^{(k-1)}, \lambda^*, s) = \lambda^* \text{Diag}\left[\left(2\Phi(\frac{\beta_i^{(k-1)}}{s}, 0, \frac{1}{\sqrt{2}}) - 1 + 2\frac{\beta_i^{(k-1)}}{s}\phi(\frac{\beta_i^{(k-1)}}{s}, 0, \frac{1}{\sqrt{2}})\right)/\beta_i^{(k-1)}, i = 1, \dots, r\right].$$

In what follows we show the derivations of the equation (3.11) from the Taylor expansion of the penalized likelihood in (3.7).

Recalling the log-likelihood from (3.7),

$$\sum_{i=1}^{T}(y_i - x_i\beta)^2 + \lambda^* \sum_{j=1}^{r} \beta_j \left(2\Phi(\frac{\beta_j}{s}, 0, \frac{1}{\sqrt{2}}) - 1\right). \tag{3.12}$$

Let $\beta^{(0)}$ be an initial estimation of the parameters. Since the function is differentiable in any point, the first order Taylor approximation of the penalty function around $\beta^{(0)}$ is given by,

$$\begin{aligned}\beta(2\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) - 1) \approx \beta^{(0)}(2\Phi(\frac{\beta^{(0)}}{s}, 0, \frac{1}{\sqrt{2}}) - 1) + \\ \left(2\Phi(\frac{\beta^{(0)}}{s}, 0, \frac{1}{\sqrt{2}}) - 1 + \frac{2\beta^{(0)}}{s}\phi(\frac{\beta^{(0)}}{s}, 0, \frac{1}{\sqrt{2}}\right)(\beta - \beta^{(0)}).\end{aligned}$$

We should recognize that this approximation is always possible since the dlasso is differentiable in any value of β.
Given $\beta \approx \beta^{(0)}$ it leads to,

$$\begin{aligned}\beta(2\Phi(\frac{\beta}{s}, 0, \frac{1}{\sqrt{2}}) - 1) \approx \beta^{(0)}(2\Phi(\frac{\beta^{(0)}}{s}, 0, \frac{1}{\sqrt{2}}) - 1) + \\ \frac{1}{\beta^{(0)}}\left(2\Phi(\frac{\beta^{(0)}}{s}, 0, \frac{1}{\sqrt{2}}) - 1 + \frac{2\beta^{(0)}}{s}\phi(\frac{\beta^{(0)}}{s}, 0, \frac{1}{\sqrt{2}}\right)(\beta^2 - \beta^{(0)^2}).\end{aligned} \tag{3.13}$$

Substituting (3.13) in (3.12) results in

$$\begin{aligned}\sum_{i=1}^{T}(y_i - x_i\beta)^2 + \lambda^* \sum_{j=1}^{r} \Bigg[\beta_j^{(0)}(2\Phi(\frac{\beta_j^{(0)}}{s}, 0, \frac{1}{\sqrt{2}}) - 1) + \\ \frac{1}{\beta_j^{(0)}}\left(2\Phi(\frac{\beta_j^{(0)}}{s}, 0, \frac{1}{\sqrt{2}}) - 1 + \frac{2\beta_j^{(0)}}{s}\phi(\frac{\beta_j^{(0)}}{s}, 0, \frac{1}{\sqrt{2}}\right)(\beta_j^2 - \beta_j^{(0)^2})\Bigg],\end{aligned}$$

that minimizing with respect to β leads to the iterative function in (3.11).

3.9 Model selection using generalized information criteria

In this section we derive the Generalized Information Criteria (GIC) in [Konishi and Kitagawa, 1996] for dlasso by,

$$GIC_s(\lambda) = -2\text{logLik} + 2\,tr\left[X\left(X'X + \lambda\, diag(\frac{\partial^2}{\partial\beta^2}\hat{\beta}'(2\Phi(\frac{\hat{\beta}}{s},0,\frac{1}{\sqrt{2}})-1)\right)^{-1}X'\right], \tag{3.14}$$

where *loglik* denotes the non-penalized likelihood evaluated at dlasso estimations. We should recognize that the differentiability of the penalty term is a necessary condition for deriving GIC as pointed out in [Konishi and Kitagawa, 1996, Section 3.3].

In what follows we show the derivation of GIC in (3.14) from the penalized likelihood in (3.6)

$$Q(\beta, s) = \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta) + \lambda\beta'\,\text{sign}_s(\beta),$$

where $\text{sign}_s(\beta) = 2\Phi(\frac{\beta}{s},0,\frac{1}{\sqrt{2}}) - 1$. Without loss of generality we assume that $\sigma^2 = 1$. The first two derivatives of the penalized likelihood with respect to the parameters are given by,

$$\psi(X, y, \beta, \lambda, s) = \frac{\partial}{\partial b}\text{logLik} = -2X'(y - X\beta) + \lambda\left(\text{sign}_s(\beta) + \beta \circ \frac{\partial}{\partial\beta}\text{sign}_s(\beta)\right), \tag{3.15}$$

$$\frac{\partial}{\partial\beta}\psi(X, y, \beta, \lambda, s) = 2X'X + \lambda diag\left(\frac{\partial^2}{\partial\beta^2}\beta'\,\text{sign}_s(\beta)\right), \tag{3.16}$$

where $(\circ)$ is Hadamard product and $\frac{\partial^2}{\partial\beta_j^2}\beta_j\,\text{sign}_s(\beta_j) = \frac{4}{s^3}\phi(\frac{\beta_j}{s},0,\frac{1}{\sqrt{2}})(s^2 - \beta_j^2)$ for $j = 1, 2, \ldots, r$. Referring to [Konishi and Kitagawa, 1996], the general form of GIC is given by

$$GIC_s(\lambda) = -2\text{logLik}\,(\hat{\beta}) + 2tr\left(R_{s,\lambda}^{-1}U_{s,\lambda}\right),$$

with

$$R_{s,\lambda} = -\frac{1}{T}\sum_{i=1}^{T}\frac{\partial\psi(X_{i.}, y_i, \beta, \lambda, s)}{\partial\beta}\Big|_{\hat{\beta}}$$
$$U_{s,\lambda} = \frac{1}{T}\sum_{i=1}^{T}\psi(X_{i.}, y_i, \beta, \lambda, s).\frac{\partial}{\partial\beta}\text{logLik}'\Big|_{\hat{\beta}},$$

where $X_{i.}$ denotes the i^{th} row of X for $i = 1, 2, \ldots, T$, $\hat{\beta}$ is dlasso estimation of the parameters and $\psi(.)$ is given in (3.15). In particular, we show that for the dlasso case $R_{s,\lambda}$ and $U_{s,\lambda}$ are given by,

$$R_{s,\lambda} = -\frac{1}{T}\left(2X'X + \lambda diag\left(\frac{\partial^2}{\partial\beta^2}\beta'\,\text{sign}_s(\beta)\right)\right),$$
$$U_{s,\lambda} = \frac{1}{T}X'X.$$

In the first step we find $U_{s,\lambda} = \psi(X_{i.}, y_i, \beta, \lambda, s)\psi'(X_{i.}, y_i, \beta, 0, s)$ with $\psi(.)$ is given in (3.15). Then,

$$U_{s,\lambda} = \left[X'(y - X\beta) - \lambda\left(\text{sign}_s(\beta) + \beta \circ \text{sign}_s^{(1)}(\beta)\right)\right]\left[X'(y - X\beta)\right]'$$

$$= X'(y - X\beta)(y - X\beta)'X - \lambda\Big(\operatorname{sign}_s(\beta) + \beta \circ \operatorname{sign}_s^{(1)}(\beta)\Big)(y - X\beta)'X. \tag{3.17}$$

where the superscripts denotes the order of the derivatives. Let $\beta_\circ$ be the true (but unknown) parameter and taking the expectation of (3.17) with respect to the true *generating distribution* of the data denoted by g leads to

$$\mathbb{E}_g\Big(X'(y - X\beta)(y - X\beta)'X - \lambda\Big(\operatorname{sign}_s(\beta) + \beta \circ \operatorname{sign}_s^{(1)}(\beta)\Big)(y - X\beta)'X\Big) =$$
$$X'\mathbb{E}_g(ee')X - \lambda\Big(\operatorname{sign}_s(\beta_\circ) + \beta_\circ \circ \operatorname{sign}_s^{(1)}(\beta_\circ)\Big)\mathbb{E}_g(e)'X.$$

The first expectation on the right hand side of this equation is I since errors are assumed to be independent, and the second term is zero. Then the equation reduces to, $U_{\lambda,s} = X'X$.

For the second term, $R_{s,\lambda}$, we have

$$R_{s,\lambda} = X'X + \lambda diag\Big(\frac{\partial^2}{\partial\beta^2}\beta' \operatorname{sign}_s(\beta)\Big).$$

Putting altogether, $GIC_s(\lambda)$ is given by

$$GIC_s(\lambda) = -2\text{logLik} + 2\,tr\Big[X\Big(X'X + \lambda\, diag(\frac{\partial^2}{\partial\beta^2}\beta' \operatorname{sign}_s(\beta))\Big)^{-1}X'\Big].$$

3.9.1 Tuning parameter

In this section we focus on selecting the optimum value for the tuning parameter λ^*. To this end, we set an upper and lower bound for λ^* denoted by $\lambda^*_{\max}$ and $\lambda^*_{\min}$ respectively and make a grid of values. Then a value of λ^* that minimizes GIC, AIC, BIC, GCV, MSPE etc. in the grid is the choice for λ^*_{opt}.

As it is pointed out, choosing a proper value for the tuning parameter requires an upper bound for λ^* that is by itself a challenging question. Obviously choosing a value less than $\lambda^*_{\max}$ result in losing a set of estimations. In contrast, choosing a value greater than $\lambda^*_{\max}$ results in an increase in computational time and cost. Referring to [Friedman et al., 2010, Donoho and Johnstone, 1994, Friedman, 2012] in elastic-net with standardized variables we have

$$T\alpha\lambda^*_{\max} = \max_{p=1,\dots,r} |\langle x_p, y\rangle|, \qquad \text{for fixed } \alpha, \tag{3.18}$$

where $\langle x_p, y\rangle$ denotes the inner product of two variables, $\langle x_j, y\rangle = \sum_{i=1}^{T} x_{ij}y_i$ and α is the proportion of l_1 norm in the elastic-net penalty, $\sum_i(\alpha|\beta_i| + (1-\alpha)\beta_i^2)$. Setting $\alpha = 1$ in (3.18) results in $\lambda^*_{\max,lasso}$ that is the quantity of interest. On the other hand $\lambda^*_{\min}$ can be chosen arbitrary close to zero $\lambda^*_{\min} = \epsilon\lambda^*_{\max}$ for small ϵ. Then, constructing a sequence of k values decreasing from $\lambda^*_{\max}$ to $\lambda^*_{\min}$ on the log scale provides the requirements for model selection criteria. Following [Friedman et al., 2010] we choose $\epsilon = 0.001$ and $k = 50$ in real applications. Similarly, one can form a two dimensional grid for s and λ^* and choose the value of s and λ^* that minimize GIC, AIC, BIC, GCV, MSPE etc. in the grid.

3.10 R package

A complete implementation of dlasso using GIC, AIC, BIC and GCV for model selection is provided in the R package `DLASSO` that accompanies this chapter. The main function `dlasso` allows different options including setting the precision digits and performing a grid search over the value of s and/or λ^*. This package encompasses two more functions to extract coefficients, `coef`, and providing visual illustration of the outputs, `plot`. We refer to the manual https://cran.r-project.org/web/packages/DLASSO/DLASSO.pdf for more details and examples.

3.11 Simulation study

In this section, we design three scenarios similar to [Zou, 2006] where the tuning parameters are selected by minimizing CV error. The purpose of the scenarios is to show that dlasso provides similar or better results to lasso, ridge, elastic-net, SCAD and OLS in terms of prediction accuracy by gradually increasing the level of complexity in the simulations. We estimate parameters using the `msgps` package in R for lasso, ridge, elastic-net over a grid of 50 values for α and `ncvreg` package for SCAD.

In all scenarios, data are simulated from the linear model, $y = X\beta + \sigma e$, $e \sim N(0,1)$. In each simulation we divide the data into training and test. Parameter estimation and tuning parameter selection are on the basis of 10-fold cross validation over the training set. Then mean square prediction error (MSPE) is computed over the test data. Here are the details for the three scenarios:

Scenario 1. We set $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and generate 50 datasets containing 240 observations under the pairwise correlation $\mathbb{C}ov(x_i, x_j) = 0.5^{|i-j|}$ and $\sigma^2 = 3$. 40 observations are assigned to the training set and the rest of 200 are assigned to the test set.

Scenario 2. This is the same as the first scenario except that $\beta_j = 0.85, j = 1, 2, \ldots, 8$.

Scenario 3. In this scenario we consider a group structured data generating procedure. The 15 coefficients are divided into three groups, $\beta^{(1)} = c(1, 2, 3, 4, 5)$, $\beta^{(2)} = c(0.5, 0.5, 0.5, 0.5, 0.5)$ and $\beta^{(3)} = c(0, 0, 0, 0, 0)$. We consider a high correlation of 0.9 amongst the first five covariates corresponding to $\beta^{(1)}$. Further, we assume 0.5 correlation in the second group and generate the covariates in the third group

independently. Then, the final correlation matrix is

$$\rho = \begin{pmatrix}
|1.0 & 0.9 & 0.9 & 0.9 & 0.9| & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
|0.9 & 1.0 & 0.9 & 0.9 & 0.9| & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
|0.9 & 0.9 & 1.0 & 0.9 & 0.9| & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
|0.9 & 0.9 & 0.9 & 1.0 & 0.9| & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
|0.9 & 0.9 & 0.9 & 0.9 & 1.0| & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & |1.0 & 0.5 & 0.5 & 0.5 & 0.5| & 0 & 0 & 0 & 0 & 0 \\
 & & & & & |0.5 & 1.0 & 0.5 & 0.5 & 0.5| & 0 & 0 & 0 & 0 & 0 \\
 & & & & & |0.5 & 0.5 & 1.0 & 0.5 & 0.5| & 0 & 0 & 0 & 0 & 0 \\
 & & & & & |0.5 & 0.5 & 0.5 & 1.0 & 0.5| & 0 & 0 & 0 & 0 & 0 \\
 & & & & & |0.5 & 0.5 & 0.5 & 0.5 & 1.0| & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & |1 & 0 & 0 & 0 & 0| \\
 & & & & & & & & & & |0 & 1 & 0 & 0 & 0| \\
 & & & & & & & & & & |0 & 0 & 1 & 0 & 0| \\
 & & & & & & & & & & |0 & 0 & 0 & 1 & 0| \\
 & & & & & & & & & & |0 & 0 & 0 & 0 & 1|
\end{pmatrix}.$$

To make the situation more complex, we set $\sigma^2 = 15$. Similar to Scenario 1 and 2, 240 observations are generated and divided into 40 and 200 for training and test respectively.

From Table (3.1) and Figure (3.6), dlasso shows slightly better or similar results to lasso and elastic-net in the second and third scenarios where the coefficients are small and there is grouped structure in data. Lasso shows the best result in Scenario 1. From this table, ridge and OLS show the worse results in all scenarios. Dlasso shows significantly better results compared to ridge, the only differentiable penalty, and SCAD, which is the only non-convex opponent.

Method	**MSPE**		
	Scenario 1	**Scenario 2**	**Scenario 3**
dlasso	3.23(0.046)	3.49*(0.11*)	18.71*(0.39)
lasso	3.21*(0.043*)	3.71(0.11*)	18.80(0.39)
elastic-net	3.24(0.048)	3.51(0.11*)	18.85(0.35*)
ridge	3.63(0.057)	4.50(0.20)	26.30(0.94)
SCAD	3.29(0.053)	3.55(0.15)	23.58(1.48)
OLS	3.31(0.056)	3.56(0.12)	22.69(0.51)

TABLE 3.1: Comparing dlasso, lasso, ridge, elastic-net, SCAD and OLS from three scenarios on the basis of median of MSPE over the test set. The values in parentheses are the corresponding standard errors of the medians result from bootstrap with 5000 iterations. The asterisk denotes the minimum value.

3.12 Real data illustration

For the first real data demonstration we use the Diabetes dataset that is previously studied in [Efron et al., 2004]. This dataset contains 442 measurements on 10 variables, namely age, sex, body mass index, average blood pressure and six blood serum measurements as well as a quantitative measure of disease progression, which is the response. We normalize all the covariates to have zero mean and unit variance and the response to have zero mean.

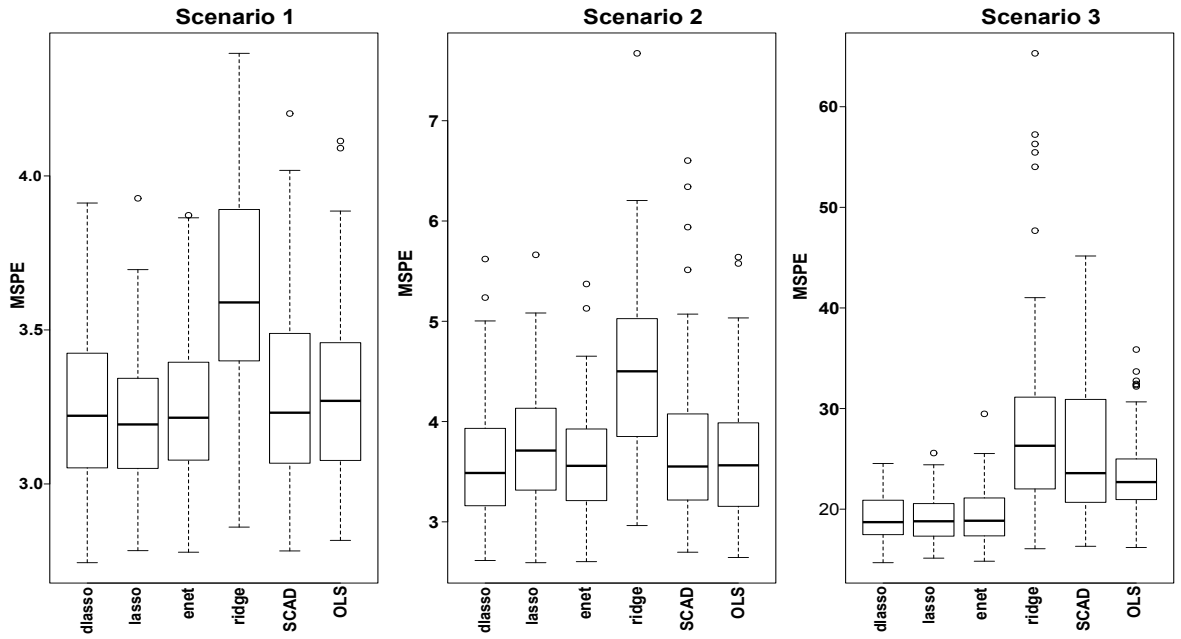

FIGURE 3.6: Comparing dlasso, lasso, elastic-net, ridge, SCAD and OLS estimations with respect to means square prediction error over the test set in the three scenarios.

We compare dlasso to lasso, SCAD, OLS and elastic-net on the basis of AIC, BIC and the number of non-zero estimations. We should stress that GIC is not defined for lasso, elastic-net and SCAD. Then we compare the results on the basis of AIC and BIC. For dlasso and elastic-net we consider a sequence of 50 values for s, α in $(0.001, 1)$ interval. The tuning parameter in all models is selected by minimizing BIC. The results are shown in Table (3.2).
Comparing BIC and AIC and the number of non-zero coefficients presented in Table (3.2) shows that dlasso for $s = 0.001$, lasso and elastic-net for selected $\alpha = 0.001$ perform similarly and better than SCAD. Further dlasso, lasso and elastic-net select sex, bmi, map, tc, hdl, ltg and glu whereas SCAD does not select the last one, glu. We should stress that the selection of the tuning parameter in dlasso is on the basis of a sequence of discrete values for λ^*, precisely 100 values, that results in slightly different results to lasso. We should stress that the solution path for lasso and dlasso are highly similar (except tch) regardless of the tuning parameter selection criteria as it is shown in Figure (3.7). Then the different estimations from both methods differ mainly based on the selection of tuning parameter.

For the second application, we apply the new penalty to the prostate data previously studied in the original paper of elastic-net [Zou and Hastie, 2005]. Prostate data is first introduced by [Stamey et al., 1989] to examine the correlation between the level of prostate specific antigen and a number of clinical measurements in 97 men who were about to receive a radical prostatectomy and includes eight covariates: log cancer volume (lcanvol), log prostate weight (lweight), log benign prostatic hyperplasia amount (lbph), log capsular penetration (lcp), age, Gleason score (gleason), percentage Gleason scores 4 or 5 (pgg45), seminal vesicle invasion (svi) as well as prostate specific antigen (lpsa) for the response.

Lasso, ridge, SCAD, OLS, elastic-net for a range of α in $(0.001, 1)$ interval as well as dlasso for $s \in (0.001, 1.5)$, $s = 1$ and $s = 100$ are applied to the data. Covariates are normalized to have zero mean and unit variance and the response to have zero mean. BIC is used to select the tuning parameters. Results

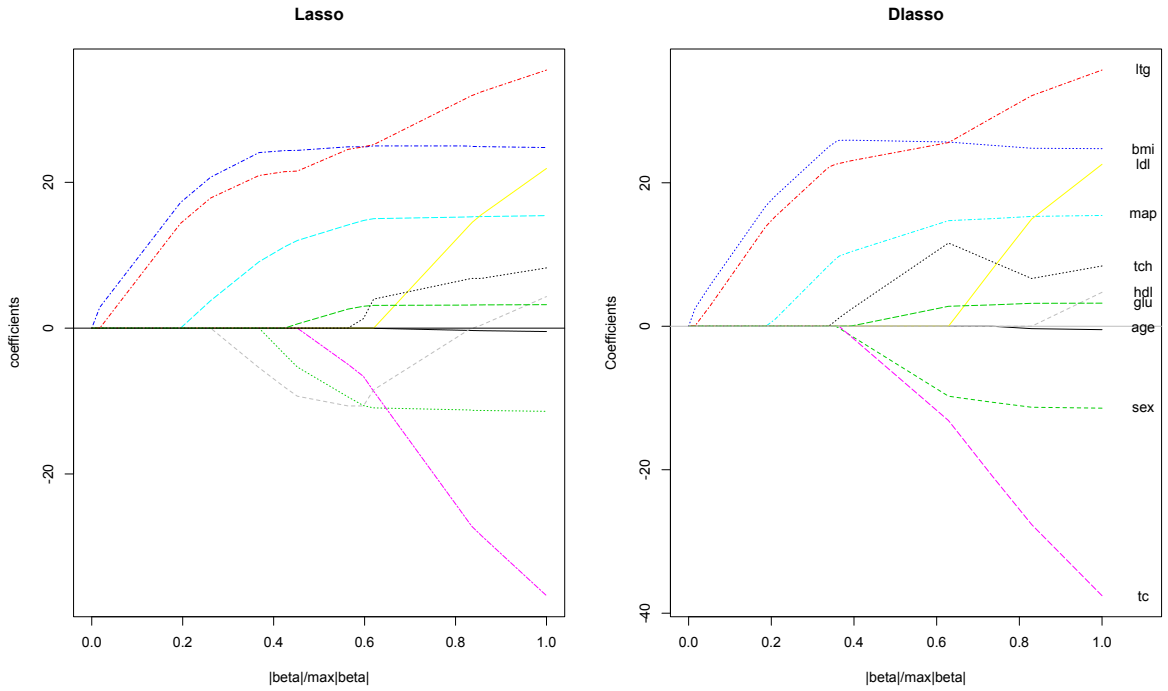

FIGURE 3.7: Comparison of lasso and dlasso in term of solution path for Diabetes dataset. The right plot is drawn by `DLASSO` package whereas the left one is drawn by `MSGPS` package in R

Model Comparison

Method	Final precision	AIC	BIC	df
dlasso	$s = 10^{-3}$	4786.4	4808.2	7
lasso	-	4784.7	4806.4	7
elastic-net	$\alpha = 10^{-3}$	4784.9	4806.5	7
SCAD	-	4792.2	4825.5	6
OLS	-	4794.0	4839.0	10

Parameter Estimation

Method	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu
dlasso	0	-6.9	24.6	12.8	-1.9	0	-9.9	0	22.7	1.3
lasso	0	-7.3	24.6	13.0	-2.2	0	-9.8	0	23.0	1.5
elastic-net	0	-7.2	24.5	13.0	-2.1	0	-10.0	0	22.7	1.6
SCAD	0	-10.9	25.3	15.6	-4.0	0	-12.3	0	25.0	0
OLS	-0.47	-11.4	24.8	15.5	-37.7	22.7	4.8	8.3	35.8	3.2

TABLE 3.2: Comparing dlasso, lasso, OLS, elastic-net and SCAD on the basis of BIC and AIC and the number of non-zero estimations for the diabetes dataset.

are provided in Table (3.3).

Dlasso for $s = 0.001$ shows the best result amongst the rest of the values for s and it is the same as both lasso and optimal elastic-net ($\alpha = 0.001$) with respect to AIC, BIC and df, and all three models are better than SCAD. Dlasso for $s = 1$ performs similar to ridge as it is shown in the middle of the table and predicted in theory. It shows better BIC compared to OLS that can be due to existence of (small) regularization in dlasso. In terms of the number of non-zero estimations, dlasso, lasso and elastic-net select the same number of covariates whereas SCAD selects 4 variables. Similar to previous example, Figure (3.8) show the similarity in solution path of dlasso and lasso.

Method	Precision	AIC	BIC	df	Non-zero variables
Lasso					
dlasso	s=0.001	207.6	216.1	5	lcavol, ibph, lweight, pgg45,svi
lasso	-	206.7	215.3	5	lcavol, ibph, lweight, pgg45,svi
elastic-net	α=0.001	206.8	215.3	5	lcavol, ibph, lweight, pgg45,svi
SCAD	-	214.6	231.0	4	lcavol, ibph, lweight, svi
Ridge					
dlasso	1	207.4	227	8	all variables
ridge	-	207.1	226	8	all variables
OLS					
dlasso	100	204	228	8	all variables
OLS	-	202	233	8	all variables

TABLE 3.3: Comparison of lasso, ridge, SCAD, OLS, elastic-net and new penalty for $s = 0.001, 1, 100$ and the result from a grid search over $s, \lambda^* \in (10^{-3}, 1)$ for Prostate dataset. Methods are compared based on AIC, BIC and sparsity.

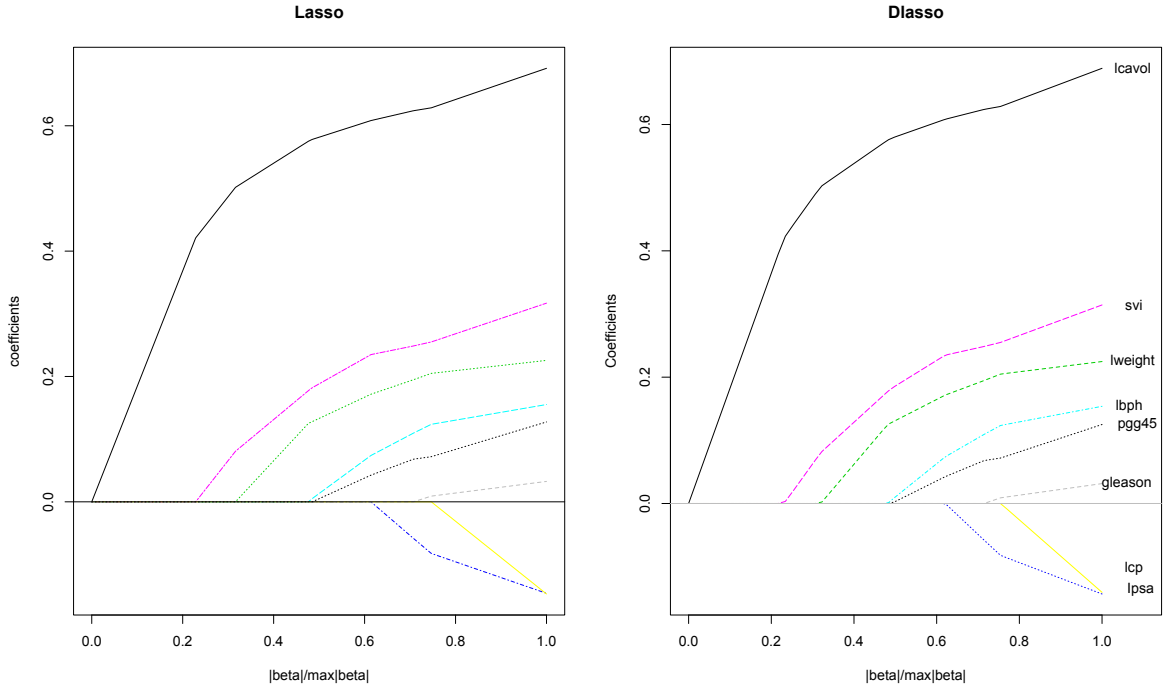

FIGURE 3.8: Comparison of lasso and dlasso in term of solution path for Prostate dataset.

3.13 Conclusion remarks

In this chapter we have proposed a novel differentiable penalty term that is capable of producing similar results to lasso, ridge and to some extent elastic-net. That is, this new penalty can be used in smooth situations to select more variables than lasso. We discussed this new penalty from a theoretical point of view, computational complexity and practical perspective by proposing an efficient algorithm and providing an R package. We have shown the applicability of this new penalty by means of simulations

and real applications in Diabetics and Prostate datasets, and compared the results to well-known models such as lasso, ridge, elastic-net and SCAD.

3.13.1 Future study

In Section § 3.4 we focused on a special case of the following function,

$$f(x,s,\alpha,\gamma) = x\left[\frac{2}{\sqrt{\pi}}\int_0^{(\frac{x}{s})^\alpha} e^{-t^2}dt\right]^\gamma = x \times \left[\text{ erf }\left((\frac{x}{s})^\alpha\right)\right]^\gamma, \qquad \{\gamma,\alpha\} \in \mathbb{R}, s > 0, x \geq 0,$$

for $\alpha = \gamma = 1$. We suggest setting $\gamma = 1$ and an odd value for α to preserve symmetry of the entire function over the vertical axis. Then, studying this function in the context of penalized likelihood can be an interesting subject for future research.

Chapter 4

A Bayesian approach to discrete Weibull regression for counts

4.1 Introduction

Data in the form of counts appear in many application areas, from medicine, social and natural sciences to econometrics, finance and industry [Cameron and Trivedi, 2013]. In medicine, two examples of this are the number of days that patients stay in hospital, commonly used as an indicator of the quality of care and planning capacity within a hospital [Atienza et al., 2008, Carter and Potts, 2014], and the number of visits to a specialist [Machado and Santos Silva, 2005], often taken as a measure of demand in healthcare. Other examples are high-throughput genomic data generated by next generation sequencing experiments [Ozsolak and Milos, 2011, Bao et al., 2014, Robinson and Smyth, 2008] or lifetime data, such as the number of cycles before a machine breaks down [Nagakawa and Osaki, 1975].

Similar to Weibull regression, which is widely used in lifetime data analysis and survival analysis for continuous response variables, [Kalktawi et al., 2016] have recently proposed a regression model for a discrete response based on the discrete Weibull distribution. A number of studies have found a good fit of this distribution in comparison to other distributions for count data [Bracquemond and Gaudoin, 2003, Englehardt and Li, 2011, Lai, 2013]. In the context of regression, [Kalktawi et al., 2016] show two important features of a discrete Weibull (DW) distribution that make this a valuable alternative to the more traditional Poisson and Negative Binomial distributions and their extensions such as Poisson mixtures [Hougaard et al., 1997], Poisson-Tweedie [Mikel Esnaola et al., 2013], zero-inflated regression[Lam et al., 2006] and COMPoisson [Sellers and Shmueli, 2010]: the ability to capture over or under-dispersion and a closed-form analytical expression of the quantiles of the conditional distribution.

In [Kalktawi et al., 2016], maximum likelihood is used for the estimation of the parameters. This is in general the most common approach for parameter estimation in regression analysis of counts. Among the contributions to Bayesian estimation of discrete regression models, [El Sayyad, 1973] consider the case of Poisson regression, [Zhou et al., 2012] provide an efficient Bayesian implementation of negative Binomial regression, [Mohebbi et al., 2014] develop Bayesian estimation for a Poisson and negative Binomial regression with a conditional autoregressive correlation structure whereas [Angers and Biswas,

2003, Ghosh et al., 2006, Neelon et al., 2010, Liu and Powers, 2012] study zero-inflated Poisson regression. In this chapter, we contribute to this literature, by providing the first Bayesian approach for parameter estimation in discrete Weibull regression. For the choice of prior distributions, we consider both the case of non-informative priors and the case of Laplace priors with a hyper penalty parameter. The choice of Laplace priors induces parameter shrinkage [Park and Casella, 2008, Kyung et al., 2010b], and, with the use of Bayesian credible intervals, leads to variable selection, similar to alternative approaches such as reversible jumps Markov chain Monte Carlo [Green, 1995] and spike and slab [Ishwaran and Rao, 2005].

The aim of this chapter is two-fold. Firstly, we highlight the role that the discrete Weibull distribution has in modelling count data from a variety of applications, beyond its current limited use to lifetime data. We particularly emphasize applications in the medical domain, using several datasets of medical records. Secondly, we present a novel Bayesian regression model for counts based on the assumption of a discrete Weibull conditional distribution. The remainder of this chapter is organized as follows. Section § 4.2 describes the discrete Weibull regression model, with a more general parametrization than that presented in the literature. Section § 4.3 describes Bayesian parameter estimation for a discrete Weibull regression model. Theoretical properties of the posterior under non-informative uniform priors is discussed in Section § 4.4. Section § 4.6 presents an extensive simulation study, whereas Section § 4.7 shows the analysis of real data via Bayesian discrete regression model and a comparison with existing approaches. Finally, we draw some conclusions in Section § 4.8.

4.2 Discrete Weibull regression

4.2.1 Discrete Weibull distribution

The discrete Weibull distribution was introduced by [Nagakawa and Osaki, 1975], as a discretized form of a continuous Weibull distribution, similarly to the geometric distribution, which is the discretized form of the exponential distribution, and the negative Binomial, which is the discrete alternative of a Gamma distribution. In some papers, this is referred to as a type I discrete Weibull, as two other distributions were subsequently defined,

$$\begin{array}{lll} \text{Type I} & : F(y;q,\beta) = 1 - q^{(y+1)^\beta} & q \in (0,1), \beta > 0, y = 0,1,2,\dots \\ \text{Type II} & : F(y;c,\beta) = \sum_{k=1}^{y<m} ck^{\beta-1} \prod_{j=1}^{k-1}(1-cj^{\beta-1}) & c \in (0,1), \beta > 0, y = 1,2,\dots,m \\ \text{Type III} & : F(y;c,\beta) = 1 - e^{-c\sum_{i=0}^{y+1} i^\beta} & c > 0, \beta \geq -1, y = 0,1,2,\dots. \end{array}$$

[Bracquemond and Gaudoin, 2003] review the three different distributions and point out the advantages of using the type I distribution: It has an unbounded support, differently than the type II distribution, and it has a more straightforward interpretation, differently than the type III distribution.

Throughout this chapter, we will refer to the Type I discrete Weibull distribution as $\text{DW}(q,\beta)$. A similar definition can be given on the support $1,2,\dots$. In this case, $F(y;q,\beta) = 1 - q^{y^\beta}$, for $y = 1,2,\dots$. Comparing this cdf with that of a continuous Weibull distribution with parameters α and γ,

$$F(y;\alpha,\gamma) = \begin{cases} 1 - e^{-\alpha y^\gamma} & \text{if } y \geq 0 \\ 0 & \text{if } y < 0 \end{cases},$$

one can see that there is a direct correspondence between β and γ, whereas q in the discrete case corresponds to $\exp(-\alpha)$ in the continuous case [Khan et al., 1989].

Given the form of the cumulative distribution function, the DW distribution has the following probability mass function:

$$p(y; q, \beta) = q^{y^\beta} - q^{(y+1)^\beta}, \quad y = 0, 1, 2, \ldots$$

with q and β denoting the shape parameters.

4.2.2 Inference for Discrete Weibull: Existing Approaches

[Khan et al., 1989] derive estimators of DW parameters q and β using the method of moments and a new method which they call *the method of proportions*, and they find a good performance for the latter. Let $y_1, \ldots, y_n$ be a random sample from a DW(q, β) distribution and denote $z = \sum_{i=1}^n I(y_i = 0)$ and $u = \sum_{i=1}^n I(y_i = 1)$. Using the method of proportions, the following estimators of q and β are proposed:

$$\begin{aligned} \hat{q} &= 1 - \frac{z}{n} \\ \hat{\beta} &= \ln\left[\ln\left(1 - \frac{z}{n} - \frac{u}{n}\right) / \ln\left(1 - \frac{z}{n}\right)\right] / \ln(2). \end{aligned}$$

These estimators use only the zeros and ones in the sample. [Araújo Santos and Fraga Alves, 2013] derive an improved estimator of β, by taking all observations into account. In particular, let d_m be the maximum observed value of y and let $k = d_m - 1$. If $d_m > 2$, then the following improved estimator is proposed:

$$\hat{\beta} = \frac{1}{k} \sum_{d=1}^{k} \ln\left[\ln\left(1 - \hat{F}(d)\right) / \ln(\hat{q})\right] / \ln(d+1),$$

where $\hat{F}$ denotes the empirical cdf. When $d_m = 2$, this estimator is equivalent to the one from [Khan et al., 1989]. Note that in both cases, no estimates of β can be obtained when $\hat{q} = 1$, i.e. there are no zero counts in the observed data, or $\hat{q} = 0$, i.e. all counts are zero.

[Kulasekera, 1994] considers maximum likelihood for the estimation of q and β. The likelihood function for a discrete Weibull sample is given by:

$$L(q, \beta) = \prod_{i=1}^{n} \left(q^{y_i^\beta} - q^{(y_i+1)^\beta}\right),$$

the maximum of which can be found numerically.

There is no explicit work in the literature for building confidence intervals for discrete Weibull parameters, although standard asymptotic likelihood and bootstrap approaches can be used. The Bayesian approach that we devise in this chapter will lead naturally to credible intervals for the parameters.

4.2.3 Regression via a discrete Weibull distribution

Let y be the response with possible values $0, 1, \ldots$, and let $X = (x_1, \ldots, x_p)$ be a vector of p covariates. We assume that the conditional distribution of y given X follows a DW distribution with parameters

q and β. There are a number of possible choices to link the parameters q and β to the covariates. In particular, we propose the following link functions:

1. q dependent on X via

$$\log(-\log(q)) = X\theta \text{ or}$$
$$\log\left(\frac{q}{1-q}\right) = X\theta,$$

where $\theta = (\theta_0, \ldots, \theta_p)'$. Both transformations restrict the value of q to lie in $(0,1)$ interval. However, the log-log transformation is asymmetric while the logit is symmetric around $q = 0.5$. Moreover, log-log transformations contains two logarithm operators that leads to numerical instability for q close to 0 and 1. The applicability of log-log transformation is motivated by continuous-time models for the occurrence of events including continuous Weibull regression [Hosmer et al., 2011, Cox, 1992, Gimenez et al., 1997] whereas the logit transformation is motivated by the applications in bounded outcome scores and has proved to be rather effective for statistical inference, see e.g. [Hosmer et al., 2011, Lesaffre et al., 2007].

2. We assume a logarithmic link for the second parameter β and the covariates, in order to capture more complex dependencies. Thus, β dependent on X via

$$\log(\beta) = X\gamma,$$

where $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_p)'$ contains the same number of parameters as θ.

In general, there are no identifiability issues in the model, as the part of the likelihood from zero observations depends only on q. However, in our simulation and real data analyses, we found the logit function to be only marginally superior to the log-log transformation, whereas the additional β parametrization is often not selected due to over-parametrization.

4.3 Bayesian inference for discrete Weibull regression

In this section, we discuss Bayesian estimation of the regression parameters $\theta = (\theta_0, \ldots, \theta_p)'$ and $\gamma = (\gamma_0, \ldots, \gamma_p)'$. The advantage of choosing Bayesian approaches over classical maximum likelihood inference is two-fold. Firstly, the possibility of taking prior information into account, such as sparsity or information from historical data and, secondly, the procedure returns credible intervals for all parameters.

Given n observations y_i and $(x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, for the response and the covariates, respectively, and letting $x_{i\cdot}$ be the row vector $x_{i\cdot} = (1, x_{i1}, \ldots, x_{ip})$, the likelihood for the most general case under a logit transformations is given by

$$L(X, y|\theta, \gamma) = \prod_{i=1}^{n} \left(\left(\frac{e^{x_{i\cdot}\theta}}{1+e^{x_{i\cdot}\theta}}\right)^{y_i^{x_{i\cdot}\gamma}} - \left(\frac{e^{x_{i\cdot}\theta}}{1+e^{x_{i\cdot}\theta}}\right)^{(y_i+1)^{x_{i\cdot}\gamma}} \right),$$

where we allow the same $x_{i\cdot}$ for both $x_{i\cdot}\theta$ and $x_{i\cdot}\gamma$. Following the same procedure, one can form the corresponding likelihood under the log-log transformation. We consider different prior distributions on

θ and γ. Unfortunately, in the context of discrete Weibull regression, there are no conjugate priors. However, we will show theoretically how a uniform non-informative prior leads to posterior distribution which is proper with finite moments and, in simulation and real data study, we show how this prior achieves an acceptable rate of mixing as well as comparable estimation to maximum likelihood. In addition, we consider a prior on the regression coefficients that encourages sparsity. In particular, we consider a Laplace prior for θ and γ, of the form

$$\begin{aligned} p(\theta|\lambda) &= \frac{\lambda}{2} e^{-\lambda|\theta|}, \qquad \lambda \geq 0, \\ p(\gamma|\tau) &= \frac{\tau}{2} e^{-\tau|\gamma|}, \qquad \tau \geq 0. \end{aligned} \tag{4.1}$$

For a given choice of λ and τ, maximising the posterior probability under these priors corresponds to maximising the l_1 penalised log-likelihood

$$\log L(X, y|\theta, \gamma) - \lambda \sum_{j=1}^{p} |\theta_j| - \tau \sum_{k=1}^{p} |\gamma_k|,$$

as in the traditional lasso approach [Park and Casella, 2008, Tibshirani, 1996]. We further assume an InverseGamma(a,b) hyper prior for both λ and τ, leading to the posterior distribution

$$p(\theta, \gamma|y, X) \propto L(y, X|\theta, \gamma) \times p(\theta|\lambda) \times p(\gamma|\tau) \times p(\lambda) \times p(\tau).$$

Since conditional posterior of DW distribution given the parameters is not belonging to a class of known distributions, Gibbs sampler is not applicable. Alternatively, we choose a Metropolis-Hastings (MH) sampler [Hastings, 1970] to draw samples from the full conditional posterior. However, this does not lead to exactly zero estimation of the parameters. But imposing Laplace priors shrink the marginal posterior of the parameters and using HPD interval encourages the sparsity.

Reviewing the literature reveals that MCMC samplers have been used before in the continuous Weibull regression context by [Newcombe et al., 2014], which utilizes a Reversible Jump MCMC, and [Soliman et al., 2012] which uses a hybrid method consisting of Metropolis-Hastings and Gibbs sampler to estimate parameters in a three parameters continuous Weibull distribution. Moreover, [Polpo et al., 2009] make use of a Metropolis-Hasting sampler to make inference for a continuous two-parameters Weibull distribution in a censoring framework.

We use MH algorithm in the following steps:

Step 1. Initializing the algorithm with MLE estimation of the parameters or random values within the space of the parameters.

Step 2. Set a proposal distribution $g(.)$ on the full set of parameters $\pi = (\theta, \gamma)$. We choose a multivariate normal proposal with covariance matrix set to the fisher information matrix of the likelihood, but other choices are possible.

Step 3. Draw a random sample from the proposal distribution, e.g. π_k at iteration k.

Step 4. Evaluate the acceptance probability

$$\alpha = \min\left(1, \frac{L(X,y|\pi_k)p(\pi_k)g(\pi_{k-1}|\pi_k)}{L(X,y|\pi_{k-1})p(\pi_{k-1})g(\pi_k|\pi_{k-1})}\right),$$

where $L(X,y|\pi_k)$ is the conditional DW likelihood given the proposal values and $p(.)$ is the prior.

Step 5. Accept the proposal π_k with probability of α.

Step 6. Following the adaptive-MH in [Haario et al., 2001], we update the covariance of the proposal by computing the sample covariance of the chain.

Step 7. Stop if the algorithm is met the maximum iterations.

Step 8. If required, adjust the initial scale of the proposal so that the acceptance rate lies in the recommended $(20,30)\%$ interval [Bedard, 2008] for non-adaptive proposals.

Step 9. Remove $\omega\%$ of the estimation chain for burn-in, e.g. $\omega = 10\%, 25\%, \ldots$

To reduce the computational time of the algorithm, Step 6 can be performed regularly on specific iterations, e.g., every 20, 50 iteration and so on. Further, the burn-in procedure in Step 9 removes the effect of randomly chosen initial values. Adjusting the proposal scale in Step 8 guarantees that the chain is not stationary in one state. Alternatively Step 3 can be written as,

$$\begin{aligned}\pi_k = & I\left(u, \min\left(1, \frac{L(X,y|\pi_k)p(\pi_k)g(\pi_{k-1}|\pi_k)}{L(X,y|\pi_{k-1})p(\pi_{k-1})g(\pi_k|\pi_{k-1})}\right)\right)\pi_k + \\ & I^c\left(u, \min\left(1, \frac{L(X,y|\pi_k)p(\pi_k)g(\pi_{k-1}|\pi_k)}{L(X,y|\pi_{k-1})p(\pi_{k-1})g(\pi_k|\pi_{k-1})}\right)\right)\pi_{k-1}\end{aligned} \tag{4.2}$$

where u is a random sample from uniform distribution over $(0,1)$ and $I(.,.)$ is the indicator function,

$$I(a,b) = \begin{cases} 1 & b < a \\ 0 & o.w. \end{cases},$$

and $I^c = 1 - I$ is the complementary indicator function. [Hastings, 1970] shows that repeating Step 2 to 4 leads to an estimation for the posterior, provided the proposal is carefully configured in Step 5. We should stress that for a symmetric distribution, $P(a|b) = P(b|a)$ in mean. As a result, normal proposals can be cancelled in Step 3 that leads to an improvement in the algorithm speed. From the posterior distribution, the mode of the marginal densities can be used as point estimate of the parameters, whereas the whole distribution is used for building credible intervals. In the case of Laplace priors, the inclusion or not of zero in the Highest Posterior Density (HPD) interval is used for variable selection. In terms of computational complexity, DW and NB distributions have the same number of parameters, but there are fewer operations involved in the evaluation of the DW distribution than in the NB distribution, leading to an expected lower computational complexity for DW.

4.4 Some key theoretical results

Although a standard conjugate prior distribution is not available for the discrete Weibull regression model, MCMC methods can be used to draw samples from the posterior distributions, as described in

the previous section. This, in principal, allows us to use virtually any prior distribution. However, in the case of non-informative priors, we should select only those that yield proper posteriors. In this section, we show some key theoretical results on this. In particular, we prove that the choice of uniform non-informative priors on the parameters, i.e. $p(\theta) \propto 1$ and $p(\gamma) \propto 1$, leads to a proper posterior distribution with finite moments.

Thus, as a first result, we show that, under uniform non-informative priors, the posterior is proper, that is

$$0 < \int_{\boldsymbol{\theta}} \int_{\gamma} L(x, y|\boldsymbol{\theta}, \boldsymbol{\gamma}) d\boldsymbol{\gamma}\, d\boldsymbol{\theta} < \infty.$$

For simplicity, we consider the case where there is no regression model on β, i.e. $p(\beta) \propto 1$ for $\beta > 0$. In addition, we consider the logit link for q, although the proof will cover also the log-log case.

Lemma 4.1. Let

$$f(y) = 1 - (e^{-a})^{y^\beta} - (\frac{a}{1+a})^{y^\beta},$$

and assuming $y > 0$, $\beta > 0$ and $a > 0$, then $f(y)$ is an increasing function of y.

Proof. The derivative of f with respect to y is

$$\begin{aligned}\frac{d\, f(y)}{dy} &= -\beta y^{\beta-1} e^{-ay^\beta} \log(e^{-a}) - \beta y^{\beta-1} (\frac{a}{1+a})^{y^\beta} \log(\frac{a}{1+a}) \\ &= \beta y^{\beta-1} \left(a e^{-ay^\beta} - (\frac{a}{1+a})^{y^\beta} \log(\frac{a}{1+a}) \right).\end{aligned}$$

Since $a > 0$ and $\log(\frac{a}{1+a}) < 0$, the derivative is always positive.

□

Theorem 4.1 (Proper posterior). Let $y = (y_1 \dots y_n)$, $x = (1\, x_1 \dots x_p)$ and $\boldsymbol{\theta} = (\theta_0 \dots \theta_p)'$ be response, covariates and regression parameters, respectively. Under the DW regression model $Y|x \sim DW\left(\frac{e^{x\boldsymbol{\theta}}}{1+e^{x\boldsymbol{\theta}}}, \beta\right)$ and choosing non-informative priors on $\boldsymbol{\theta}$ and β, the posterior distribution is proper, i.e.

$$\int_{\beta} \int_{\boldsymbol{\theta}} \prod_{i=1}^{n} \left\{ \left(\frac{e^{x_i\boldsymbol{\theta}}}{1+e^{x_i\boldsymbol{\theta}}} \right)^{y_i^\beta} - \left(\frac{e^{x_i\boldsymbol{\theta}}}{1+e^{x_i\boldsymbol{\theta}}} \right)^{(y_i+1)^\beta} \right\} d\boldsymbol{\theta}\, d\beta < \infty.$$

Proof. Let $S = \{i; y_i \neq 0, y_i \neq 1\}$ be the set of indices for which the response y is different from zero and one. Let $m \leq n$ be the cardinality of S and assuming $S \neq \emptyset$. This excludes the case where the data contain only zeros and ones, a special case that is normally modelled by a Bernoulli conditional distribution.

Under this assumption, it follows that

$$\prod_{i=1}^{n} \left\{ \left(\frac{e^{x_i\theta}}{1+e^{x_i\theta}} \right)^{y_i^\beta} - \left(\frac{e^{x_i\theta}}{1+e^{x_i\theta}} \right)^{(y_i+1)^\beta} \right\} \leq \prod_{i \in S} \left\{ \left(\frac{e^{x_i\theta}}{1+e^{x_i\theta}} \right)^{y_i^\beta} - \left(\frac{e^{x_i\theta}}{1+e^{x_i\theta}} \right)^{(y_i+1)^\beta} \right\}.$$

Choosing any $k \in S$, such that $\min |x_{kj}| \neq 0, j = 1, \ldots, p$, results in

$$\begin{aligned}\prod_{i \in S} \left\{ \left(\frac{e^{x_i\theta}}{1+e^{x_i\theta}} \right)^{y_i^\beta} - \left(\frac{e^{x_i\theta}}{1+e^{x_i\theta}} \right)^{(y_i+1)^\beta} \right\} &\leq \left(\frac{e^{|x_k|\theta}}{1+e^{|x_k|\theta}} \right)^{y_k^\beta} - \left(\frac{e^{|x_k|\theta}}{1+e^{|x_k|\theta}} \right)^{(y_k+1)^\beta} \\ &= \left(\frac{e^{|x_k|\theta}}{1+e^{|x_k|\theta}} \right)^{y_k^\beta} \left(1 - \left(\frac{e^{|x_k|\theta}}{1+e^{|x_k|\theta}} \right)^{(y_k+1)^\beta - y_k^\beta} \right) \\ &\leq \left(\frac{e^{|x_k|\theta}}{1+e^{|x_k|\theta}} \right)^{y_k^\beta}.\end{aligned}$$

Without loss of generality we assume $p = 1$ so, $\theta = (\theta_0, \theta_1)$ and $x_k = (x_{k0}, x_{k1})$. Then we consider the cases where $\theta_j > 0$ or $\theta_j \leq 0$ for $j = 0, 1$. Then we consider the four cases where the θs are both positive, negative or of different signs, respectively.

Assuming $\theta_j \leq 0, j = 0, 1$ we get,

$$\left(\frac{e^{|x_k|\theta}}{1+e^{|x_k|\theta}} \right)^{y_k^\beta} \leq (e^{|x_k|\theta})^{y_k^\beta},$$

where the integral over θ and β is bounded ($\leq \frac{1}{2|x_{k0}x_{k1}|\log(y_k)}$).
Similarly, for $\theta_0 \leq 0$ and $\theta_1 > 0$ we have,

$$\left(\frac{e^{|x_k|\theta}}{1+e^{|x_k|\theta}} \right)^{y_k^\beta} \leq (e^{|x_{k0}|\theta_0})^{y_k^\beta} \left(1 + e^{-|x_{k1}|\theta_1} \right)^{-\beta},$$

and

$$\begin{aligned}\int_0^\infty \int_{\theta_1=0}^\infty \int_{\theta_0=-\infty}^0 (e^{|x_{k0}|\theta_0})^{y_k^\beta} \left(1 + e^{-|x_{k1}|\theta_1} \right)^{-\beta} d\theta_0 d\theta_1 d\beta = \\ \int_{\theta_1=0}^\infty \int_{\beta=0}^\infty \frac{1}{y_k^\beta} \left(1 + e^{-|x_{k1}|\theta_1} \right)^{-\beta} d\beta \, d\theta_1 \\ = \int_{\theta_1=0}^\infty \frac{1}{\log\left(y_k(1 + e^{-|x_{k1}|\theta_1}) \right)} d\theta_1.\end{aligned}$$

The function $\log^{-1}\left(y_k(1 + e^{-|x_{k1}|\theta_1}) \right)$ is continuous and bounded over the domain of θ_1, provided $y_k > 1$. Thus, the integral is bounded. A similar derivation would hold for the case $\theta_0 > 0$, $\theta_1 \leq 0$.

For the final case, $\theta_j > 0, j = 0, 1$, we have

$$\begin{aligned}\prod_{i=1}^n \left\{ \left(\frac{e^{x_i\theta}}{1+e^{x_i\theta}} \right)^{y_i^\beta} - \left(\frac{e^{x_i\theta}}{1+e^{x_i\theta}} \right)^{(y_i+1)^\beta} \right\} &\leq \left(1 + e^{-|x_k|\theta} \right)^{-y_k^\beta} - \left(1 + e^{-|x_k|\theta} \right)^{-(y_k+1)^\beta} \\ &\leq e^{-e^{|x_k|\theta}(y_k+1)^\beta} - e^{-e^{|x_k|\theta} y_k^\beta}\end{aligned}$$

$$
\begin{aligned}
&= e^{-e^{|x_{k0}|\theta_0}(y_k+1)^\beta} e^{-e^{|x_{k1}|\theta_1}(y_k+1)^\beta} - e^{-e^{|x_{k0}|\theta_0} y_k^\beta} e^{-e^{|x_{k1}|\theta_1} y_k^\beta} \\
&\leq e^{-|x_{k0}|\theta_0 (y_k+1)^\beta} e^{-|x_{k1}|\theta_1 (y_k+1)^\beta} - e^{-|x_{k0}|\theta_0 y_k^\beta} e^{-|x_{k1}|\theta_1 y_k^\beta},
\end{aligned}
$$

where the last term is a direct result of Lemma (4.1) with $a = e^{|x_{kj}|\theta_j}, j = 0, 1$. Thus,

$$
\begin{aligned}
&\int_{\theta_1=0}^{\infty} \int_{\theta_0=0}^{\infty} e^{-e^{|x_{k0}|\theta_0}(y_k+1)^\beta} e^{-e^{|x_{k1}|\theta_1}(y_k+1)^\beta} - e^{-e^{|x_{k0}|\theta_0} y_k^\beta} e^{-e^{|x_{k1}|\theta_1} y_k^\beta} \, d\theta_0 \, d\theta_1 \\
&\qquad \leq \int_{\theta_1=0}^{\infty} \int_{\theta_0=0}^{\infty} e^{-|x_{k0}|\theta_0 (y_k+1)^\beta} e^{-|x_{k1}|\theta_1 (y_k+1)^\beta} - e^{-|x_{k0}|\theta_0 y_k^\beta} e^{-|x_{k1}|\theta_1 y_k^\beta} \, d\theta_0 \, d\theta_1 \\
&\qquad = \frac{1}{|x_{k0} x_{k1}|} \left(\frac{1}{(y_k+1)^{2\beta}} - \frac{1}{y_k^{2\beta}} \right),
\end{aligned}
$$

and

$$
\begin{aligned}
\int_{\beta=0}^{\infty} \frac{1}{|x_{k0} x_{k1}|} \left(\frac{1}{(y_k+1)^{2\beta}} - \frac{1}{y_k^{2\beta}} \right) d\beta &= \frac{1}{|x_{k0} x_{k1}|} \left(\frac{-(y_k+1)^{-2\beta}}{2\log(y_k+1)} + \frac{y_k^{-2\beta}}{2\log(y_k)} \right) \Bigg|_{\beta=0}^{\beta=\infty} \\
&= \frac{1}{2|x_{k0} x_{k1}|} \frac{\log(\frac{y_k+1}{y_k})}{\log(y_k)\log(y_k+1)} < \infty,
\end{aligned}
$$

which completes the proof. Similar derivations can be carried out in the general case of $p > 1$. □

Having proved that the posterior is proper, in the following remark we show that the posterior moments exist and are finite.

Remark 4.1. (Proper moments) Under the same conditions of Theorem (4.1), the posterior distribution of $(\boldsymbol{\theta}, \beta)$ has finite $(m_0, m_1, \ldots, m_p, m_\beta)$ moments, that is

$$
\int_\beta \int_{\boldsymbol{\theta}} \prod_{i=1}^{n} \left\{ \left(\frac{e^{x_i \boldsymbol{\theta}}}{1 + e^{x_i \boldsymbol{\theta}}} \right)^{y_i^\beta} - \left(\frac{e^{x_i \boldsymbol{\theta}}}{1 + e^{x_i \boldsymbol{\theta}}} \right)^{(y_i+1)^\beta} \right\} \theta_0^{m_0} \ldots \theta_p^{m_p} \beta^{m_\beta} \, d\boldsymbol{\theta} d\beta < \infty.
$$

Proof. This proof is similar to the proof of theorem (4.1). Without loss of generality, we consider $p = 1$. Then, for example in the last case, assuming $\theta_j > 0, j = 0, 1$,

$$
\begin{aligned}
&\int_\beta \int_{\theta=(\theta_0,\theta_1)>0} \prod_{i=1}^{n} \left\{ \left(\frac{e^{x_i \theta}}{1 + e^{x_i \theta}} \right)^{y_i^\beta} - \left(\frac{e^{x_i \theta}}{1 + e^{x_i \theta}} \right)^{(y_i+1)^\beta} \right\} \theta_0^{m_0} \theta_1^{m_1} \beta^{m_\beta} \, d\theta d\beta \leq \\
&\qquad \left(\prod_{i=0}^{1} \frac{\Gamma(m_i+1)}{|x_{ki}^{m_i+1)}|} \right) \frac{1}{(m_0+m_1+2)^{m_\beta+1}} \left(\frac{1}{\log^{m_\beta+1}(y_k)} - \frac{1}{\log^{m_\beta+1}(y_k+1)} \right).
\end{aligned}
$$

In general, for $p+2$ parameters $(\theta_0, \ldots, \theta_p, \beta)$ and corresponding moments $(m_0, \ldots, m_p, m_\beta)$ we have,

$$
\begin{aligned}
&\int_\beta \int_{\theta=(\theta_0,\ldots,\theta_p)>0} \prod_{i=1}^{n} \left\{ \left(\frac{e^{x_i \theta}}{1 + e^{x_i \theta}} \right)^{y_i^\beta} - \left(\frac{e^{x_i \theta}}{1 + e^{x_i \theta}} \right)^{(y_i+1)^\beta} \right\} \theta_0^{m_0} \ldots \theta_p^{m_p} \beta^{m_\beta} \, d\theta d\beta \leq \\
&\qquad \left(\prod_{i=0}^{p} \frac{\Gamma(m_i+1)}{|x_{ki}^{m_i+1)}|} \right) \frac{1}{(\sum_{i=0}^{p} m_i + 1)^{m_\beta+1}} \left(\frac{1}{(\log(y_k))^{m_\beta+1}} - \frac{1}{(\log(y_k+1))^{m_\beta+1}} \right),
\end{aligned}
$$

which completes the proof. □

Theorem (4.1) and Remark (4.1) refer to the model with logit link on q and constant β. In fact, the results apply also to the case of log-log link, given Lemma (4.1). In the next sections, we consider empirical results on simulated and real data using non-informative priors. Of course, any proper prior distribution can also be used when prior information is available. In particular, in the next sections, we consider the case of sparsity and variable selection. In this case, we use Laplace priors as defined in Equation 4.1.

4.5 R package

The R package that accompany this chapter provides Bayesian implementation of the discrete Weibull (BDWreg) regression under both transformations, logit and log-log . Estimating the marginal densities is fulfilled in the main function of the package, *bdw*. Several options including arbitrary penalizations via different priors on parameters as well as different hyper priors are implemented in this package. In particular, this package contains the routines to run Reversible Jumps Metropolis Hastings (RJMH) [Green, 1995] for simultaneous model selection and parameter estimation. To take the benefit of multicore processing, a *multicore* routine is implemented in this package. The aim is simultaneously generating several Markov chains that lead to increasing precision of the final results. Two extra functions for producing visual illustrations and summary help diagnostic checking and comparison among different models. This package is publicly available in `https://cran.r-project.org/web/packages/BDWreg/index.html`.

4.6 Simulations study

In this section, we perform a simulation study where we show the effectiveness of the Bayesian estimation procedure, both in the case of data drawn from a DW regression model and in the case of model misspecification, where the generating model is that of Poisson or Negative Binomial (NB). Finally, we test the use of Laplace priors in a variable selection scenario.

4.6.1 Simulation from a DW regression model

Table (4.1) shows six configurations of parameters used in the simulation, where we consider the two link functions for q, and the link function for β described in Section § 4.2, i.e. imposing a linear model on logit (q) or log-log (q), and on $\log(\beta)$. We choose the regression and distribution parameters in such a way to obtain different shapes of the distribution. For Case 2 to 6, we generate 500 observations from

Case	Model	True Parameters	
1	$DW(q, \beta)$	$q = .41$	$\beta = 1.1$
2	$DW(q,\ \log : reg\beta)$	$q = .8$	$\gamma_0 = .1\ ,\ \gamma_1 = -.15\ ,\ \gamma_2 = .5$
3	$DW(\text{logit}\ : regQ, \beta)$	$\theta_0 = .4\ ,\ \theta_1 = -.1\ ,\ \theta_2 = .34$	$\beta = .7$
4	$DW(\text{logit}\ : regQ,\ \log : reg\beta)$	$\theta_0 = .4\ ,\ \theta_1 = -.1\ ,\ \theta_2 = .34$	$\gamma_0 = .1\ ,\ \gamma_1 = -.15\ ,\ \gamma_2 = .5$
5	$DW(\text{log-log}\ : regQ, \beta)$	$\theta_0 = .4\ ,\ \theta_1 = -.1\ ,\ \theta_2 = .34$	$\beta = .7$
6	$DW(\text{log-log}\ : regQ,\ \log : reg\beta)$	$\theta_0 = .4\ ,\ \theta_1 = -.1\ ,\ \theta_2 = .34$	$\gamma_0 = .1\ ,\ \gamma_1 = -.15\ ,\ \gamma_2 = .5$

TABLE 4.1: The configuration of DW regression models used in the simulations.

uniform distribution $U(-1.5, 1.5)$ for each predictor. For the Bayesian estimation of the parameters,

we use non-informative priors, $p(x) \propto 1$, and make use of a Metropolis-Hastings algorithm with an independent Gaussian proposal to draw samples from the posterior. The scale of the proposal is adjusted so that a recommended acceptance rate lies in the interval $(22, 25)\%$ [Bedard, 2008]. We consider 25,000 iterations of the sampler that is far more than enough for some cases, e.g. Case 1, and remove the first 25% of the chains for burn-in.

Figure (4.1) shows the posterior distribution of the parameters and the chain convergence in the first case, when no exogenous variables are presented. Similar plots are obtained for the other cases. From this figure, the sampler shows a promising mixing and rapid convergence as confirmed by the chains and sample ACF plot respectively. Similar plots are obtained for the other cases. Figure (4.2) shows the marginal densities of the parameters and the 95% HPD interval for all six cases, as well as the maximum likelihood point estimate and the true value of the parameters. Overall, the plots show convergence of the chains and accurate estimation for the parameters.

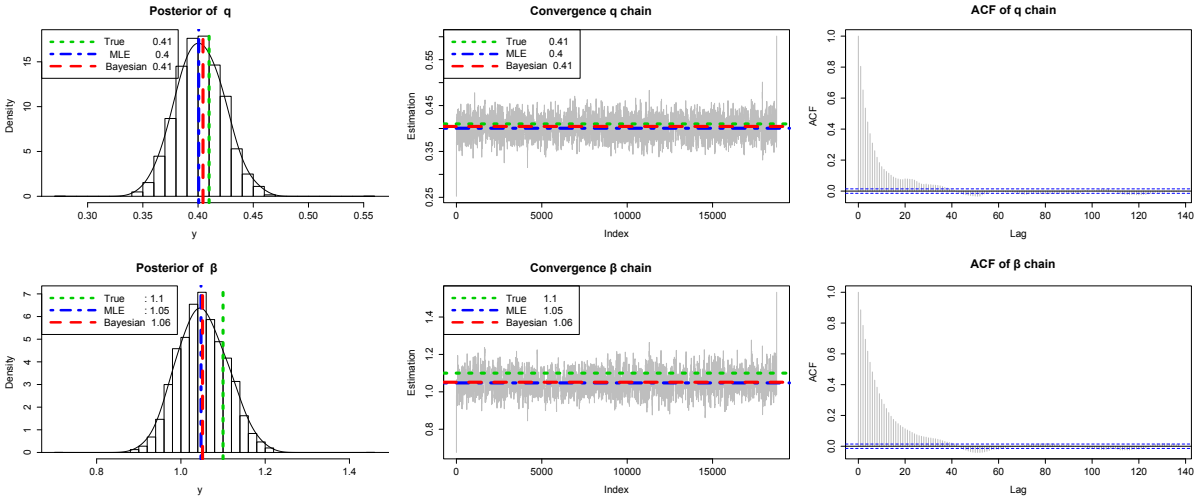

FIGURE 4.1: Marginal densities and chain convergence for q (top) and β (bottom), for Case 1 where there are no exogenous variables in model.

4.6.2 Simulation from a Poisson and NB regression model

The aim of this section is to test the fitting of a DW regression model to data generated from a Poisson and NB regression. To this end, we design two experiments using two explanatory variables, $x = (x_1, x_2)$, and $n = 500$ data points. We simulate data for the predictors from uniform distributions, namely $x_1 \sim U(0, 1)$ and $x_2 \sim U(0, 1.5)$. In the first experiment, we assume that the conditional distribution of y given x is $\text{Poisson}(e^{X\alpha})$, whereas in the second experiment, we assume it to be a NB distribution with mean $\mu = e^{X\alpha}$ and variance $\mu + \mu^2/\theta$ with $\theta = 4.5$. We fix the intercept and the regression parameters to $\alpha = (-0.5, 4.3, -2.2)$, with values chosen to cover a wide range of shapes for the target distribution. Figures (4.3) shows the conditional distribution fitted by $DW(regQ, \beta)$ for a fixed value of $x_1 = 0.5$ and sliding values of x_2 in the $[0, 0.7]$ interval. The figure shows how the estimation improves as the mean of the target distribution decreases, both for Bayesian and frequentist approaches. In addition, the logit link shows a better fit compared to the log-log link in both Poisson and NB experiments that can be

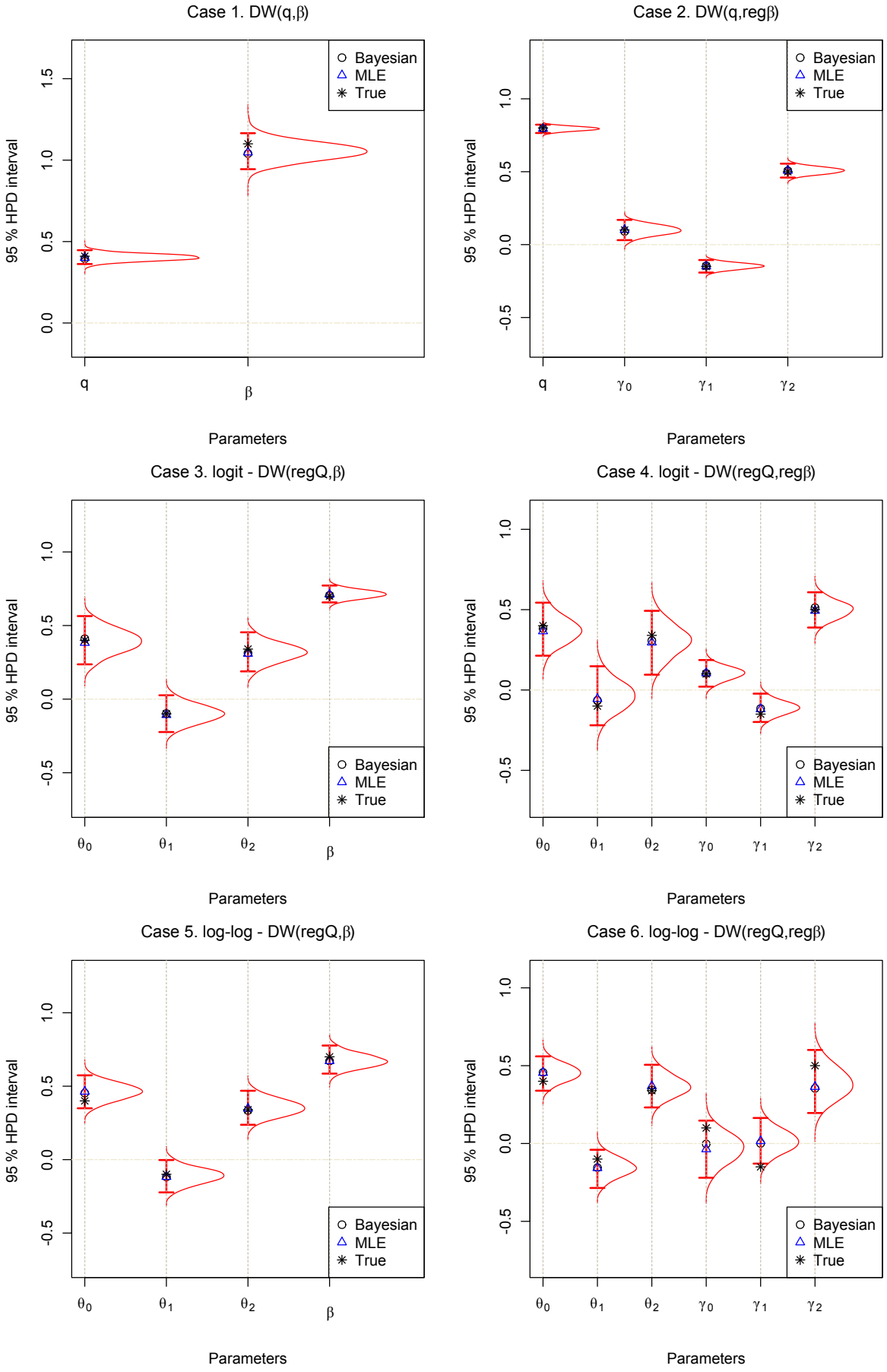

FIGURE 4.2: Marginal densities and 95% high probability density interval for Cases 1-6 in Table (4.1).

a consequence of numerical instability in log-log transformation. For the frequentist estimation, we use the R package `DWreg` [Kalktawi et al., 2016].

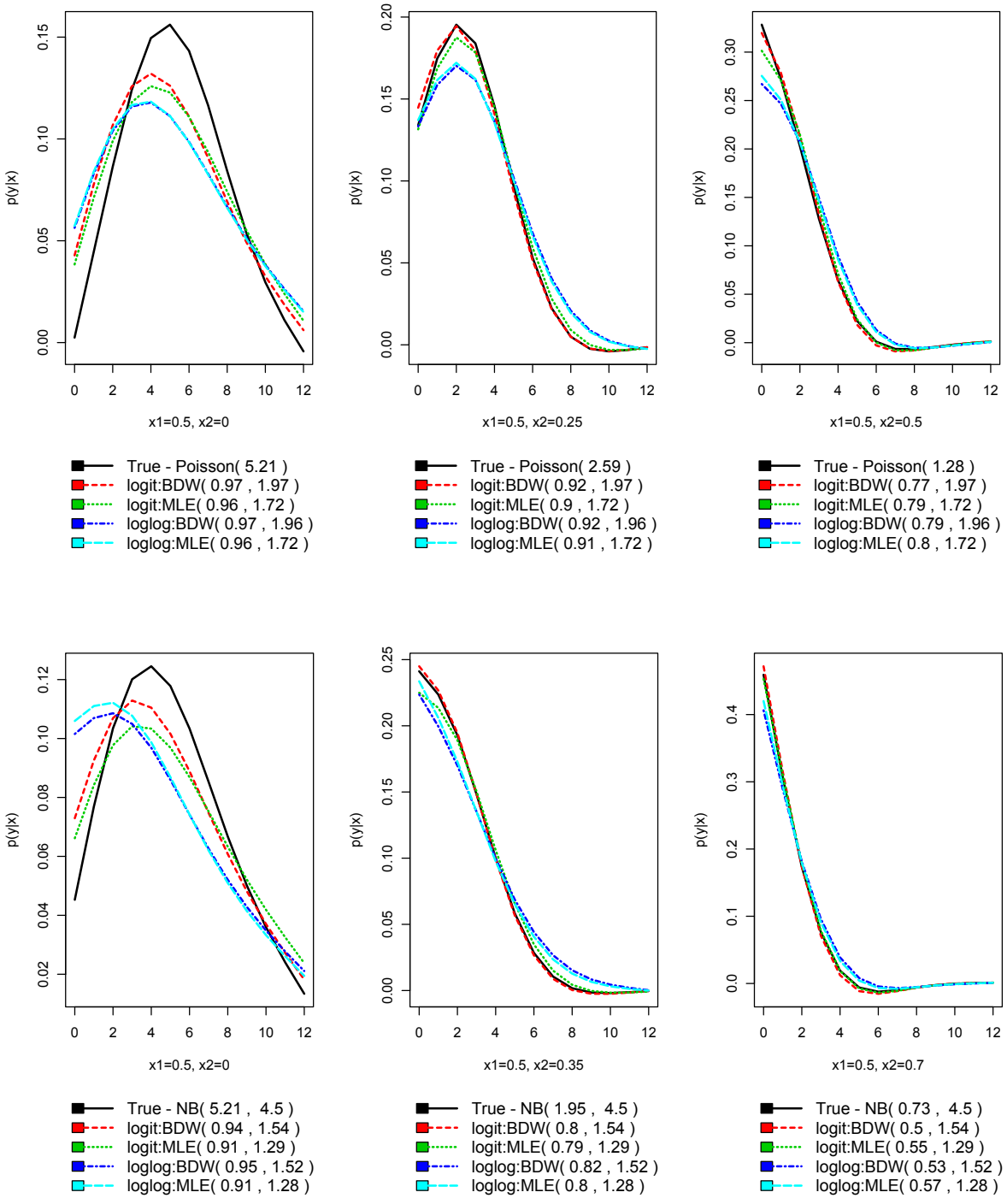

FIGURE 4.3: Fitting Poisson (top) and NB (bottom) by $DW(regQ, \beta)$ for a range of values of x_2 and fixed $x_1 = 0.5$. The plots show the true conditional pmf (black) together with the conditional pmf fitted by the Bayesian DW model proposed in this chapter, with the logit (q) (red) and log-log (q) (blue) links, and by the corresponding frequentist approaches (green and light blue, respectively).

4.6.3 Simulation on Variable Selection

In this simulation, we show the performance of DW regression for variable selection. To this end, we consider a simulation with 50 predictors and assume that 75% of the parameters, 37 out of 50, are zero. We generate the remaining non-zero parameters uniformly in the $[-0.5, 0.5]$ interval. We simulate 500 observations for each predictor from a $U(-1.5, 1.5)$ distribution, and the response variable from a DW distribution using a logit link for q or the log link for β. Similar results are obtained with the log-log link function. For parameter estimation, we keep the average rate of acceptance in the $(20, 30)$% interval for a total of $50,000$ iterations. We choose an InverseGamma(2,1) hyper prior for the penalty parameters. This prior allows to cover a large range of penalties, in the (0.02,70) interval, with a tendency to small penalties (i.e. sparsity) due to a mean of 1 and a median of 0.5. Variable selection is performed by considering the 95% HPD interval for each parameter.

Table (4.2) shows the performance of the method in terms of selection of variables under six different generating models. In particular, the table reports the True Negative Rate (TNR), Recall $\left(\frac{\text{TP}}{\text{TP}+\text{FN}}\right)$, Precision $\left(\frac{\text{TP}}{\text{TP}+\text{FP}}\right)$ and F_1 score $\left(\frac{2\text{TP}}{2\text{TP}+\text{FN}+\text{FP}}\right)$, averaged over 20 simulations. The table shows a good performance overall, particularly for the $BDW(regQ, \beta)$ models. The model with the $\log(\beta)$ link does not perform very well when q decreases, i.e. when the number of zeros in the sample increases. In these cases, the models show a low recall, that is a high false negative rate.

Model	**Avr.TNR**	**Avr.Recall**	**Avr.Precision**	**Avr.F_1**
$BDW(\text{logit} : regQ, \beta = .1)$	93%	90%	93%	91%
$BDW(\text{logit} : regQ, \beta = .8)$	95%	89%	95%	92%
$BDW(\text{logit} : regQ, \beta = 1.6)$	93%	91%	93%	92%
$BDW(\text{logit} : regQ, reg\beta)$	97%	68%	96%	79%
$BDW(q = .85, reg\beta)$	90%	92%	91%	91%
$BDW(q = .50, reg\beta)$	93%	37%	84%	52%

TABLE 4.2: Performance of BDW with Laplace priors. Variables are selected based on the 95% HPD interval and the selection is compared to the truth on the basis of the average True Negative Rate (TNR), recall, precision and F_1 score.

4.7 Real data illustration

In this section, we show the performance of the Bayesian discrete Weibull regression model on real datasets from the medical domain. We compare the proposed model with the Bayesian Poisson (BPoisson), Bayesian Negative Binomial (BNB) and in the case of excessive zeros with Bayesian zero-inflated Poisson (BZIP) and negative binomial (BZINB) on the basis of a number of commonly used criteria: BIC [Dayton, 2003], AIC [Dayton, 2003], Deviance Information Criterion (DIC) [Spiegelhalter et al., 2002], Quasi-likelihood Information Criteria (QIC) [Pan, 2001], Consistent AIC (CAIC) [Bozdogan, 1987], Bayesian Predictive Information Criterion (BPIC) [Ando, 2007] and the Prior Predictive Density (PPD) used in the Bayes factor [Kass, 1993].

Since BIC, AIC, QIC and CAIC are commonly used in the frequentist framework, we adopt them to the Bayesian framework by estimating the parameters using the *mode* of the marginal densities of the parameters that corresponded to MLE estimations. Next HPD interval is applied to identifying the

insignificant (zero) parameters. Then the likelihood is re-estimated and degree of freedom is defined by the number of non-zero estimations.

4.7.1 Comparison with Bayesian generalised linear models

To show the ability of BDW to estimate parameters in the presence of under-dispersion, over-dispersion and excessive zeros in count data, we choose the following three medical datasets:

1. The data on inhaler usage from [Grunwald et al., 2011], with 5209 observations. The response is the daily counts of inhalers usage, whereas the covariates are humidity, barometric pressure, daily temperature, air particles level. The sample mean and variance of the data are 1.3 and 0.8 respectively, so this is a case of under-dispersion relative to Poisson [Kalktawi et al., 2016].

2. The German health survey dataset available in the R package `COUNT` under the name `badhealth`, with 1127 observations. The response is the number of visits to doctors during the year 1988 and the predictors are whether the patient claims to be in bad health or not, and the age of the patient. The response variable ranges from 0 to 40 visits and has a sample mean of 2.4 and variance of 12, suggesting over-dispersion relative to Poisson.

3. The German health registry dataset available in the R package `COUNT` under the name `rwm`, with 27326 observations. The response is the number of visits to doctors for the years 1984-1988 and the predictors are age, years of education and household yearly income. The response variable, number of visits, has about 37% of zeros, a sample mean of 3.2 and a variance of 32.4, pointing again to a case of over-dispersion and excessive zeros.

We fit a BDW model with a non-informative prior on the regression parameters, 25,000 iterations for the Metropolis-Hastings algorithm and an acceptance rate in the $(20, 30)$% interval. For the case of BPoisson, BNB, BZIP and BZINB, we make use of the `MCMCpack` R package [Martin et al., 2011] using the same configurations as with our approach. Table (4.3) shows a comparison of the models on the three datasets. We only report the results of the BDW(regQ,β) models, which show superior performance to the other BDW models on these datasets. Of the two links on q, the logit (q) link performs better than the log-log (q) link with respect to BIC, AIC, CAIC, DIC, BPIC and log(PPD) in Table (4.3). As for a comparison with the other models, Poisson has the worst performance for all cases, while NB has a performance comparable to the logit -BDW model in the over-dispersed scenario, while it does not perform well in the under-dispersed and excessive zero scenario. In the latter case, zero-inflated negative Binomial has a performance comparable to DW. This is promising and it points to a future extension of DW to a zero-inflated DW model.

4.7.2 Comparison with Bayesian penalised regression

In this section, we compare the performance of BDW to BPoisson and BNB regression for variable selection on a dataset with several variables. In particular, we consider the multivariate data of [Machado and Santos Silva, 2005]. The data consist of 5096 observations from the 1985 wave of the German Socioeconomic Panel. As in [Machado and Santos Silva, 2005], we measure the demand in healthcare

by the number of visits to a specialist (except gynecology or pedriatics) in the last quarter. The 20 covariates are listed in full in Table (4.4) and they are the same considered in [Machado and Santos Silva, 2005]. This is an extreme example of excessive zeros as the response variable contains 67.82% of zeros.

We fit a BDW model with a Laplace prior on the regression parameters and an InverseGamma(2,1) hyper-prior on the shrinkage parameters. We consider 175000 iterations for the MCMC routine and similar configurations for the Bayesian Poisson and NB models. We also extend the comparison by including Bayesian zero-inflated models and frequentist L_1 regularized models. For the latter, we use the `glmnet` package [Friedman et al., 2010] to fit regularized Poisson regression and the `glm.nb` R function to fit regularized negative Binomial regression. In both cases, the penalty parameter is chosen by BIC. According to the results in Table (4.5), $DW(regQ, \beta)$ with the log-log link achieves overall the best performance compared with the others BDW models and with NB and Poisson models.

Figure (4.4) shows the marginal densities of the parameters for the $DW(regQ, \beta)$ with the log-log link. Highlighted in red are those variables that are found to be significant based on the 95% HPD interval. The selection is overall in accordance with the results obtained by [Machado and Santos Silva, 2005] using a jittering approach, with variables such as gender, chronic complaints, sick leave and disability found to be significant, and other variables like unemployment, private insurance and those related to job characteristics, such as heavy labor, stress, variety on job, self-determined and control found not to be significant. Figure (4.5) shows the effect of the variable chronic complaints on the conditional distribution, suggesting that the probability of a large number of visits is higher for the case of chronic complaints than for the case of no complaints. Table (4.6) further compares the selection of variables with those selected by Poisson and NB regression models. Overall, there is high agreement between DW and NB, with the exception of the variable control which is found significant by NB (both in the Bayesian and frequentist estimation) but not by DW. Poisson and BPoisson tend to select many more variables.

4.8 Conclusion remarks

In this chapter we have proposed a novel Bayesian regression model for count data, by assuming a discrete Weibull conditional distribution. A discrete Weibull regression model was originally proposed by [Kalktawi et al., 2016] in a frequentist context and a number of desirable features of the model compared to existing ones were highlighted. The Bayesian implementation in this chapter is based on a more general model, where both parameters can be linked to the predictors. We have experimented with different link functions and have found the models with the link on q and constant β to work particularly well, with the logit (q) link displaying superior performance than the log-log (q) link in the simulations and in three of the four applications considered in this chapter. Including a link to both q and β was found to lead to over-parametrization for the applications considered, but it may be useful for other applications showing more complex dependencies. In terms of the Bayesian inferential approach, we have shown theoretically how the posterior is proper and with finite moments under a uniform non-informative prior on the parameters.

We have shown the applicability of the Bayesian discrete Weibull model to count data from the medical domain. In particular, we have analysed datasets on the number of visits to doctors/specialists, a quantity that is often used as an indicator of healthcare demand. The response variable in the examples considered

Model	AIC	BIC	CAIC	QIC	DIC	BPIC	log(PPD)	df
Inhaler Use (under-dispersed)								
log-log : *BDW*	13497.22	13536.57	13542.57	2.59*	13487.63	13493.88	-6745.93	6
logit : *BDW*	13494.19*	13533.54*	13539.54*	2.59*	13484.92*	13490.49*	-6739.41*	6
BPoisson	14009.01	14041.80	14046.80	2.69	13822.54	13734.31	-6960.64	5
BNB	13952.85	13992.33	13998.20	2.68	13771.0	13686.47	-6960.81	6
German Health Survey (over-dispersed)								
log-log : *BDW*	4478.9	4499.0	4502.0	3.98	4474.60	4478.33	-2245.75	4
logit : *BDW*	4475.2*	4495.3*	4449.3*	3.97*	4474.16*	4477.70*	-2242.23*	4
BPoisson	5638.9	5654.02	5656.10	5.01	5638.14	5641.18	-2826.88	3
BNB	4475.9	4495.9	4499.97	3.97*	4474.66	4478.10	-2243.87	4
German Health Registry (excessive zeros)								
log-log : *BDW*	120340.1	120381.2	120386.2	4.4*	120334.6	120339.2	-60187.6	5
logit : *BDW*	120339.2*	120380.3*	120385.3*	4.4*	120327.0*	120331.9*	-60181.8*	5
BPoisson	209636.4	209669.2	209673.2	7.7	209635.8	209639.6	-104836.7	4
BNB	120658.7	120708.0	120714.0	4.4*	129125.8	133365.3	-60344.0	5
BZIP	169417.7	169450.6	169454.6	6.2	169402.1	169398.3	-83522.3	5
BZINB	120649.5	120682.4	120686.4	4.4*	120629.1	120622.9	-60245.2	6

TABLE 4.3: Comparison of Bayesian DW, Poisson, Zero-Inflated Poisson, Negative Binomial and Zer-Inflated Negative Binomial on three datasets and under a number of information criteria. (*) denotes the minimum value.

Variable	Description
Age	Age in decades
Chronic complaints	1 if has chronic complaints for at least 1 year
Control	1 if has a job where work performance is strictly controlled
Degree of disability $> 20\%$	1 if the degree of disability is greater than 20%
Education	Number of years in education after age 16
HH-income	Net monthly household income
Hospitalized > 7 days	1 if was more than 7 days hospitalized in the previous year
Marital Status	1 if single
Month of unemployment	Number of months of unemployment in the previous year
Physically heavy labour	1 if has a job in which physically heavy labour is required
Physician density	Number of physicians per 100,000 inhabitants in the place of residence
Population < 5000	1 if place of residence has less than 5,000 inhabitants
Population 20000-100000	1 if place of residence has between 20,000 and 100,000 inhabitants
Population 5000-20000	1 if place of residence has between 5,000 and 20,000 inhabitants
Private insurance	1 if had private medical insurance in the previous year
Self-determined	1 if has a job where the individual can plan and carry out job tasks
Sex	1 if female
Sick leave > 14 days	1 if missed more than 14 work days due to illness in the previous year
Stress	1 if has a job with high level of stress
Variety on job	1 if job offers a lot of variety

TABLE 4.4: List of the variables and descriptions in *the number of visits to a specialist* dataset [Machado and Santos Silva, 2005].

Model	AIC	BIC	CAIC	QIC	DIC	BPIC	log(PPD)	df
logit :BDW(regQ,β)	12720.4	12864.2	2.5*	12886.2	12710.8	12731.5	-6392.3	11
log-log :BDW(regQ,β)	12698.5*	12842.3*	2.5*	12864.3*	12693.3*	12713.6*	-6383.3*	11
BDW(q,regβ)	13256.0	13399.8	2.6	13421.8	13250.4	13270.3	-6665.8	6
logit :BDW(regQ,regβ)	12750.3	12920.7	2.5*	12951.7	12713.0	12744.9	-6516.3	19
log-log :BDW(regQ,regβ)	12748.9	12924.1	2.5*	12955.2	12715.4	12741.3	-6519.1	19
BPoisson	21588.2	21705.8	4.2	21723.8	21594.6	21615.8	-10832.6	17
BNB	12867.3	12939.2	2.5*	12950.2	12838.3	12834.8	-6452.3	11
BZIP	16677.1	16760.1	3.2	16760.0	16698.7	16720.6	-8385.6	15
BZINB	12850.0	12921.1	2.5*	12932.2	12872.0	12894.1	-6456.8	13
Poisson (glmnet)	21571.1	21706.1	4.2	21724.1	-	-	-	17
NB (glm.nb)	12839.3	12911.2	2.5*	12922.6	-	-	-	12

TABLE 4.5: Comparison of BDW with Bayesian and regularized NB and Poisson on the number of visits to a specialist dataset of [Machado and Santos Silva, 2005]. (∗) denotes the minimum value, whereas df is the number of non-zero coefficients. For the Bayesian models, these are based on the 95% HPD interval.

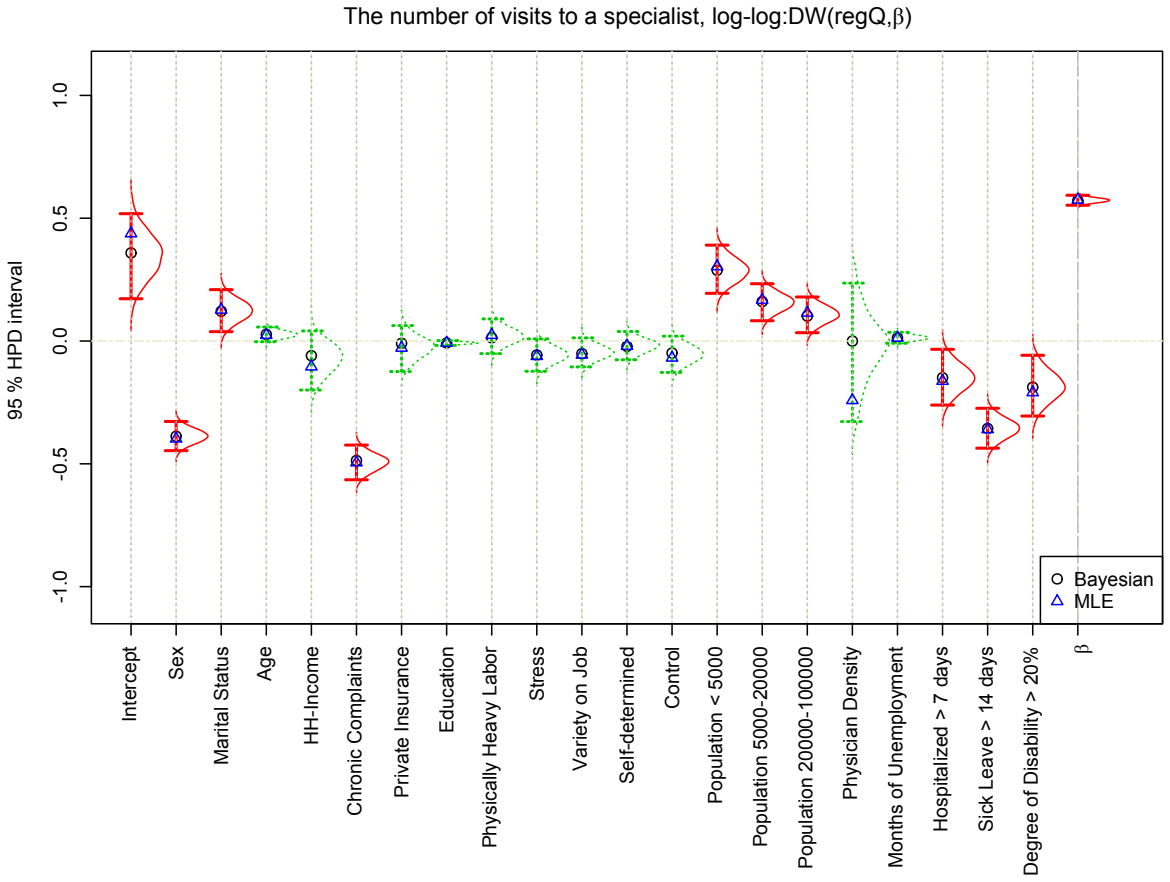

FIGURE 4.4: Marginal densities of the parameters for the $BDW(reg Q, \beta)$ model with log-log (q) link on the number of visits to a specialist dataset. The red lines are for the cases where the 95% HDP interval does not contain zero (significant variable). Green dotted lines for the opposite.

Variable	**BDW(regQ,β)**	**NB**	**BNB**	**BZINB**	**Poisson**	**BPoisson**	**BZIP**
Sex	*	*	*	*	*	*	*
Marital status	*	*	*		*	*	*
Age					*		
HH-income					*	*	
Chronic complaints	*	*	*	*	*	*	*
Private insurance							
Education					*		*
Physically heavy labour					*	*	*
Stress					*	*	*
Variety on job					*	*	*
Self-determined							
Control		*	*	*	*	*	*
Population $<$ 5000	*	*	*	*	*	*	*
Population 5000-20000	*	*	*	*	*	*	
Population 20000-100000	*	*		*	*	*	*
Physician density						*	
Months of unemployment			*	*		*	
Hospitalized $>$ 7 days	*	*	*	*	*	*	*
Sick Leave $>$ 14 days	*	*	*	*	*	*	*
Degree of disability $>$ 20	*	*	*	*	*	*	

TABLE 4.6: Significant (non-zero) covariates that are selected by $BDW(regQ, \beta)$ with log-log link, Bayesian and regularized *NB* and Poisson regression models, and Bayesian zero-inflated Poisson and NB, for the number of visits to a specialist dataset. An $(*)$ indicates a non-zero coefficient.

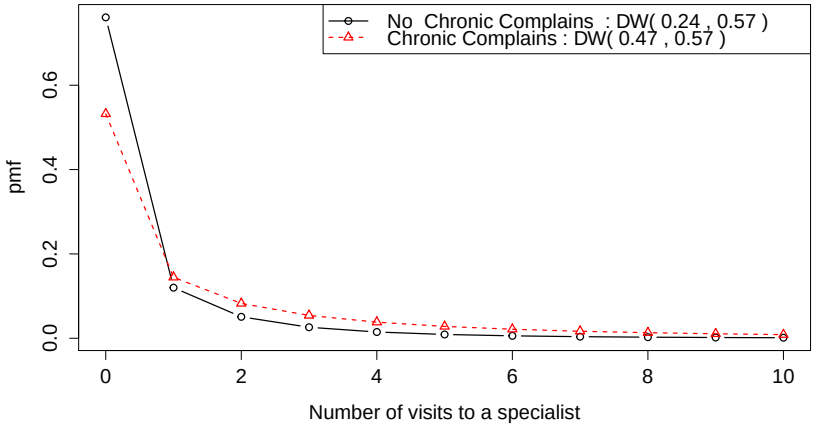

FIGURE 4.5: Effect of the variable Chronic Complaints on the conditional distribution for the healthcare data, when all other variables are held constant.

is discrete and is characterized by a skewed distribution, making the whole conditional distribution of interest and not only the conditional mean. We have tested the inference procedure on simulated and real data with various characteristics, such as under-dispersion, over-dispersion and excess of zeros. Overall, we have found a good performance of the method in comparison with Poisson and NB regression models, on the basis of a number of information criteria and of the selection of influential variables.

4.8.1 Future study

Future work will explore an extension of the approach proposed in this chapter to more flexible DW regression models, such as zero-inflated, multilevel and mixture DW models, in a similar spirit to the existing models for continuous responses [Dunson et al., 2007]. Moreover, an interesting topic for future work can be performing the Bayesian optimization methods such as simulated annealing [Hwang, 1988] for estimating the DW regression parameters. We expect that using the maximum posterior methods (MAP) can lead to exactly zero estimation of the parameters.

Chapter 5

Conclusions

This thesis has focused on the problem of variable selection and parameter estimation within both frequentist and Bayesian frameworks. Clear advantages over existing methods include proposing a penalized estimation of the parameters in a time-dependent framework in Chapter § 2, proposing a differentiable penalty term and discussing its theoretical properties in Chapter § 3, and proposing Bayesian solution to discrete response regression under discrete Weibull distribution in Chapter § 4. The main contributions are listed below.

5.1 Main Contributions

1 A l_1 penalized approach for estimating parameters in the dynamic regression in the presence of autocorrelated residuals is proposed in Chapter § 2. This model is extensively discussed from both a theoretical and practical point of view and by means of simulations and real data applications. The proposed two-step algorithm for estimating the parameters shows promising results and low bias. Comparisons with the existing methods in the literature have also shown that this model is beneficial in real applications. An R package that accompanies this chapter provides an implementation of the method, for other statisticians and practitioners. The inferential procedure presented in this chapter can be extended to cover a wider range of penalty functions, including l_2 and a combination of l_1 and l_2 penalties.

2 A fully differentiable and novel penalty is proposed in Chapter § 3. This penalty allows to go from a flat (OLS) to a very sharp (lasso) regularization of the parameters. We have discussed this novel penalty theoretically, including asymptotic properties in regularized linear regression, and practically by proposing an efficient algorithm as well as preparing an R package. We have discussed this penalty from a computation point of view, and proposed a simple approximation for it. Simulation studies and real data applications confirm the advantage of this penalty over the competitors in the literature. We keep the research open by proposing a more general form of this penalty as well as exploring different scenarios where the differentiability of the penalty function is particularly advantageous.

3 Chapter § 4 of this thesis has addressed the problem of regression for discrete response within a Bayesian framework. We have introduced the discrete Weibull distribution and the corresponding regression model. Two link functions are proposed for connecting the distribution parameters to exogenous variables, precisely a log-log and logit transformations. Then a fully Bayesian approach is applied in order to estimate the posterior conditional distribution of the response given the covariates and over a range of priors. We have proved that under the non-informative prior, the DW posterior and the moments are bounded. In other words, we have proved that the DW posterior is proper under the non-informative prior for the parameter and for any moment. We have shown the usefulness of the new regression model in a number of applications and by comparing it to the existing methods in the literature. The results of this chapter show that imposing independent Laplace priors on the parameters encourage shrinking parameters toward zero that by itself leads to model selection in frequency domain. The former is discussed numerically using simulations and in a number of real data applications. Further, the R package that accompanies this chapter provides a Bayesian solution to the problem of DW regression. The provided R package covers the whole contents that are discussed in this chapter and some extras for implementing RJMCMC and parallel processing. The latter results in a significant improvement in estimations by simultaneously estimating several Markov chains at the same time.

Appendix A

Asymptotic properties of non-penalized DREGAR

A.0.1 Asymptotic properties of non-penalized DREGAR

In this section we focus on the limiting distribution of OLS estimations in a non-penalized DREGAR given $T > r + p + q$. In particular, this section provides essential proofs that are used in Chapter § 2 of the thesis.

We start with the product of two infinite geometric series.

Lemma A.1 (Product of two infinite geometric series). Let $S_1 = \sum_{i=0}^{\infty}(\frac{L}{a})^i, a > 1$ and $S_2 = \sum_{i=0}^{\infty}(\frac{L}{b})^i, b > 1$ be two geometric series, then $S_1 S_2$ is a linear function of S_1 and S_2, provided $a \neq b$.

Proof.

$$
\begin{aligned}
S_1 S_2 &= \sum_{i=0}^{\infty}(\frac{L}{a})^i \sum_{i=0}^{\infty}(\frac{L}{b})^i \\
&= \sum_{i=0}^{\infty}\sum_{k=0}^{i}(\frac{L}{a})^k(\frac{L}{b})^{i-k} \\
&= \sum_{i=0}^{\infty}(\frac{L}{b})^i \sum_{k=0}^{i}(\frac{b}{a})^k \\
&= \sum_{i=0}^{\infty}(\frac{L}{b})^i(\frac{1-(\frac{b}{a})^{i+1}}{1-(\frac{b}{a})}) \\
&= \frac{a}{a-b}\sum_{i=0}^{\infty}(\frac{L}{b})^i + \frac{b}{b-a}\sum_{i=0}^{\infty}(\frac{L}{a})^i \qquad a \neq b.
\end{aligned}
$$

□

By induction one can extend the result of Lemma (A.1) to m geometric series, $S_1S_2S_3\dots S_m$. The rest of this section is focused on the block matrices in H_1 and limit distribution of $H_2'e$ in equation (2.5),

$$H_1 = \left(\frac{1}{n}(X', H_{(p)}, H_{(q)})'(X', H_{(p)}, H_{(q)})\right)^{-1}$$
$$H_2'e = \frac{1}{\sqrt{n}}(X', H_{(p)}, H_{(q)})'e.$$

Recall the general form of (non-penalized) DREGAR(p,q),

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + x_t'\beta + \epsilon_t$$
$$\epsilon_t = \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + e_t.$$

Using backward shift operator,

$$(1 - \sum_{i=1}^{p} L^i\phi_i)y_t = x_t'\beta + \epsilon_t \to y_t = \frac{1}{A}x_t'\beta + \frac{1}{A}\epsilon_t$$
$$(1 - \sum_{l=1}^{q} L^l\theta_l)\epsilon_t = e_t \to \epsilon_t = \frac{1}{B}e_t,$$

where $A = (1 - \sum_{i=1}^{p} L^i\phi_i)$ and $B = (1 - \sum_{i=1}^{q} L^i\theta_i)$.
From H_1, $\frac{1}{n}H_{(p)}'H_{(q)}$ is

$$\frac{1}{n}H_{(p)}'H_{(q)} = \frac{1}{n}\begin{pmatrix} \sum_{i=T_0+1}^{T} y_{i-1}\epsilon_{i-1} & \sum_{i=T_0+1}^{T} y_{i-1}\epsilon_{i-2} & \dots & \sum_{i=T_0+1}^{T} y_{i-1}\epsilon_{i-q} \\ \sum_{i=T_0+1}^{T} y_{i-2}\epsilon_{i-1} & \sum_{i=T_0+1}^{T} y_{i-2}\epsilon_{i-2} & \dots & \sum_{i=T_0+1}^{T} y_{i-2}\epsilon_{i-q} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=T_0+1}^{T} y_{i-p}\epsilon_{i-1} & \sum_{i=T_0+1}^{T} y_{i-p}\epsilon_{i-2} & \dots & \sum_{i=T_0+1}^{T} y_{i-p}\epsilon_{i-q} \end{pmatrix}_{p\times q},$$

where each element of this matrix comes from $\frac{1}{n}\sum_{i=T_0+1}^{T} y_{i-s_1}\epsilon_{i-s_2}$ for $s_1 \in \{1,2,3,\dots,p\}$ and $s_2 \in \{1,2,3,\dots,q\}$. Thus,

$$\frac{1}{n}\sum_t y_{t-s_1}\epsilon_{t-s_2} = \frac{1}{n}\sum_t (\frac{L^{s_1}}{A}x_t'\beta + \frac{L^{s_1}}{A}\epsilon_t)L^{s_2}\epsilon_t$$
$$\text{where}: \epsilon_t = \frac{1}{B}e_t.$$

Then,

$$\begin{aligned}\frac{1}{n}\sum_t y_{t-s_1}\epsilon_{t-s_2} &= \frac{1}{n}\sum_t (\frac{L^{s_1}}{A}x_t'\beta + \frac{L^{s_1}}{A}\frac{1}{B}e_t)L^{s_2}\frac{1}{B}e_t \\ &= \frac{1}{n}\sum_t (\frac{L^{s_1}}{A}x_t'\beta)(\frac{L^{s_2}}{B}e_t) + \frac{1}{n}\sum_t (\frac{L^{s_1}}{A}\frac{1}{B}e_t)(\frac{L^{s_2}}{B}e_t). \end{aligned} \tag{A.1}$$

$x_{.i}, i = 1,2,\dots,r$ are independent of the error term by the assumptions, then the first term in (A.1) tends to have zero mean. If $\mathbb{E}\{(\frac{L^{s_1}}{A}x_t')(\frac{L^{s_1}}{A}x_t)\} < \infty$, then the variance of the first term in (A.1)

tends to zero at the speed of n. Consequently, the first term in (A.1) tends to zero, $o_p(1)$. Moreover, $\forall i \in \{1,2,3,\ldots,r\}, x_{.i}$s are mutually independent and covariance stationary with finite moments and in particular satisfying the conditions in Theorem (2.1). From this, $\mathbb{E}\{(\frac{L^{s_1}}{A}x_t')(\frac{L^{s_1}}{A}x_t)\} < \infty$.

For the second term, by assumptions $t \neq t'$, $\mathbb{C}ov(e_t, e_{t'}) = 0$. As a result, this term is non-zero in similar orders of $(\frac{L^{s_1}}{A}\frac{t}{B})$ and $(\frac{L^{s_2}t}{B})$. Precisely, let $a_1, a_2, a_3, \ldots, a_p$ and $b_1, b_2, b_3, \ldots, b_q$ be the roots for $1 - \sum_{i=1}^{p}\phi_i L^i$ and $1 - \sum_{i=1}^{q}\theta_i L^i$ respectively. Then,

$$\frac{L^{s_1}}{AB} = \left(\prod_{j=1}^{p}\sum_{i=0}^{\infty}(\frac{L}{a_j})^i\right)\left(\prod_{k=1}^{q}\sum_{i=0}^{\infty}(\frac{L}{b_k})^i\right)L^{s_1} \tag{A.2}$$

$$\frac{L^{s_2}}{B} = \left(\prod_{j=1}^{q}\sum_{i=0}^{\infty}\frac{1}{b_j}(\frac{L}{b_j})^i\right)L^{s_2}. \tag{A.3}$$

Using remark (A.1), given $i, i' \in \{1,2,3,\ldots,p\}$ and $j, j' \in \{1,2,3,\ldots,q\}; a_i \neq a_{i'}, b_j \neq b_{j'}$ and $a_i \neq b_j$, then all terms in RHS of (A.2) and (A.3) can be rewritten in the form of linear functions of individual elements. As a result, there are infinite terms with similar orders between (A.2) and (A.3), provided the process is started from $-\infty$. Consequently, the second term in the RHS of equation (A.1) is a non-zero function of σ^2. Clearly, the model orders p, q do not affect (A.2) and (A.3), provided there are enough observations. On the other hand, if the process starts at zero, then, $[H'_{(p)}H_{(q)}]_{s_1 s_2} = 0$ for $\min(s_1, s_2) \geq n$. As a result, p and/or q can freely tend to infinity, provided H_1 is non-singular and there are enough observations.

For a simple example, let the true underlying model be DREGAR(1,1). Then,

$$(A.2) = L^{s_1}(\sum_{i=0}^{\infty}\phi^i L^i)(\sum_{i=0}^{\infty}\theta^i L^i)$$
$$(A.3) = L^{s_2}(\sum_{i=0}^{\infty}\theta^i L^i).$$

If $\theta \neq \phi$,

$$\begin{aligned}(A.2) &= L^{s_1}(\sum_{i=0}^{\infty}\phi^i L^i)(\sum_{i=0}^{\infty}\theta^i L^i)\\ &= L^{s_1}\left((\frac{\phi}{\phi-\theta})\sum_{i=0}^{\infty}\phi^i L^i + (\frac{\theta}{\theta-\phi})\sum_{i=0}^{\infty}\theta^i L^i\right)\\ (A.3) &= L^{s_2}(\sum_{i=0}^{\infty}\theta^i L^i).\end{aligned}$$

Then,

$$\begin{aligned}&\mathbb{C}ov\left(L^{s_1}\left((\frac{\phi}{\phi-\theta})\sum_{i=0}^{\infty}\phi^i L^i + (\frac{\theta}{\theta-\phi})\sum_{i=0}^{\infty}\theta^i L^i\right)e_t, L^{s_2}(\sum_{i=0}^{\infty}\theta^i L^i)e_t\right) =\\ &\mathbb{C}ov\left(L^{s_1}(\frac{\phi}{\phi-\theta})\sum_{i=0}^{\infty}\phi^i L^i e_t, L^{s_2}(\sum_{i=0}^{\infty}\theta^i L^i)e_t\right) +\\ &\mathbb{C}ov\left(L^{s_1}(\frac{\theta}{\theta-\phi})\sum_{i=0}^{\infty}\theta^i L^i e_t, L^{s_2}(\sum_{i=0}^{\infty}\theta^i L^i)e_t\right).\end{aligned}$$

But $s_1 = s_2 = 1$ then,

$$\begin{aligned}
&\mathbb{C}ov\left(L^{s_1}\left((\frac{\phi}{\phi-\theta})\sum_{i=0}^{\infty}\phi^i L^i + (\frac{\theta}{\theta-\phi})\sum_{i=0}^{\infty}\theta^i L^i\right)e_t, L^{s_2}(\sum_{i=0}^{\infty}\theta^i L^i)e_t\right)\\
&= \sigma^2(\frac{\phi}{\phi-\theta}\sum_{i=0}^{\infty}(\theta\phi)^i) + \sigma^2(\frac{\theta}{\theta-\phi}\sum_{i=0}^{\infty}\theta^{2i})\\
&= \sigma^2\left(\frac{\phi}{(\phi-\theta)(1-\theta\phi)} + \frac{\theta}{(\theta-\phi)(1-\theta^2)}\right)\\
&= \frac{\sigma^2}{(1-\theta\phi)(1-\theta^2)}.
\end{aligned} \tag{A.4}$$

Moreover, in DREGAR(1,1), $\mathbb{E}\{(\frac{L^{s_1}}{A}x'_t)(\frac{L^{s_1}}{A}x_t)\}$ is,

$$\mathbb{E}\left\{(\frac{L^{s_1}}{A}x'_t)(\frac{L^{s_1}}{A}x_t)\right\} = \mathbb{E}\left\{(L^{s_1}\sum_{i=0}^{\infty}\phi^i L^i x'_t)(L^{s_1}\sum_{i=0}^{\infty}\phi^i L^i x_t)\right\},$$

where $s_1 = 1$. Then,

$$\begin{aligned}
\mathbb{E}\left\{(\frac{L^{s_1}}{A}x'_t)(\frac{L^{s_1}}{A}x_t)\right\} &= \mathbb{E}\left\{(\sum_{i=0}^{\infty}\phi^i L^{i+1}x'_t)(\sum_{i=0}^{\infty}\phi^i L^{i+1}x_t)\right\}\\
&= \mathbb{E}\left\{(\sum_{i=0}^{\infty}\phi^i x'_{t-i-1})(\sum_{i=0}^{\infty}\phi^i x_{t-i-1})\right\}\\
&= \mathbb{E}\left(\sum_{i=0}^{\infty}\phi^i\sum_{k=0}^{i}x'_{t-k-1}x_{t-(i-k)-1}\right)\\
&= \sum_{i=0}^{\infty}\phi^i\sum_{k=0}^{i}\mathbb{E}(x'_{t-k-1}x_{t-(i-k)-1})\\
&= \sum_{i=0}^{\infty}\phi^i\sum_{k=0}^{i}\gamma_{(x)_{i-2k}} \quad ,
\end{aligned}$$

where $\gamma_{(x)_k} = \gamma_{(x)_{-k}}$ is the k^{th} order auto-covariance of x'_i. By assumptions, in particular ergodicity, $\forall z \in \mathbb{N}$, $\sum_{i=0}^{z}\gamma_{(x)_i} < M < \infty$, then

$$\mathbb{E}\{\sum_{i=0}^{\infty}\phi^i x'_{t-i-1}\sum_{i=0}^{\infty}\phi^i x_{t-i-1}\} < \frac{M}{1-\phi} < \infty.$$

For the second block matrix in H_1, $\frac{1}{n}XH_{(p)}$, we have

$$\frac{1}{n}XH_{(p)} = \frac{1}{n}\begin{pmatrix}
\sum_{t=T_0+1}^{T}x_{1t}y_{t-1} & \sum_{t=T_0+1}^{T}x_{1t}y_{t-2} & \cdots & \sum_{t=T_0+1}^{T}x_{1t}y_{t-p}\\
\sum_{t=T_0+1}^{T}x_{2t}y_{t-1} & \sum_{t=T_0+1}^{T}x_{2t}y_{t-2} & \cdots & \sum_{t=T_0+1}^{T}x_{2t}y_{t-p}\\
\vdots & \vdots & \ddots & \vdots\\
\sum_{t=T_0+1}^{T}x_{rt}y_{t-1} & \sum_{t=T_0+1}^{T}x_{rt}y_{t-2} & \cdots & \sum_{t=T_0+1}^{T}x_{rt}y_{t-p}
\end{pmatrix}.$$

Considering the case $\frac{1}{n}\sum_{t=T_0+1}^{T}x_{s_3t}y_{t-s_4}$ where $s_3 \in \{1,2,3,\ldots,r\}$ and $s_4 \in \{1,2,3,\ldots,p\}$.

$$\frac{1}{n}\sum_{t=T_0+1}^{T}x_{s_3t}y_{t-s_4} = \frac{1}{n}\sum_{t=T_0+1}^{T}x_{s_3t}(\frac{L^{s_4}}{A}x'_t\beta + \frac{L^{s_4}}{AB}e_t)$$

$$= \frac{1}{n}\sum_{t=T_0+1}^{T} x_{s_3 t}\left(\frac{L^{s_4}}{A}x_t'\beta\right) + \frac{1}{n}\sum_{t=T_0+1}^{T} x_{s_3 t}\frac{L^{s_4}}{AB}e_t. \tag{A.5}$$

Since $x_{.i}, i = 1,2,\ldots,r$ is independent of the error, the final term in (A.5) tends to zero, $\frac{1}{n}x_{s_3}e = N(0, \frac{\mathbb{E}(x_{s_3}'x_{s_3})}{n}) \to o_p(1)$. For the first term,

$$\frac{1}{n}\sum_{t=T_0+1}^{T} x_{s_3 t}\left(\frac{L^{s_4}}{A}x_t'\beta\right) = g(\gamma_x)\beta < \infty,$$

where $g(.)$ is a bounded function. This is due to the fact that $\forall i = 1,2,3,\ldots,r; x_{.i}$s are mutually independent and the entire process is stable. As a result, autocorrelations of $x_{.i}, 1 \le i \le r$ die quickly after a finite number of lags, provided the coefficients are far from the boundary of stationarity. Otherwise, the decreasing speed of auto-covariances may take longer. In a special case where $x_{tr}, t = 1,2,3,\ldots$ are a random sample from $x_{.r}$, the inner correlation is zero and the last equation is zero provided $\forall i \in \{1,2,3,\ldots,r\}, \beta_i < \infty$.

For $\frac{1}{n}XH_{(q)}$,

$$\frac{1}{n}XH_{(q)} = \frac{1}{n}\begin{pmatrix} \sum_{t=T_0+1}^{T} x_{1t}\epsilon_{t-1} & \sum_{t=T_0+1}^{T} x_{1t}\epsilon_{t-2} & \cdots & \sum_{t=T_0+1}^{T} x_{1t}\epsilon_{t-q} \\ \sum_{t=T_0+1}^{T} x_{2t}\epsilon_{t-1} & \sum_{t=T_0+1}^{T} x_{2t}\epsilon_{t-2} & \cdots & \sum_{t=T_0+1}^{T} x_{2t}\epsilon_{t-q} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{t=T_0+1}^{T} x_{rt}\epsilon_{t-1} & \sum_{t=T_0+1}^{T} x_{rt}\epsilon_{t-2} & \cdots & \sum_{t=T_0+1}^{T} x_{rt}\epsilon_{t-q} \end{pmatrix}.$$

Taking $\frac{1}{n}\sum_{t=T_0+1}^{T} x_{rt}\epsilon_{t-q}$ for example. If $\mathbb{E}(x_r'x_r) < \infty$,

$$\frac{1}{n}\sum_{t=T_0+1}^{T} x_{rt}\epsilon_{t-q} = \frac{1}{n}\sum_{t=T_0+1}^{T} x_{rt}\frac{L^q}{B}e_t \propto N\left(0, \frac{\mathbb{E}(x_r'x_r)}{n}\right) \xrightarrow{p} o_p(1).$$

It is straightforward to show that $\frac{1}{n}X'X \to \sigma_x^2 I_{r\times r} = I_{r\times r}$ as data are assumed to be normalized prior to the analysis.

Finally for $\frac{1}{n}H_{(p)}'H_{(p)}$ and $\frac{1}{n}H_{(q)}'H_{(q)}$ we have

$$\frac{1}{n}H_{(p)}'H_{(p)} = \frac{1}{n}\begin{pmatrix} \sum_{i=T_0+1}^{T} y_{i-1}y_{i-1} & \sum_{i=T_0+1}^{T} y_{i-1}y_{i-2} & \cdots & \sum_{i=T_0+1}^{T} y_{i-1}y_{i-p} \\ \sum_{i=T_0+1}^{T} y_{i-2}y_{i-1} & \sum_{i=T_0+1}^{T} y_{i-2}y_{i-2} & \cdots & \sum_{i=T_0+1}^{T} y_{i-2}y_{i-p} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=T_0+1}^{T} y_{i-p}y_{i-1} & \sum_{i=T_0+1}^{T} y_{i-p}y_{i-2} & \cdots & \sum_{i=T_0+1}^{T} y_{i-p}y_{i-p} \end{pmatrix}.$$

Each element of this matrix belongs to a general form of $\frac{1}{n}\sum_{i=T_0+1}^{T} y_{i-s_7}y_{i-s_8}$ where $s_7, s_8 \in \{1,2,3,\ldots,p\}$. Thus,

$$\frac{1}{n}\sum_{i=T_0+1}^{T} y_{i-s_7}y_{i-s_8} = \frac{1}{n}\sum_{t=T_0+1}^{T}\left(\frac{L^{s_7}}{A}x_t'\beta + \frac{L^{s_7}}{AB}e_t\right)\left(\frac{L^{s_8}}{A}x_t'\beta + \frac{L^{s_8}}{AB}e_t\right)$$

$$= \frac{1}{n}\sum_{t=T_0+1}^{T}(\frac{L^{s_7}}{A}x_t'\beta)(\frac{L^{s_8}}{A}x_t'\beta) + \frac{1}{n}\sum_{t=T_0+1}^{T}(\frac{L^{s_7}}{AB}e_t\frac{L^{s_8}}{AB}e_t), \tag{A.6}$$

where the cross terms tend to zero. The second term in (A.6) is non-zero because there is an infinite number of common terms, provided the process is started from $-\infty$. The first term in (A.6) tends to $\mathbb{E}\left((\frac{L^{s_7}}{A}x_t')(\frac{L^{s_8}}{A}x_t')'\right)$, provided the expectation exists. If $s_7 = s_8$, the formula above produces the diagonal elements with the same values (remembering that all exogenous variables are independent and have unit variance). As it is pointed out, covariance stationary of $\{x_t\}$ is sufficient for the last expectation to be bounded. In the special case of DREGAR(1,1), it results in

$$\begin{aligned}
\mathbb{E}\left((\frac{L^{s_7}}{A}x_t')(\frac{L^{s_8}}{A}x_t')'\right) &= \mathbb{E}\left((\sum_{i=0}^{\infty}\phi^i L^i x_{t-1}')(\sum_{i=0}^{\infty}\phi^i L^i x_{t-1}')'\right)\\
&= \mathbb{E}(\sum_{i=0}^{\infty}\phi^i x_{t-i-1}'\sum_{i=0}^{\infty}\phi^i x_{t-i-1}')\\
&= \mathbb{E}(\sum_{i=0}^{\infty}\phi^i\sum_{k=0}^{i}x_{t-k-1}'x_{t-(i-k)-1})\\
&= \sum_{i=0}^{\infty}\phi^i\sum_{k=0}^{i}\gamma_{(x)_{i-2k}} < \infty.
\end{aligned}$$

For $\mathbb{E}(\frac{L^{s_7}}{AB}e_t\frac{L^{s_8}}{AB}e_t)$,

$$\begin{aligned}
\mathbb{E}(\frac{L^{s_7}}{AB}e_t\frac{L^{s_8}}{AB}e_t) &= \mathbb{E}\left\{(\sum_{i=0}^{\infty}\phi^i L^i\sum_{j=0}^{\infty}\theta^i L^i e_{t-1})(\sum_{i=0}^{\infty}\phi^i L^i\sum_{j=0}^{\infty}\theta^i L^i e_{t-1})\right\}\\
&= \mathbb{E}\left\{(\frac{\phi}{\phi-\theta}\sum_{i=0}^{\infty}\phi^i L^i + \frac{\theta}{\theta-\phi}\sum_{i=0}^{\infty}\theta^i L^i)e_{t-1}(\frac{\phi}{\phi-\theta}\sum_{i=0}^{\infty}\phi^i L^i + \frac{\theta}{\theta-\phi}\sum_{i=0}^{\infty}\theta^i L^i)e_{t-1}\right\}\\
&= \sigma^2\left((\frac{\phi}{\phi-\theta})^2\frac{1}{1-\phi^2} + (\frac{\theta}{\theta-\phi})^2\frac{1}{1-\theta^2} + 2(\frac{\theta}{\theta-\phi})(\frac{\phi}{\phi-\theta})\frac{1}{1-\theta\phi}\right)\\
&= \frac{\sigma^2}{(\phi-\theta)^2}\left(\frac{\phi^2}{1-\phi^2} + \frac{\theta^2}{1-\theta^2} - 2\frac{\phi\theta}{1-\phi\theta}\right) < \infty.
\end{aligned}$$

Finally, a similar calculation for $\frac{1}{n}H_{(q)}'H_{(q)}$ results in

$$\frac{1}{n}H_{(q)}'H_{(q)} = \frac{1}{n}\begin{pmatrix}
\sum_{i=T_0+1}^{T}\epsilon_{i-1}\epsilon_{i-1} & \sum_{i=T_0+1}^{T}\epsilon_{i-1}\epsilon_{i-2} & \cdots & \sum_{i=T_0+1}^{T}\epsilon_{i-1}\epsilon_{i-q}\\
\sum_{i=T_0+1}^{T}\epsilon_{i-2}\epsilon_{i-1} & \sum_{i=T_0+1}^{T}\epsilon_{i-2}\epsilon_{i-2} & \cdots & \sum_{i=T_0+1}^{T}\epsilon_{i-2}\epsilon_{i-q}\\
\vdots & \vdots & \ddots & \vdots\\
\sum_{i=T_0+1}^{T}\epsilon_{i-q}\epsilon_{i-1} & \sum_{i=T_0+1}^{T}\epsilon_{i-q}\epsilon_{i-2} & \cdots & \sum_{i=T_0+1}^{T}\epsilon_{i-q}\epsilon_{i-q}
\end{pmatrix}.$$

Every element of this matrix is of the form $\frac{1}{n}\sum_{i=T_0+1}^{T}\epsilon_{i-s_9}\epsilon_{i-s_{10}}$ where $s_9, s_{10} \in \{1,2,3,\ldots,q\}$. Following the previous proofs, one can show that $\frac{1}{T}\sum_{t=T_0+1}^{T}\frac{L^{s_9}}{B}e_t\frac{L^{s_{10}}}{B}e_t$ is non-zero.

All in all, H_1 tends to Q that is,

$$Q = \begin{pmatrix}
I_{r\times r} & \mathbb{E}(x_{s_3 t}(\frac{L^{s_4}}{A}x_t'\beta)|s_3,s_4,t) & O_{r\times q}\\
 & \mathbb{E}\left((\frac{L^{s_7}}{A}x_t'\beta)(\frac{L^{s_8}}{A}x_t'\beta) + (\frac{L^{s_7}}{AB}e_t\frac{L^{s_8}}{AB}e_t)|s_7,s_8,t\right) & \mathbb{E}\left((\frac{L^{s_1}}{A}\frac{1}{B}e_t)(\frac{L^{s_2}}{B}e_t)|s_1,s_2,t\right)\\
 & & \mathbb{E}\left(\frac{L^{s_9}}{B}e_t\frac{L^{s_{10}}}{B}e_t|s_9,s_{10},t\right)
\end{pmatrix}, \tag{A.7}$$

with $(s_1, s_4, s_7, s_8) \in \{1, 2, \ldots, p\}$, $(s_2, s_9, s_{10}) \in \{1, 2, \ldots, q\}$ and $s_3 \in \{1, 2, \ldots, r\}$. We should stress that Q consists of only x'_{ti} and e_t, $t = T_0 + 1, T_0 + 2, T_0 + 3, \ldots, T$ and $i = 1, 2, 3, \ldots, r$. Furthermore, in all discussed cases p and q can freely increase to infinity, provided H_1 is non-singular and there are enough observations.

In the following lines, we focus on $H_2'e$ and introduce the central limit theorem for martingales and find the limit distribution of H_2e.

Theorem A.1 (Central limit theorem for martingales). If $\{y_t\}$ is a martingale difference sequence with mean and variance given by

$$\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t, \quad \bar{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n} \sigma_t^2,$$

and provided that higher order moments are bounded,

$$\mathbb{E}(|y_t|^{2+\delta}) < \infty, \quad \delta > 0,$$

and

$$\frac{1}{n}\sum_{t=1}^{n} y_t^2 - \bar{\sigma}_n^2 \xrightarrow{p} 0.$$

Then

$$\sqrt{n}(\frac{\bar{y}}{\bar{\sigma}}) \xrightarrow{d} N(0, 1).$$

Proof. [Martin et al., 2012, p.51]. □

Let H_t° be $\frac{1}{\sqrt{n}}H_2'e$,

$$H_t^\circ = \frac{1}{\sqrt{n}}(X', H_{(p)}, H_{(q)})'e.$$

H_t° is a martingale difference sequence (MDS) because

$$\begin{aligned}\mathbb{E}\left(H_t^\circ \middle| t = t-1, t-2, \ldots, t-(p+q)\right) &= \mathbb{E}\left((X_t', H_{(p)_t}, H_{(q)_t})'e_t \middle| \mathcal{F}_{t-1}\right), \\ &= (X_t', H_{(p)_t}, H_{(q)_t})'\mathbb{E}(e_t) = 0,\end{aligned}$$

where $\mathcal{F}_{t-1}$ contains the information up to time $t-1$. We show that the central limit theorem for MDS holds for H°,

$$\bar{\mu} = \frac{1}{n}\sum_{t=T_0+1}^{T} H_t^\circ$$

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{t=T_0+1}^{T} \mathbb{V}ar(H_t^\circ) = \sigma^4 Q.$$

To establish the boundedness condition in the martingales central limit theorem, a convenient option is choosing $\delta = 2$ so that

$$\mathbb{E}(|H_t^{\circ}|^4) = \mathbb{E}(e_t^4) E\left(X_t', H_{(p)_t}, H_{(q)_t}\right)^4.$$

Using assumption [a], $\mathbb{E}(e_t^4) < \infty$ and it can be shown that $E\left(X_t', H_{(p)_t}, H_{(q)_t}\right)^4 < \infty$, provided y_t and $x_{.i}', i = 1, 2, \ldots, r$ are stationary and ergodic. Moreover,

$$\begin{aligned}\frac{1}{n}\sum_{t=T_0+1}^{T} e_t^2\left(X_t, H_{(p)_t}', H_{(q)_t}'\right)^2 = \\ \frac{1}{n}\sum_{t=T_0+1}^{T}(e_t^2-\sigma^2)\left(X_t', H_{(p)_t}, H_{(q)_t}\right)^2 + \sigma^2\frac{1}{n}\sum_{t=T_0+1}^{T}\left(X_t', H_{(p)_t}, H_{(q)_t}\right)^2. \end{aligned} \tag{A.8}$$

The first term in (A.8) is a mean zero MDS. Using the weak law of large numbers (WLLN), we have $\frac{1}{n}\sum_{t=T_0+1}^{T}(e_t^2-\sigma^2)\left(X_t', H_{(p)_t}, H_{(q)_t}\right)^2 \xrightarrow{d} 0$. The second term in RHS of (A.8) tends to $\sigma^2 Q$ where Q is defined in (A.7). As a result

$$\frac{1}{n}\sum_{t=T_0+1}^{T} e_t^2\left(X_t, H_{(p)_t}', H_{(q)_t}'\right)^2 \xrightarrow{d} \sigma^2 Q.$$

Therefore, the central limit theorem for martingales results in $\frac{1}{\sqrt{n}} H_2' e \xrightarrow{d} N(0, \sigma^2 Q)$.

A.0.2 Source of the bias

In the previous section, we relied on the assumption that ϵ_t, and as a result $H_{(q)}$, are known, whereas this is not the case in reality. In fact both ϵ and θ are unknown in real applications. Consequently, ϵ must be estimated from a primary step precisely from $(Y - H_{(p)}\phi - X\beta)$. Then we concentrate on the theoretical properties of estimating ϕ in the presence of autocorrelated residuals.

Let the initial model be

$$y = H_{(p)}\phi + X'\beta + \epsilon,$$

where we assume an AR(q) process for ϵ. Estimating parameters using OLS leads to

$$\begin{aligned}\hat{\phi} &= \begin{pmatrix} H_{(p)}'H_{(p)} & H_{(p)}'X \\ X'H_{(p)} & X'X \end{pmatrix}_p^{-1} H_{(p)}'y \\ &= \phi + \begin{pmatrix} H_{(p)}'H_{(p)} & H_{(p)}'X \\ X'H_{(p)} & X'X \end{pmatrix}_p^{-1} H_{(p)}'\epsilon, \end{aligned} \tag{A.9}$$

where $(M)_p$ represents the first p rows of the corresponding matrix M. The second term in RHS of (A.9) is the source of the bias, which we show by using asymptotic results. To this end, the asymptotic form

of the estimations is defined by:

$$\sqrt{n}(\hat{\phi}-\phi) = n\begin{pmatrix} H'_{(p)}H_{(p)} & H'_{(p)}X \\ X'H_{(p)} & X'X \end{pmatrix}^{-1}_{p} \frac{1}{\sqrt{n}} H'_{(p)}\epsilon \overset{n\to\infty}{=} (\Sigma_y)^{-1}\frac{1}{\sqrt{n}}H'_{(p)}\epsilon$$
$$\propto \sqrt{n}H'_{(p)}\epsilon,$$

where Σ_y is the covariance matrix of the corresponding element in the inverse term. On the other hand, $H'_{(p)}$ and ϵ are not independent because of the inner-correlations in ϵs. For instance the first column of $H'_{(p)}$ and ϵ_t are correlated via $\theta\epsilon_{k-1}$ and all former lags. As a result $\sqrt{n}H'_{(p)}\epsilon$ is not a proper martingale and this results in a bias and complicated structure for the distribution of estimations.

Bibliography

Abbruzzo, A., Vujačić, I., Wit, E., and Mineo, A. M. Generalized information criterion for model selection in penalized graphical models. *arXiv preprint arXiv:1403.1249*, 2014. (Cited on page 46.)

Abramowitz, M. and Stegun, I. *Handbook of mathematical functions: With formulas, graphs, and mathematical tables*. Dover Books on Mathematics. Dover Publications, 2012. ISBN 9780486158242. (Cited on page 50.)

Ahsanullah, M., Kibria, B., and Shakil, M. *Normal and Student´s t distributions and their applications*. Atlantis Studies in Probability and Statistics. Atlantis Press, 2014. ISBN 9789462390614. (Cited on page 49.)

An, H., Huang, D., Yao, Q., and Zhang, C.-H. Stepwise searching for feature variables in High-Dimensional linear regression. 2008. (Cited on page 9.)

Ando, T. Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, 94(2):443–458, 2007. (Cited on page 82.)

Angers, J.-F. and Biswas, A. A Bayesian analysis of Zero-inflated generalized Poisson model. *Computational statistics & data analysis*, 42(1):37–46, 2003. (Cited on page 69.)

Araújo Santos, P. and Fraga Alves, M. I. Improved shape parameter estimation in a discrete Weibull model. In *Recent Developments in Modeling and Applications in Statistics . Studies in Theoretical and Applied Statistics.*, pages 71–80. Springer-Verlag, 2013. (Cited on page 71.)

Aravkin, A. Y., Burke, J. V., and Pillonetto, G. Linear system identification using stable spline kernels and PLQ penalties. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 5168–5173. IEEE, 2013. (Cited on page 17.)

Arlot, S., Celisse, A., et al. A survey of Cross Validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010. (Cited on page 35.)

Atienza, N., Garcia Heras, J., Munoz Pichardo, J. M., and Villa, R. An application of mixture distributions in modelization of length of hospital stay. *Statistics in Medicine*, 27(9):1403–1420, 2008. (Cited on page 69.)

Augugliaro, L., Mineo, A. M., and Wit, E. C. Differential geometric least angle regression: A differential geometric approach to sparse generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):471–498, 2013. (Cited on page 9.)

Bańbura, M., Giannone, D., and Reichlin, L. Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92, 2010. (Cited on page 17.)

Bao, Y., Vinciotti, V., Wit, E., and 't Hoen, P. Joint modeling of ChIP-seq data via a Markov random field model. *Biostatistics*, 15(2):296–310, 2014. (Cited on page 69.)

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009. (Cited on page 9.)

Bedard, M. Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234. *Stochastic Processes and their Applications*, 118(12):2198–2222, 2008. (Cited on pages 74 and 79.)

Berk, R. and MacDonald, J. M. Overdispersion and Poisson regression. *Journal of Quantitative Criminology*, 24(3):269–284, 2008. (Cited on page 11.)

Bloomfield, P. and Steiger, W. *Least Absolute Deviations: Theory, applications and algorithms*. Progress in Probability. Birkhäuser Boston, 2012. ISBN 9781468485745. (Cited on page 2.)

Borjesson, P. S. et al. Simple approximations of the error function Q (x) for communications applications. *Communications*, 1979. (Cited on page 59.)

Box, C. D. R., George EP. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26(2):211–252, 1964. (Cited on page 15.)

Box, G. E. and Pierce, D. A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332):1509–1526, 1970. (Cited on page 40.)

Bozdogan, H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987. (Cited on pages 40 and 82.)

Bracquemond, C. and Gaudoin, O. A survey on discrete lifetime distributions. *International Journal of Reliability, Quality and Safety Engineering*, 10(1):69–98, 2003. (Cited on pages 69 and 70.)

Cameron, A. C. and Trivedi, P. K. *Regression analysis of count data*. Cambridge university press, 2013. (Cited on pages 10 and 69.)

Candes, E. and Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007. (Cited on page 4.)

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. *Measurement error in nonlinear models: A modern perspective, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2006. ISBN 9781420010138. (Cited on page 2.)

Carter, E. and Potts, H. Predicting length of stay from an electronic patient record system: A primary total knee replacement example. *BMC Medical Informatics and Decision Making*, 14(26), 2014. (Cited on page 69.)

Chang, S.-H., Cosman, P. C., and Milstein, L. B. Chernoff-type bounds for the Gaussian error function. *Communications, IEEE Transactions on*, 59(11):2939–2944, 2011. (Cited on page 50.)

Chen, S. S., Donoho, D. L., and Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998. (Cited on page 8.)

Chevillard, S. and Revol, N. Computation of the error function erf in arbitrary precision with correct rounding. *JD Bruguera et M. Daumas (editeurs): RNC*, 8:27–36, 2008. (Cited on page 59.)

Chiuso, A. and Pillonetto, G. A Bayesian approach to sparse dynamic network identification. *Automatica*, 48(8):1553–1565, 2012. (Cited on page 17.)

Cody, W. J. Performance evaluation of programs for the error and complementary error functions. *ACM Trans. Math. Softw.*, 16(1):29–37, March 1990. ISSN 0098-3500. doi: 10.1145/77626.77628. (Cited on page 59.)

Cody, W. J. Rational Chebyshev approximations for the error function. *Mathematics of Computation*, 23(107):631–637, 1969. (Cited on page 59.)

Cox, D. R. Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer, 1992. (Cited on page 72.)

Daubechies, I., Defrise, M., and De Mol, C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on pure and applied mathematics*, 57(11):1413–1457, 2004. (Cited on page 9.)

Davino, C., Furno, M., and Vistocco, D. *Quantile Regression: Theory and applications.* Wiley Series in Probability and Statistics. Wiley, 2013. ISBN 9781118752715. (Cited on page 2.)

Dayton, C. M. Model comparisons using information measures. *Journal of modern applied statistical methods*, 2(2):2, 2003. (Cited on page 82.)

Dellaportas, P., Forster, J. J., and Ntzoufras, I. Bayesian variable selection using the Gibbs sampler. *BIOSTATISTICS-BASEL*, 5:273–286, 2000. (Cited on page 10.)

Donoho, D. L. and Johnstone, J. M. Ideal spatial adaptation by Wavelet shrinkage. *Biometrika*, 81(3): 425–455, 1994. (Cited on page 62.)

Dos Santos, P. *Linear parameter-varying system identification: New developments and trends.* Advanced series in electrical and computer engineering. World Scientific Publishing Company, 2012. ISBN 9789814355452. (Cited on page 16.)

Du, D. and Pardalos, P. *Minimax and applications.* Nonconvex Optimization and Its Applications. Springer US, 2013. ISBN 9781461335573. (Cited on page 2.)

Dunson, D. B., Pillai, N., and Park, J.-H. Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):163–183, 2007. (Cited on page 88.)

Efron, B. Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81(395):709–721, 1986. (Cited on page 11.)

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004. (Cited on pages 9, 33, and 64.)

El Sayyad, G. Bayesian and classical analysis of Poisson regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 445–451, 1973. (Cited on page 69.)

Englehardt, J. D. and Li, R. The discrete Weibull distribution: An alternative for correlated counts with confirmation for microbial counts in water distributions. *Risk Analysis*, 31(3):370–381, 2011. (Cited on page 69.)

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. *Regression: Models, methods and applications*. Springer Berlin Heidelberg, 2013. ISBN 9783642343339. (Cited on page 10.)

Fan, J. and Li, R. Variable selection via Nonconcave penalized likelihood and its Oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. (Cited on pages 5, 7, 30, 45, and 60.)

Fan, J., Lv, J., and Qi, L. Sparse High-Dimensional models in economics. *Annual review of economics*, 3:291, 2011. (Cited on page 14.)

Fan, L. R., Jianqing. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. (Cited on page 4.)

Frank, L. E. and Friedman, J. H. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993. (Cited on page 2.)

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007. (Cited on page 8.)

Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010. (Cited on pages 62 and 84.)

Friedman, J. H. Fast sparse regression and classification. *International Journal of Forecasting*, 28(3): 722–738, 2012. (Cited on page 62.)

Fu, W. J. Penalized regressions: The Bridge versus the Lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998. (Cited on page 9.)

Fuller, W. *Measurement error models*. Wiley Series in Probability and Statistics. Wiley, 2009. ISBN 9780470317334. (Cited on page 2.)

Galton, F. *Natural inheritance*. Macmillan, 1894. (Cited on page 1.)

George, E. I. and McCulloch, R. E. Variable selection via Gibbs sampling. *Journal of the American Atatistical Association*, 88(423):881–889, 1993. (Cited on page 10.)

Ghosh, S. K., Mukhopadhyay, P., and Lu, J.-C. J. Bayesian analysis of Zero-inflated regression models. *Journal of Statistical planning and Inference*, 136(4):1360–1375, 2006. (Cited on page 70.)

Gibbons, J. and Chakraborti, S. *Nonparametric statistical inference*. Statistics, textbooks and monographs. Marcel Dekker Incorporated, 2003. ISBN 9780824755225. (Cited on page 2.)

Gimenez, P., Bolfarine, H., and Colosimo, E. *Estimation in Weibull Regression Model with Measurement Error*. RT-MAE. IME-USP, 1997. (Cited on page 72.)

Green, P. J. Reversible Jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995. (Cited on pages 10, 70, and 78.)

Grunwald, G. K., Bruce, S. L., Jiang, L., Strand, M., and Rabinovitch, N. A statistical model for under-or overdispersed clustered and longitudinal count data. *Biometrical Journal*, 53(4):578–594, 2011. (Cited on page 83.)

Gustafson, P. *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 2003. ISBN 9781135441234. (Cited on page 2.)

Haario, H., Saksman, E., and Tamminen, J. An adaptive Metropolis algorithm. *Bernoulli*, pages 223–242, 2001. (Cited on page 74.)

Hamilton, J. *Time series analysis*. Princeton University Press, 1994. ISBN 9780691042893. (Cited on pages 15 and 20.)

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: Data mining, inference, and prediction*. Springer Series in Statistics. Springer New York, 2013. ISBN 9780387216065. (Cited on page 9.)

Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. (Cited on pages 10, 73, and 74.)

Hebiri, M. and Lederer, J. How correlations influence Lasso prediction. *Information Theory, IEEE Transactions on*, 59(3):1846–1854, March 2013. doi: 10.1109/tit.2012.2227680. (Cited on page 33.)

Hebiri, M. Regularization with the Smooth-Lasso procedure. *arXiv preprint arXiv:0803.0668**, 2008. (Cited on pages 3 and 45.)

Hebiri, M., van de Geer, S., et al. The Smooth-Lasso and other $l_1+ l_2$ penalized methods. *Electronic Journal of Statistics*, 5:1184–1226, 2011. (Cited on page 45.)

Hesterberg, T., Choi, N. H., Meier, L., Fraley, C., et al. Least angle and l_1 penalized regression: A review. *Statistics Surveys*, 2:61–93, 2008. (Cited on page 9.)

Hibbs Jr, D. A. Problems of statistical estimation and causal inference in time-series regression models. *Sociological methodology*, 1974:252–308, 1973. (Cited on page 38.)

Hirose, K., Tateishi, S., and Konishi, S. Efficient algorithm to select tuning parameters in sparse regression modelling with regularization. *arXiv preprint arXiv:1109.2411**, 2011. (Cited on page 35.)

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. (Cited on pages 2, 10, and 45.)

Hollander, M., Wolfe, D., and Chicken, E. *Nonparametric statistical methods*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN 9781118553299. (Cited on page 2.)

Hosmer, D., Lemeshow, S., and May, S. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley Series in Probability and Statistics. Wiley, 2011. ISBN 9781118211588. (Cited on page 72.)

Hougaard, P., Lee, M. T., and Whitmore, G. A. Analysis of overdispersed count data by mixtures of Poisson variables and Poisson processes. *Biometrics*, 53:1225–1238, 1997. (Cited on page 69.)

Hu, M.-C., Pavlicova, M., and Nunes, E. V. Zero-inflated and Hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial. *The American journal of drug and alcohol abuse*, 37(5):367–375, 2011. (Cited on page 11.)

Huang, M. S. Z. C.-H., Jian. Adaptive Lasso for sparse High-Dimensional regression models. *Statistica Sinica*, 18(4):1603, 2008. (Cited on pages 5, 6, and 16.)

Hwang, C.-R. Simulated annealing: theory and applications. *Acta Applicandae Mathematicae*, 12(1): 108–111, 1988. (Cited on page 88.)

Ishwaran, H. and Rao, J. S. Spike and Slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics*, pages 730–773, 2005. (Cited on pages 10 and 70.)

James, G. M., Radchenko, P., and Lv, J. DASSO: Connections between the Dantzig selector and Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):127–142, 2009. (Cited on page 9.)

Kalktawi, H. S., Vinciotti, V., and Yu, K. A Simple and Adaptive Dispersion Regression Model for Count Data. *arXiv:1511.00634**, November 2016. (Cited on pages 11, 69, 81, 83, and 84.)

Kass, R. E. Bayes factors in practice. *The Statistician*, pages 551–560, 1993. (Cited on page 82.)

Keesman, K. *System identification: An introduction*. Advanced Textbooks in Control and Signal Srocessing. Springer London, 2011. ISBN 9780857295224. (Cited on page 16.)

Khan, M. S. A., Khalique, A., and Abouammoth, A. M. On estimating parameters in a discrete Weibull distribution. *IEEE transactions on Reliability*, 38(3):348–350, 1989. (Cited on pages 11 and 71.)

Kim, J. and Pollard, D. Cube Root Asymptotics. *The Annals of Statistics*, 18(1):191–219, 03 1990. doi: 10.1214/aos/1176347498. (Cited on page 28.)

Knight, K. and Fu, W. Asymptotics for Lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000. (Cited on pages 5, 7, 16, 28, 29, 31, 53, and 56.)

Koenker, R. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005. ISBN 9780521608275. (Cited on page 2.)

Konishi, S. and Kitagawa, G. Generalised information criteria in model selection. *Biometrika*, 83(4): 875–890, 1996. (Cited on pages 46 and 61.)

Kulasekera, K. B. Approximate MLE's of the parameters of a discrete Weibull distribution with Type I censored data. *Microelectronics Reliability*, 34(7):1185–1188, 1994. (Cited on page 71.)

Kuo, L. and Mallick, B. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81, 1998. (Cited on page 10.)

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of econometrics*, 54(1):159–178, 1992. (Cited on page 42.)

Kyung, M., Gill, J., and Ghosh, C. G., Malay. Penalized regression, standard errors, and Bayesian Lasso. *Bayesian Analysis*, 5(2):369–411, 2010a. (Cited on pages 5 and 10.)

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. Penalized regression, standard errors, and Bayesian Lassos. *Bayesian Analysis*, 5(2):369–411, 2010b. (Cited on page 70.)

Lai, C. D. Issues concerning constructions of discrete lifetime models. *Qualitative technology of quantitative managment*, 10(2):251–262, 2013. (Cited on page 69.)

Lam, K. F., Xue, H., and Bun Cheung, Y. Semiparametric analysis of Zero-Inflated count data. *Biometrics*, 62(4):996–1003, 2006. (Cited on page 69.)

Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992. (Cited on page 11.)

Lawrence, K. D. and Arthur, J. L. *Robust regression: Analysis and applications*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 1990. (Cited on page 2.)

Lee, C.-I. C. On Laplace continued fraction for the normal integral. *Annals of the Institute of Statistical Mathematics*, 44(1):107–120, 1992. (Cited on page 59.)

Leng, L. Y. W. G., Chenlei. A note on the Lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273, 2006. (Cited on page 5.)

Lesaffre, E., Rizopoulos, D., and Tsonaka, R. The logistic transform for bounded outcome scores. *Biostatistics*, 8(1):72–85, 2007. (Cited on page 72.)

Liu, H. and Powers, D. A. Bayesian inference for Zero-inflated Poisson regression models. *Journal of Statistics: Advances in Theory and Applications*, 7(2):155–188, 2012. (Cited on page 70.)

Ljung, L. *System Identification: Theory for the user*. Pearson Education, 1998. ISBN 9780132440530. (Cited on page 16.)

Lozano, M. N., Aurélie C. Minimum distance estimation for robust High-Dimensional regression. *arXiv preprint arXiv:1307.3227**, 2013. (Cited on page 36.)

Machado, J. and Santos Silva, M. Quantiles for counts. *JASA*, 100(472):1226–1237, 2005. (Cited on pages xi, 10, 69, 83, 84, 85, and 86.)

Martin, A. D., Quinn, K. M., Park, J. H., et al. Mcmcpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, 42(9):1–21, 2011. (Cited on page 83.)

Martin, V., Hurn, S., and Harris, D. *Econometric modelling with time series: Specification, estimation and testing*. Themes in Modern Econometrics. Cambridge University Press, 2012. ISBN 9780521139816. (Cited on page 97.)

Medeiros, M. E. F., Marcelo C. Estimating High-Dimensional time series models. Technical report, 2012. (Cited on pages 13 and 35.)

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006. (Cited on pages 5 and 6.)

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Euation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953. (Cited on page 10.)

Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. On the Lasso and Its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 1999. (Cited on page 8.)

Mikel Esnaola, Pedro Puig, David Gonzalez, Robert Castelo, and Juan R Gonzalez. A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments. *BMC Bioinformatics*, 14:254, 2013. (Cited on page 69.)

Miller, A. *Subset selection in regression*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 2002. ISBN 9781420035933. (Cited on page 5.)

Mohebbi, M., Wolfe, R., and Forbes, A. Disease mapping and regression with count data in the presence of overdispersion and spatial autocorrelation: A Bayesian model averaging approach. *International journal of environmental research and public health*, 11(1):883–902, 2014. (Cited on page 69.)

Morten Arendt Rasmussen, R. B. A tutorial on the Lasso approach to sparse modeling. *Chemometrics and Intelligent Laboratory Systems*, 119(0):21 – 31, 2012. ISSN 0169-7439. (Cited on page 5.)

Nagakawa, T. and Osaki, S. The discrete Weibull distribution. *IEEE transactions on reliability*, R-24 (5), 1975. (Cited on pages 10, 69, and 70.)

Nawata, K. *The least absolute deviations estimators in generalized regression models*, volume 88 of *Economics and econometrics*. University of Tokyo, Komaba, Dept. of Social and International Relations, 1988. (Cited on page 2.)

Neelon, B. H., OMalley, A. J., and Normand, S.-L. T. A Bayesian model for repeated measures Zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling*, 10(4): 421–439, 2010. (Cited on page 70.)

Nelder, J. A. and Wedderburn, R. W. Generalized linear models. *Journal of the Royal Statistical Society. Series A*, pages 370–384, 1972. (Cited on page 10.)

Nelles, O. *Nonlinear system identification: From classical approaches to Neural networks and Fuzzy models*. Springer Berlin Heidelberg, 2013. ISBN 9783662043233. (Cited on page 16.)

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005. (Cited on page 46.)

Newcombe, P., Ali, H. R., Blows, F., Provenzano, E., Pharoah, P., Caldas, C., and Richardson, S. Weibull regression with Bayesian variable selection to identify prognostic tumour markers of breast cancer survival. *Statistical methods in medical research*, page 0962280214548748, 2014. (Cited on page 73.)

Olver, F., National Institute of Standards, and Technology (U.S.). *NIST Handbook of mathematical functions*. Cambridge University Press, 2010. ISBN 9780521192255. (Cited on pages 46 and 58.)

Osgood, D. W. Poisson-based regression analysis of aggregate crime rates. *Journal of quantitative criminology*, 16(1):21–43, 2000. (Cited on page 10.)

Ozsolak, F. and Milos, P. M. RNA sequencing: Advances, challenges and opportunities. *Nature Review Genetics*, 12:87–98, 2011. (Cited on page 69.)

Pan, W. Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125, 2001. (Cited on pages 40 and 82.)

Park, T. and Casella, G. Bayesian Lasso. *Journal of the American Statistical Association*, 103(482): 681–686, 2008. (Cited on pages 5, 10, 70, and 73.)

Parker, K. F. Industrial shift, polarized labor markets and urban violence: Modeling the dynamics between the economic transformation and disaggregated homicide. *Criminology*, 42(3):619–646, 2004. (Cited on page 10.)

Paternoster, R. and Brame, R. Multiple routes to delinquency. A test of developmental and general theories of crime. *Criminology*, 35(1):49–84, 1997. (Cited on page 10.)

Perkins, S., Lacker, K., and Theiler, J. Grafting: Fast, incremental feature selection by gradient descent in function space. *The Journal of Machine Learning Research*, 3:1333–1356, 2003. (Cited on page 8.)

Pillonetto, G. and Aravkin, A. A new kernel-based approach for identification of time-varying linear systems. In *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*, pages 1–6. IEEE, 2014. (Cited on page 17.)

Pillonetto, G. and Chiuso, A. Tuning complexity in regularized kernel-based regression and linear system identification. *Automatica (Journal of IFAC)*, 58(C):106–117, 2015. (Cited on page 17.)

Pillonetto, G., Chen, T., Chiuso, A., De Nicolao, G., and Ljung, L. Regularized linear system identification using atomic, nuclear and kernel-based norms: The role of the stability constraint. *arXiv preprint arXiv:1507.00564**, 2015. (Cited on page 17.)

Pintelon, R. and Schoukens, J. *System identification: A frequency domain approach*. Wiley, 2004. ISBN 9780471660958. (Cited on page 16.)

Polpo, A., Coque Jr, M., and Pereira, C. Statistical analysis for Weibull distributions in presence of right and left censoring. In *Reliability, Maintainability and Safety, 2009. ICRMS 2009. 8th International Conference on*, pages 219–223. IEEE, 2009. (Cited on page 73.)

Pourahmadi, M. *High-Dimensional covariance estimation: With High-Dimensional data*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN 9781118573662. (Cited on page 5.)

Press, W. *Numerical Recipes in C: The art of scientific computing*. Number v. 4. Cambridge University Press, 1992. ISBN 9780521437202. (Cited on page 59.)

Qian, Y. Y., Wei. Model selection via standard error adjusted adaptive Lasso. *Annals of the Institute of Statistical Mathematics*, 65(2):295–318, 2013. (Cited on page 6.)

Ramirez, C., Sanchez, R., et al. $\sqrt{x^2+m}$ is the most computationally efficient smooth approximation to $|x|$. *Journal of Uncertain Systems*, 8, 2014. (Cited on page 46.)

Rish, I. and Grabarnik, G. *Sparse modeling: Theory, algorithms, and applications*. CRC Press, 2014. (Cited on page 9.)

Robinson, M. D. and Smyth, G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332, 2008. (Cited on page 69.)

Rosset, S. and Zhu, J. Piecewise linear regularized solution paths. *The Annals of Statistics*, pages 1012–1030, 2007. (Cited on page 8.)

Sáez-Castillo, A. and Conde-Sánchez, A. A hyper-Poisson regression model for overdispersed and underdispersed count data. *Computational Statistics & Data Analysis*, 61:148–157, 2013. (Cited on page 11.)

Sampson, R. J. and Laub, J. H. Socioeconomic achievement in the life course of disadvantaged men: Military service as a turning point, circa 1940-1965. *American Sociological Review*, pages 347–367, 1996. (Cited on page 10.)

Schmidt, M., Fung, G., and Rosales, R. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *Machine Learning: ECML 2007*, pages 286–297. Springer, 2007a. (Cited on page 45.)

Schmidt, M., Fung, G., and Rosales, R. Fast optimization methods for l_1 regularization: A comparative study and two new approaches. In Kok, J., Koronacki, J., Mantaras, R., Matwin, S., Mladenič, D., and Skowron, A., editors, *Machine Learning: ECML 2007*, volume 4701 of *Lecture Notes in Computer Science*, pages 286–297. Springer Berlin Heidelberg, 2007b. ISBN 978-3-540-74957-8. doi: 10.1007/978-3-540-74958-5_28. (Cited on page 46.)

Sellers, K. F. and Shmueli, G. A flexible regression model for count data. *The Annals of Applied Statistics*, 4(2):943–961, 2010. (Cited on pages 11 and 69.)

Shao, J. Linear model selection by Cross Validation. *Journal of the American statistical Association*, 88 (422):486–494, 1993. (Cited on page 35.)

Soliman, A. A., Abd Ellah, A. H., Abou Elheggag, N. A., and Ahmed, E. A. Modified Weibull model: A Bayes study using MCMC approach based on progressive censoring data. *Reliability Engineering & System Safety*, 100:48–57, 2012. (Cited on page 73.)

Song, S. and Bickel, P. J. Large vector auto regressions. *ArXiv e-prints*, 2011. (Cited on page 13.)

Sourav, C. Assumptionless consistency of the Lasso. *arXiv*, 5817v3*, 2013. (Cited on pages 5 and 6.)

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 583–639, 2002. (Cited on page 82.)

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., and Yang, N. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *The Journal of urology*, 141(5):1076–1083, 1989. (Cited on page 65.)

Suo, X. and Tibshirani, R. An ordered Lasso and sparse time-lagged regression. *Technometrics*, (just-accepted), 2015. (Cited on page 13.)

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. (Cited on pages 3, 6, 16, 35, 36, and 73.)

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005. (Cited on page 4.)

Turlach, B. A. On algorithms for solving least squares problems under an L_1 Penalty or an L_1 Constraint. *Proc. American Statistical Association; Statistical Computing Section*, 2005. (Cited on page 9.)

Ulgen, B. J. J.-J. B. K. R., Onur M. Simulation methodology: A practitioner's perspective. *International Journal of Industrial Engineering, Applications and Practice*, 1(2), 1994. (Cited on page 36.)

Usai, M. G., Goddard, M. E., and Hayes, B. J. Lasso with Cross Validation for genomic selection. *Genetics research*, 91(06):427–436, 2009. (Cited on page 35.)

Vazquez Leal, H., Castaneda Sheissa, R., Filobello Nino, U., Sarmiento Reyes, A., and Sanchez Orea, J. High accurate simple approximation of normal distribution integral. *Mathematical problems in engineering*, 2012, 2012. (Cited on page 58.)

Wang, H., Li, G., and Tsai, C.-L. Regression coefficient and autoregressive order shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):63–78, 2007. (Cited on pages 13, 14, 15, 16, 25, 29, 30, 31, and 44.)

Wang, L. C., Hansheng. Unified Lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479), 2007. (Cited on page 35.)

Wasserman, L. *All of nonparametric statistics*. Springer Texts in Statistics. Springer New York, 2006. ISBN 9780387306230. (Cited on page 2.)

White, H. *Asymptotic theory for econometricians*. Economic theory, econometrics and mathematical economics. Academic Press, 2001. ISBN 9780127466521. (Cited on page 20.)

Wu, R. and Wang, Q. Shrinkage estimation for linear regression with ARMA errors. *Journal of Statistical Planning and Inference*, 142(7):2136–2148, 2012. (Cited on pages 13, 14, 15, 16, 39, 40, and 41.)

Wu, Y. *Minimax estimation of nonparametric regression through white noise problem*. Cornell University, January, 1997. (Cited on page 2.)

Y. Nardi, A. R. Autoregressive process modeling via the Lasso procedure. *Journal of Multivariate Analysis*, 102(3):528 – 549, 2011. ISSN 0047-259X. (Cited on pages 5, 13, and 36.)

Yoon, P. C. L.-T., Young Joo. Penalized regression models with autoregressive error terms. *Journal of Statistical Computation and Simulation*, (ahead-of-print):1–17, 2012. (Cited on page 5.)

Zhang, L. R. T. C.-L., Yiyun. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323, 2010. (Cited on page 35.)

Zhao, Y. B., Peng. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006. (Cited on pages 5 and 6.)

Zhou, M., Li, L., Dunson, D., and Carin, L. Lognormal and Gamma mixed negative binomial regression. In *Proceedings of the 29th International Conference on Machine Learning*, volume 2012, page 1343. NIH Public Access, 2012. (Cited on page 69.)

Zhou, v. d. G. S.-B. P., Shuheng. Adaptive Lasso for High-Dimensional regression and Gaussian graphical modeling. *arXiv preprint arXiv:0903.2515**, 2009. (Cited on page 5.)

Zhu, Y. *Multivariable system identification for process control*. Elsevier Science, 2001. ISBN 9780080537115. (Cited on page 16.)

Zou, H. The adaptive Lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006. (Cited on pages 5, 6, 24, 34, and 63.)

Zou, H. and Hastie, T. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. (Cited on pages 3, 44, 45, and 65.)