Gamma and Inverse Gaussian dgLARS

Hassan Pazira

Promoter: ***Ernst Wit***

29.10.2014

Contents

Chapter 1

Introduction

Nowadays, high-dimensional data sets, namely data sets where the number of predictors, say p, is larger than the sample size N, are becoming more and more common. Modern statistical methods developed to study this kind of data sets are usually based on the idea to use a penalty function to estimate a solution curve embedded in the parameter space and then to find the point that represents the best compromise between sparsity and predictive behaviour of the model. Recent statistical literature has a great number of contributions devoted to this problem: some important examples are the L1-penalty function (Tibshirani 1996), the SCAD method (Fan and Li 2001), the Dantzig selector (Candes and Tao 2007), which was extended to generalized linear models (GLMs) in James and Radchenko (2009), and the MC+ penalty function introduced in Zhang (2010), among others.

Differently from the methods cited above, Augugliaro, Mineo, and Wit (2013) propose a new approach based on the differential geometrical representation of a GLM. The derived method, that does not require an explicit penalty function, has been called differential geometric LARS (dgLARS) method because it is defined generalizing the geometrical ideas on which the least angle regression (LARS), proposed in Efron *et al.* (2004), is based. As underlined in Augugliaro *et al.* (2013), LARS is not only "an important contribution to statistical computing", as suggested in Madigan and Ridgeway (2004), but is a proper likelihood method in its own right: it can be generalized to any model and its succes does not depend on the arbitrary match of the constraint and the objective function, as is the case in penalized inference methods. In particular, using the differential geometric characterization of the classical signed Rao score test statistic, dgLARS gains important theoretical properties that are not shared by other methods. From a computational point of view, the dgLARS method consists essentially in the computation of the implicitly defined solution curve. In Augugliaro *et al.* (2013), this problem is satisfactorily solved by using a predictor-corrector (PC) algorithm, that however has the drawback of becoming intractable when working with thousands of predictors, since in the predictor step of this algorithem the number of arithmetic operations scale as the cube of the number of predictors. To overcome this problem, Augugliaro *et al.* (2012) propose a much more efficient cyclic coordinate descend (CCD) algorithm, which connects the original dgLARS problem with an iterative reweighted least squares (IRLS) algorithm to fit the dgLARS solution curve when we work with an high-dimensional data set.

But in Augugliaro *et al.* (2012) and (2013), the authors considered a class of the *one-parameter exponential family* for a GLM. Indeed, they assumed that the dispersion parameter is fixed, $\phi = 1$. Also, Augugliaro *et al.* (2014) presented the **dglars** package that implements both the algorithms to compute the solution curve implicitly defined by dgLARS, but just based on two discrete regression models (Logistic and Poisson).

In this paper we are going to extend the dgLARS method for a GLM to a larger class of the exponential family, namely the *exponential dispersion family* (when the dispersion parameter, ϕ, is known and unknown). We are going to present the **dglars.G.IG** package that implements both the algorithms to compute the solution curve implicitly defined by dgLARS based on Gamma and Inverse Gaussian models. The object returned by these functions is a S3 class object, for which specific methods and functions have been implemented. The package **dglars.G.IG** will be available under general public licence (GPL-2) from the Comprehensive *R* Archive Network at http://CRAN.R-project.org/package=dglars.G.IG.

The paper is organized as follows. In Section 2 we give some background of the GLM for Gamma and Inverse Gaussian. In Section 3 we introduce the dgLARS method by giving some essential clues to the theory underlying a generalized linear model from a differential geometric point of view. Section 4 is devoted to the simulation studies, and in Section 5 we draw some conclusions.

Chapter 2

Gamma and Inverse Gaussian GLM

Models are abstract, simplified representations of reality, often used both in science and in technology. No one should believe that a model could be true, although much of theoretical statistical inference is based on just this assumption. Models may be deterministic or probabilistic. Models with a probabilistic component are called statistical models. The one most important class, that with which we are concerned, contains the generalized linear models. They are so called because they generalize the classical linear models based on the normal distribution. This generalization has two aspects: in addition to the linear regression part of the classical models, these models can involve a variety of distributions selected from a special family, exponential dispersion models, and they involve transformations of the mean, through what is called a "link function", linking the regression part to the mean of one of these distributions. It had been known since the time of Fisher (1934) that many of the commonly used distributions were members of one family, which he called the *exponential family.* By the end of the 1960s, the time was ripe for a synthesis of these various models (Lindsey, 1971). In 1972, Nelder and Wedderburn went the step further in unifying the theory of statistical modelling and, in particular, regression models, publishing their article on *generalized linear models* (GLM). They showed

- how many of the most common linear regression models of classical statistics, were in fact members of one family and could be treated in the same way,
- that the maximum likelihood estimates for all of these models could be obtained using the same algorithm, *iterated weighted least squares* (IWLS).

Both elements were equally important in the subsequent history of this approach. Thus, most of the models have a distribution in the *exponential dispersion family* (Jørgensen, 1987), a generalization of the exponential family, with some transformation of the mean, the link function, being related linearly to the explanatory variables.

2.1 Generalized Linear Model

We now present the general set up for the Generalized Linear Models (GLMs). Let there be n independent random *response* variables $Y_1, \ldots, Y_n$ and p non-random explanatory variables or *covariates.* The $(p+1)$-vector $\boldsymbol{x}_i$ denotes the vector of covariates for y_i. Sometimes there may only be one observation y_i for each Y_i, but on other occasions there may be several observations y_{ij} $(j = 1, \ldots, n_i)$ for each Y_i. Our goal is to investigate the relationship between the response variables Y_i and covariates. For $i = 1, \ldots, n$, this relationship is given by

$$\begin{array}{rl} (i) & Y_i \sim f_i, \\ (ii) & \eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} = \boldsymbol{x}_i^T \boldsymbol{\beta}, \\ (iii) & \mu_i = g^{-1}(\eta_i) = E(Y_i), \end{array} \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ is a vector of $p+1$ unknown constants, β_0 belonging to the explanatory variable $(1, \ldots, 1)^T$ called *intercept.* In (2.1) we have split the model for the $Y_1, \ldots, Y_n$ into a *random component* (i) and a *systematic component* (ii). The random component of the model specifies the distribution of Y_i. The systematic component consists of a vector of so-called *predictors*, η_i, one for each observation. It specifies the way in which the explanatory variables come into the model. In (iii) the so-called *link function*, denoted by g, specifies the connection between the random and the systematic component. More precisely, g expresses η_i as a function of $E[Y_i]$, that is $g(E[Y_i]) = \eta_i$. In (2.1), f_i is the probability density function of an *exponential (dispersion) family* of distributions of the form

$$f(y_i; \theta_i, \phi) = \exp\left[\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right] \tag{2.2}$$

where θ_i is called the *canonical parameter* and represents the location while ϕ is called the *dispersion parameter* and represents the scale, and b is an arbitrary monotonic, differentiable function. The specific form of f_i is determined by the functions a, b and c. For members of the exponential (dispersion) families, a special relationship exists between the mean and the variance: the latter is a precisely defined and unique function of the former for each member (Tweedie, 1947). The relationship can be shown in the following way. For any likelihood function, $L(\theta_i, \phi; y_i) = f(y_i; \theta_i, \phi)$, for one observation, the first derivative of its logarithm,

$$\mathcal{U}_i = \frac{\partial \log\left[L(\theta_i, \phi; y_i)\right]}{\partial \theta_i} = \frac{\partial \ell(\theta_i)}{\partial \theta_i}$$

is called the *score function.* (When this function, for a complete set of observations, is set to zero, the solution of the resulting equations, called the *score equations*, yields the *maximum likelihood estimates.*) From standard inference theory, it can be shown that

$$E[\mathcal{U}_i] = 0 \qquad \text{and} \qquad Var[\mathcal{U}_i] = E[\mathcal{U}_i^2] = E\left[-\tfrac{\partial \mathcal{U}_i}{\partial \theta_i}\right]$$

under mild regularity conditions that hold for these families. It can be shown that, for $i = 1, \ldots, n$,

$$\begin{aligned} E[Y_i] &= \mu_i = b'(\theta_i), \\ Var[Y_i] &= b''(\theta_i)a_i(\phi) = V(\mu_i)a_i(\phi), \end{aligned} \tag{2.3}$$

where $V(\mu_i)$ is the *variance function*, a function of μ_i (or θ_i) only. Usually, we assume that $a_i(\phi) = \frac{\phi}{\omega_i}$ where ω_i is known *prior weight* (a known positive constant, not a parameter) that can vary between observations. In the following, we take $\omega_i = 1$ so that $a_i(\phi) = \phi$ and $Var[Y_i]$ is a product of the dispersion parameter and a function of the mean only. Here, θ_i is the parameter of interest, whereas ϕ is usually a nuisance parameter which is the same for all Y_i. Also $\phi > 0$ is assumed ($\phi < 0$ would just change the sign of some equations with only trivial effect). For these families of distributions, $b(\theta_i)$ and the variance function each uniquely distinguishes among the members.

We note that (2.1) means that a GLM is completely specified by two choices, namely the choice of the exponential family and the choice of the link function, since the systematic component is the same for all GLMs.

The choice of the link function depends on the distribution of the response. For example if y is non-negative a link function is appropriate which specifies non-negative means without restricting the parameters. For each distribution within the simple exponential family there is one link function that has some technical advantages, the so-called *canonical link*. It links the linear predictor directly to the canonical parameter in the form $\theta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} = \eta_i$. Since θ_i is determined as a function $\theta(\mu_i)$ the canonical link "g" may be derived from the general form $\mathrm{g}(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ as the transformation which transforms μ_i to θ_i. In the meantime, the canonical link might not always be the best choice.

The expression (2.2) may falsely suggest that we have n unknown parameters $\theta_1, \ldots, \theta_n$ to estimate. However, there are only $p+1$ parameters to estimate, namely the unknown constants $\beta_0, \ldots, \beta_p$. Since the $\beta_0, \ldots, \beta_p$ are linked to the $\theta_1, \ldots, \theta_n$ through the equations (2.1)(ii), (2.1)(iii), and (2.3), the density functions $f_1, \ldots, f_n$ implicitly depend on $\beta_0, \ldots, \beta_p$. This is why a natural method for the estimation of the $\beta_0, \ldots, \beta_p$ is the maximum likelihood method.

We denote the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$. It is the value obtained by maximizing the log-likelihood

$$\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell(\theta_i) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right) \tag{2.4}$$

with respect to $\boldsymbol{\beta}$. Except for the classical linear regression model, where they are the same, the least squares estimator of $\boldsymbol{\beta}$ will generally have inferior performance compared to the MLE $\hat{\boldsymbol{\beta}}$. To obtain the maximum likelihood estimator for the parameter β_j we need

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \mathcal{U}_j = \sum_{i=1}^{n} \frac{\partial \ell(\theta_i)}{\partial \beta_j} = \sum_{i=1}^{n} \left[\left(\frac{y_i - b'(\theta_i)}{a(\phi)} \right) \cdot \frac{\partial \theta_i}{\partial \beta_j} \right]$$

using the chain rule for differentiation we can write

$$\frac{\partial\theta_i}{\partial\beta_j} = \frac{\partial\theta_i(\mu_i)}{\partial\mu_i}\cdot\frac{\partial\mu_i(\eta_i)}{\partial\eta_i}\cdot\frac{\partial\eta_i}{\partial\beta_j}$$

Hence the score statistics are

$$\frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_j} = \mathcal{U}_j = \frac{1}{\phi}\sum_{i=1}^{n}\left[\frac{(y_i-\mu_i)}{V(\mu_i)}x_{ij}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)\right] \tag{2.5}$$

These are the equations the computer sets equal to zero and solves to find the regression coefficients. Note that the dispersion parameter ϕ appears only multiplicatively. So it cancels when the partial derivatives are set equal to zero. Thus the regression coefficients can be estimated without estimating the dispersion (just as in linear regression).

The variance-covariance matrix of the U_j's has terms

$$\mathcal{I}_{jk} = E[\mathcal{U}_j\mathcal{U}_k] = \frac{1}{\phi}\sum_{i=1}^{N}\left[\frac{x_{ij}x_{ik}}{V(\mu_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right] \tag{2.6}$$

which form the *Fisher Information Matrix* $\mathcal{I}$.

In most cases an explicit expression of the MLE is not available, and $\hat{\boldsymbol{\beta}}$ needs to be computed numerically. Nelder and Wedderburn (1972) proposed *Fisher scoring* as a general method for the numerical evaluation of $\hat{\boldsymbol{\beta}}$ in GLMs. That is, given a trial estimate $\boldsymbol{\beta}^0$, update to $\boldsymbol{\beta}^1$ given by

$$\boldsymbol{\beta}^1 = \boldsymbol{\beta}^0 + \left\{E_{\boldsymbol{\beta}^0}\left(-\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)\right\}^{-1}\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \tag{2.7}$$

where both derivatives are evaluated at $\boldsymbol{\beta}^0$, and the expectation is evaluated as if $\boldsymbol{\beta}^0$ were the true parameter value. Then $\boldsymbol{\beta}^0$ is replaced by $\boldsymbol{\beta}^1$ and the updating is repeated yielding $\boldsymbol{\beta}^0$, $\boldsymbol{\beta}^1$, $\boldsymbol{\beta}^2$, ... with $\boldsymbol{\beta}^m$ tending to $\hat{\boldsymbol{\beta}}$ when m tends to infinity. The updating is stopped when $\boldsymbol{\beta}^m - \boldsymbol{\beta}^{m-1}$ is small enough and $\hat{\boldsymbol{\beta}}$ is taken to be the final $\boldsymbol{\beta}^m$. We notice that this method is very similar to the well-known Newton-Raphson method for finding a zero of the function $\frac{\partial\ell}{\partial\boldsymbol{\beta}}$. Except here the expected value of the derivative of $\frac{\partial\ell}{\partial\boldsymbol{\beta}}$ is used instead of the derivative itself as Newton-Raphson does.

It turns out that for a GLM, the Fisher scoring updating equations (2.7) can be written as

$$\boldsymbol{\beta}^1 = (\boldsymbol{X}^T\boldsymbol{W}^0\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^0\boldsymbol{z}^0, \tag{2.8}$$

where $\boldsymbol{X}$ is the *design* matrix with $\boldsymbol{x}_i^T$ as its i-th row, $\boldsymbol{z}^0$ is the n-vector with i-th component

$$z_i^0 = (y_i - \mu_i^0)g'(\mu_i^0) + \boldsymbol{x}_i^T\boldsymbol{\beta}^0,$$

where $g'(\mu_i^0) = \frac{\partial\eta_i}{\partial\mu_i}$ at μ_i^0, and $\boldsymbol{W}^0$ the $n \times n$ diagonal matrix with i-th diagonal element

$$w_i^0 = (g'(\mu_i^0)^2 b''(\theta_i^0))^{-1}.$$

Thus, for the iteration $\boldsymbol{z}^0$ and $\boldsymbol{W}^0$ are evaluated as if $\boldsymbol{\beta}^0$ were the true parameter value. (For instance, $\mu_i^0 = E_{\boldsymbol{\beta}^0}[Y_i]$.) Expression (2.8) means that each iteration of the Fisher scoring method for numerical evaluation of the MLE is in fact a weighted least squares regression for the "working response vector" $\boldsymbol{z}^0$ on the model matrix $\boldsymbol{X}$ with a 'working weights matrix' $\boldsymbol{W}^0$. Since both $\boldsymbol{z}^0$ and $\boldsymbol{W}^0$ are functions of the current estimate of $\boldsymbol{\beta}$, they need to be re-evaluated each iteration.

For a generalized linear model with log-likelihood ℓ, the *deviance* is defined as

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = -2\phi\left[\ell(\hat{\boldsymbol{\mu}},\phi;\boldsymbol{y}) - \ell(\boldsymbol{y},\phi;\boldsymbol{y})\right], \tag{2.9}$$

where $\ell(\hat{\boldsymbol{\mu}},\phi;\boldsymbol{y})$ is the log-likelihood function for the mean vector $\boldsymbol{\mu}$ and dispersion parameter ϕ, and $\ell(\boldsymbol{y},\phi;\boldsymbol{y})$ is the log-likelihood of the saturated model, i.e., the greatest possible value of ℓ. For the saturated model which matches the data exactly one has (usualy) $\mu_i(\hat{\boldsymbol{\beta}}) = y_i$.

If the responses Y_i are independently distributed and form an exponential family (2.2), then

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n}\left\{y_i[\theta_i(y_i) - \theta_i(\hat{\mu}_i)] - b(\theta_i(y_i)) + b(\theta_i(\hat{\mu}_i))\right\}, \tag{2.10}$$

where $\theta_i(y_i)$ is the value of θ_i that corresponds to $\mu_i = y_i$, and $\theta_i(\hat{\mu}_i)$ is the value of θ_i that corresponds to $\mu_i = \hat{\mu}_i$. In particular, the deviance does not depend on ϕ, only on the observations $\boldsymbol{y}$ and on the maximum likelihood estimate $\hat{\boldsymbol{\mu}}$.

The *scaled deviance* is known as

$$D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = \frac{D(\boldsymbol{y};\hat{\boldsymbol{\mu}})}{\phi}. \tag{2.11}$$

The deviance is a natural measure of model fit (the greater the likelihood, the smaller the deviance), but serious problems, are associated to the use of deviance in assessing the fit of a given model. Its main role is in the comparison of 'competing' models, especially in the choice of explanatory variables to be included in the model. This is also the reason behind the choice for the specific form (2.9) for the expression of the deviance.

In cases where the dispersion parameter, ϕ, is not known, an estimate can be used. To estimate the dispersion, McCullagh and Nelder (1989) recommend the use of:

$$\hat{\phi} = \frac{1}{(n-p)}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \frac{\mathcal{X}_P^2}{(n-p)}, \tag{2.12}$$

where $\mathcal{X}_P^2$ is Pearson's χ^2 statistic and $V(\hat{\mu}_i)$ is the variance function. The maximum likelihood estimator and the usual (or deviance) estimator, $\frac{D}{(n-p)}$, are both sensitive to unusually small values of the response and are not consistent estimates of the coefficient of variation when the gamma distribution assumption does not hold.

2.2 Gamma and Inverse Gaussian GLM

The binomial, Gaussian and Poisson GLMs are by far the most commonly used, but there are a number of less popular GLMs which are useful for particular types of data. The gamma and inverse Gaussian are intended for continuous, skewed responses.

2.2.1 Gamma GLM

The density of the gamma distribution , $Y_i \sim G(\nu, \lambda_i)$ is usually given by:

$$f(y_i) = \frac{1}{\Gamma(\nu)} {\lambda_i}^{-\nu} y^{\nu-1} e^{-\frac{y_i}{\lambda_i}} \quad , \qquad y_i > 0 \tag{2.13}$$

where ν describes the shape and λ_i describes the scale of the distribution. The gamma distribution can arise in various ways. The sum of ν independent and identically distributed exponential random variables with rate λ_i has a gamma distribution. The χ^2 distribution is a special case of the gamma where $\lambda_i = 2$ and $\nu = df/2$. The gamma distribution is useful for modeling a positive continuous response variable, where the conditional variance of the response grows with its mean but where the coefficient of variation of the response, $SD(Y_i)/\mu_i$, is constant. However, for the purposes of a GLM, it is convenient to reparameterize by putting $\lambda_i = \frac{\mu_i}{\nu}$, and the dispersion parameter is $\phi = \frac{1}{\nu}$.

Let us assume that the responses $Y_1, Y_2, ..., Y_n$ are Gamma random variables with $G(\nu, \frac{\mu_i}{\nu})$, so that:

$$\begin{aligned} f(y_i) &= \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu} y_i^{\nu-1} exp\left\{-\frac{y_i \nu}{\mu_i}\right\} \quad , \qquad y_i > 0 \\ &= exp\left\{\frac{-y_i \frac{1}{\mu_i} - \log(\mu_i)}{\frac{1}{\nu}} + \nu \log(y_i \nu) - \log(y_i \Gamma(\nu))\right\}, \end{aligned} \tag{2.14}$$

then we have

$$E(Y_i) = b'(\theta_i) = -\frac{1}{\theta_i} = \mu_i,$$
$$Var(Y_i) = \phi\ b''(\theta_i) = \phi\ V(\mu_i) = \frac{\mu_i^2}{\nu},$$

where $b(\theta_i) = -\log(-\theta_i) = \log(\mu_i)$, and $V(\mu_i) = \mu_i^2$ is the variance function. Notice that, the variance depends on the mean.

Since the canonical parameter is $\theta_i = -\frac{1}{\mu_i}$, so

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{V(\mu_i)} = \frac{1}{\mu_i^2}, \tag{2.15}$$

and

$$\frac{\partial^2 \theta_i}{\partial \mu_i^2} = \frac{-2}{\mu_i^3}. \tag{2.16}$$

The utility of the gamma GLM arises in two different ways. Certainly, if we believe the response to have a gamma distribution, the model is clearly applicable. However, the model can also be useful in other situations where we may be willing to speculate on the relationship between the mean and the variance of the response but are not sure about the distribution. Indeed, it is possible to grasp the mean to variance relationship from graphical displays with relatively small datasets, while assertions about the response distribution would require a lot more data. See Faraway (2006).

In general, there is not just one reasonable link function for a given response variable distribution. For parametric models, the choice of link function can lead to substantively different estimates and tests. McCullagh and Nelder (1989) support the method of minimal deviances. They also check the residuals to observe closeness of fit and the normality and independence of the standardized residuals themselves. All other factors the same, choosing the model that has the least value for the deviance is preferable.

There are three common choices of link functions, one canonical and two noncanonical links:

1. The *canonical inverse* link:

   The canonical parameter is $\theta_i = -\frac{1}{\mu_i}$, so the canonical link is $\eta_i = -\frac{1}{\mu_i}$. However, the canonical link is equivalent to the inverse link in the sense that if

   $$g(\mu_i) = -\frac{1}{\mu_i} = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

   then

   $$\begin{aligned} \frac{1}{\mu_i} &= \boldsymbol{x}_i^T(-\boldsymbol{\beta}) \\ &= \boldsymbol{x}_i^T \boldsymbol{\beta}^* \end{aligned}$$

   where $\boldsymbol{\beta}^* = -\boldsymbol{\beta}$. Since $-\infty < \eta_i < \infty$, the link does not guarantee $\mu_i > 0$ which could cause problems and might require restrictions on $\boldsymbol{\beta}$ or on the range of possible predictor values. This shows that the canonical link might not always be the best choice since $-\frac{1}{\mu_i} = \boldsymbol{x}_i^T \boldsymbol{\beta}$ or $\mu_i = -1/\boldsymbol{x}_i^T \boldsymbol{\beta}$ implies severe restrictions on $\boldsymbol{\beta}$ arising from the restriction that μ_i has to be non-negative.

   In this link function, we have

   $$\begin{aligned} g(\mu_i) &= -\frac{1}{\mu_i} = \boldsymbol{x}_i^T \boldsymbol{\beta} = \eta_i, \\ \mu_i = g^{-1}(\eta_i) &= -\frac{1}{\eta_i} \quad , \quad -\infty < \mu_i < \infty \end{aligned}$$

thus

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{\eta_i^2} = \mu_i^2, \tag{2.17}$$

$$\frac{\partial^2 \mu_i}{\partial \eta_i^2} = -\frac{2}{\eta_i^3} = 2\mu_i^3. \tag{2.18}$$

2. The *log* link:

   This link, $\eta = \log \mu$, should be used when the effect of the predictors is suspected to be multiplicative on the mean. When the variance is small, this approach is similar to a Gaussian model with a logged response. In this link function, $0 \leq g^{-1}(\eta_i) < \infty$.

   In this link function, we have

$$g(\mu_i) = \log(\mu_i) = \eta_i \ ,$$
$$\mu_i = \exp(\eta_i) \quad , \quad 0 < \mu_i < \infty$$

   thus

$$\frac{\partial \mu_i}{\partial \eta_i} = \exp(\eta_i) = \mu_i, \tag{2.19}$$

$$\frac{\partial^2 \mu_i}{\partial \eta_i^2} = \exp(\eta_i) = \mu_i. \tag{2.20}$$

3. The *identity* link:

   This link, $\eta_i = \mu_i$, is useful for modeling sums of squares or variance components which are χ^2. This is a special case of the gamma.

   In this link function, we have

$$g(\mu_i) = \mu_i = \eta_i,$$
$$\mu_i = \eta_i \quad , \quad -\infty < \mu_i < \infty$$

   thus

$$\frac{\partial \mu_i}{\partial \eta_i} = 1, \tag{2.21}$$

$$\frac{\partial^2 \mu_i}{\partial \eta_i^2} = 0. \tag{2.22}$$

These are the link functions which are available in *R*. The log and identity links are the primary noncanonical links associated with the Gamma distribution. One may fit an identity-link and log-link model. Graphs of the competing fitted values versus the outcomes can be subjectively compared. Otherwise, we can determine which model is preferable, between the log and identity links, form the value of the deviance function. Of course, we assume that all other aspects of the data and model are the same. If there is a significant difference between the deviance values between two models, than the model with the lower deviance is preferred. I there is little difference between the two, then either may be used. The same logic may be applied with respect to BIC or AIC statistics. Residual analysis may help in determining model preference as well.

From (2.10), the (unscaled) deviance is:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} \left\{ \log(\frac{\hat{\mu}_i}{y_i}) + \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i} \right\} \quad , \qquad y_i > 0. \tag{2.23}$$

When the dispersion parameter, ϕ, is unknown, it may be estimated by (2.12):

$$\begin{aligned} \hat{\phi} = \frac{1}{\hat{\nu}} &= \frac{1}{(n-p)} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \\ &= \frac{1}{(n-p)} \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right)^2, \end{aligned} \tag{2.24}$$

where $V(\hat{\mu}_i)$, the variance function, here is $\hat{\mu}_i^2$ for the Gamma distribution. It is actually a *moment estimator* and since moment estimators also have the invariance property (like MLEs), we can estimate ν by $\hat{\nu} = \frac{1}{\hat{\phi}}$. The usual (or deviance) estimator

$$\frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{(n-p)} = \frac{2}{(n-p)} \sum_{i=1}^{n} \left\{ \log(\frac{\hat{\mu}_i}{y_i}) + \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i} \right\} \quad , \qquad y_i > 0$$

and the maximum likelihood estimator are both sensitive to unusually small values of the response and are not consistent estimates of the coefficient of variation when the gamma distribution assumption does not hold. In linear regression, the dispersion parameter is not estimated by maximum likelihood but by the method of moments.

2.2.2 Inverse Gaussian GLM

The Inverse Gaussian distribution is an exponential distribution. It is one of the distributions implemented in R's Generalized Linear Model routines. Since We want a distribution that can 'reach up high' and admit some extreme values, it is a good idea to use the distribution. The density of an Inverse Gaussian random variable, $Y_i \sim IG(\mu_i, \lambda)$, is:

Table 2.1: Characteristics of the two GLM Distributions

Distribution	$b(\theta)$	$V(\mu)$	$g(\mu)$
Gamma $(y, \mu > 0)$	$\log(\mu) = -\log(-\theta)$	$\mu^2 = \left(-\frac{1}{\theta}\right)^2$	$-\frac{1}{\mu}$ (canonical) $\log(\mu)$ μ
Inverse Gaussian $(y, \mu > 0)$	$-\frac{1}{\mu} = -\sqrt{-2\theta}$	$\mu^3 = (-2\theta)^{-3/2}$	$-\frac{1}{2\mu^2}$ (canonical) $\frac{1}{\mu}$ $\log(\mu)$ μ

$$\begin{aligned} f(y_i) &= \sqrt{\frac{\lambda}{2\pi y_i^3}} e^{-\frac{\lambda(y_i-\mu_i)^2}{2\mu_i^2 y_i}} \quad , \qquad y_i > 0 \\ &= exp\left\{ \frac{y_i(-\frac{1}{2\mu_i^2}) + 1/\mu_i}{1/\lambda} - \frac{\lambda}{2y_i} - \frac{1}{2}\log(\frac{2\pi y_i^3}{\lambda}) \right\}. \end{aligned} \tag{2.25}$$

so that

$$E(Y_i) = b'(\theta_i) = \frac{1}{\sqrt{-2\theta_i}} = \mu_i,$$

$$Var(Y_i) = \phi\ b''(\theta_i) = \phi\ V(\mu_i) = \frac{\mu_i^3}{\lambda},$$

where $b(\theta_i) = -\sqrt{-2\theta_i} = -\frac{1}{\mu_i}$, and $V(\mu_i) = \mu_i^3$ is the variance function. Notice that, the variance depends on the mean, and the variance function for the Inverse Gaussian GLM increases more rapidly with the mean than the gamma GLM, making it suitable for data where this occurs.

The case of $\mu = 1$ is known as the Wald distribution. The Inverse Gaussian has found application in the modelling of lifetime distributions with non-monotone failure rates and in the first passage times of Brownian motions with drift. See Seshadri (1993).

In particular, it is most appropriate when modeling a nonnegative response having a high initial peak, rapid drop, and long right tail. If a discrete response has many different values together with the same properties, then the Inverse Gaussian may be appropriate for this type of data as well. A variety of other shapes can also be modelled using Inverse Gaussian regression.

Since the canonical parameter is $\theta_i = -\frac{1}{2\mu_i^2}$, so

$$\frac{\partial\theta_i}{\partial\mu_i} = \frac{1}{V(\mu_i)} = \frac{1}{\mu_i^3}, \tag{2.26}$$

and

$$\frac{\partial^2\theta_i}{\partial\mu_i^2} = \frac{-3}{\mu_i^4}. \tag{2.27}$$

The four most commonly used link functions for Inverse Gaussian GLMs, which are available in *R*, are:

1. The *canonical inverse-square* link:

   The canonical parameter is $\theta_i = (-2\mu_i^2)^{-1}$, so the canonical link is $\eta_i = -(\sqrt{2}\mu_i)^{-2}$. In this canonical link, we have

   $$g(\mu_i) = -\frac{1}{2\mu_i^2} = \boldsymbol{x}_i^T\boldsymbol{\beta} = \eta_i,$$
   $$\mu_i = g^{-1}(\eta_i) = \frac{1}{\sqrt{-2\eta_i}} \quad , \quad 0 < \mu_i < \infty$$

   thus

   $$\frac{\partial\mu_i}{\partial\eta_i} = (-2\eta_i)^{-\frac{3}{2}} = \mu_i^3, \tag{2.28}$$

   $$\frac{\partial^2\mu_i}{\partial\eta_i^2} = 3(-2\eta_i)^{-\frac{5}{2}} = 3\mu_i^5. \tag{2.29}$$

2. The *inverse* link ($\eta_i = \frac{1}{\mu_i}$):

   In this link function, we have

   $$g(\mu_i) = \frac{1}{\mu_i} = \eta_i,$$
   $$\mu_i = g^{-1}(\eta_i) = \frac{1}{\eta_i} \quad , \quad -\infty < \mu_i < \infty$$

   thus

   $$\frac{\partial\mu_i}{\partial\eta_i} = -\frac{1}{\eta_i^2} = -\mu_i^2, \tag{2.30}$$

   $$\frac{\partial^2\mu_i}{\partial\eta_i^2} = \frac{2}{\eta_i^3} = 2\mu_i^3. \tag{2.31}$$

3. The *log* link:

   This link, $\eta = \log\mu$, should be used when the effect of the predictors is suspected to be multiplicative on the mean. When the variance is small, this approach is similar to a Gaussian model with a logged response. In this link function, $0 \leq g^{-1}(\eta_i) < \infty$. In this link function we have

$$g(\mu_i) = \log(\mu_i) = \eta_i,$$
$$\mu_i = \exp(\eta_i) \quad , \quad 0 < \mu_i < \infty$$

thus

$$\frac{\partial \mu_i}{\partial \eta_i} = \exp(\eta_i) = \mu_i, \tag{2.32}$$

$$\frac{\partial^2 \mu_i}{\partial \eta_i^2} = \exp(\eta_i) = \mu_i. \tag{2.33}$$

4. The *identity* link:

   In this link function we have

$$g(\mu_i) = \mu_i = \eta_i,$$
$$\mu_i = \eta_i \quad , \quad -\infty < \mu_i < \infty$$

thus

$$\frac{\partial \mu_i}{\partial \eta_i} = 1, \tag{2.34}$$

$$\frac{\partial^2 \mu_i}{\partial \eta_i^2} = 0. \tag{2.35}$$

We can use other statistical tests to assess the worth of differential models among links. These tests include the BIC and AIC. Models having lower values of these criteria are preferable. With respect to BIC, if the absolute difference between the BIC for two models is less than 2, then there is only weak evidence that the model with the smaller BIC is preferable. Absolute difference between 2 and 6 give positive support for the model with the smaller BIC, whereas absolute difference between 6 and 10 offer strong support. Absolute differences more than 10 are very strong support for the model with the smaller BIC.

From (2.10), the (unscaled) deviance is given by:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu_i})^2}{\hat{\mu_i}^2 y_i}. \tag{2.36}$$

When the dispersion parameter, ϕ, is unknown, from (2.12) we have:

$$\begin{aligned}\hat{\phi} = \frac{1}{\hat{\lambda}} &= \frac{1}{(n-p)} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \\ &= \frac{1}{(n-p)} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^3}, \end{aligned} \tag{2.37}$$

where $V(\hat{\mu}_i)$, the variance function, here is $\hat{\mu}_i^3$ for the Inverse Gaussian distribution. It is actually a *moment estimator* and since moment estimators also have the invariance property (like MLEs), we can estimate λ by $\hat{\lambda} = \frac{1}{\hat{\phi}}$. The usual (or deviance) estimator

$$\frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{(n-p)} = \frac{1}{(n-p)} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i{}^2 y_i}, \qquad y_i > 0$$

and the maximum likelihood estimator are both sensitive to unusually small values of the response and are not consistent estimates of the coefficient of variation when the gamma distribution assumption does not hold. Also as in linear regression, the dispersion parameter is not estimated by maximum likelihood but by the method of moments.

Chapter 3

Gamma and Inverse Gaussian dgLARS

In this section, we give a rough overview of the dgLARS method. The reader interested in more of the differential geometric details of this method is referred to Augugliaro *et al.* (2013). The dgLARS method defines a continuous solution path foe a GLM, with on the extreme of the path the maximum likelihood estimate and on the other side the intercept-only estimate. The aim of the method is to define the most efficient model - in likelihood terms - that uses the fewest variables. In order to describe the method, we introduce the differential geometric least angle regression in section 3.1 and Gamma and Inverse Gaussian dgLARS is given in section 3.2.

3.1 Differential Geometric Least Angle Regression

As mentioned in previous chapter, we introduce the GLM (McCullagh and Nelder, 1989) from a differential geometric point of view. In our treatment, we rely heavily on Amari (1985), Kass and Vos (1997) and Amari and Nagaoka (2000). A differential geometric approach was also used in Wei (1998) to study non-linear models based on the exponential family. Essential aspects of differential and information geometry have been included to make the paper self-contained. Let $Y = (Y_1, Y_2, \cdots, Y_n)^T$ be a random vector with independent components. In what follows we shall assume that the ith element of Y, Y_i, is a random variable with probability density function belonging to the exponential dispersion family

$$p_{Y_i}(y_i; \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right], \quad y_i \in \mathcal{Y}_i \subseteq \mathbb{R}, \tag{3.1}$$

where the canonical parameter $\theta_i \in \Theta_i \subseteq \mathbb{R}$, the dispersion parameter $\phi \in \Phi \subseteq \mathbb{R}^+$ and $a(.)$, $b(.)$ and $c(.,.)$ are specific given functions. Under family (3.1), the joint probability density function of the random vector Y can be written as

$$p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta},\phi)=\prod_{i=1}^{n}p_{Y_i}(y_i;\theta_i,\phi),$$

where the canonical parameter $\boldsymbol{\theta}$ varies in the subset $\otimes_{i=1}^{n}\Theta_i=\Theta\subseteq\mathbb{R}^{+}n$. The mean value of $\mathbf{Y}$ is denoted by $\boldsymbol{\mu}=(\mu(\theta_1),\cdots,\mu(\theta_n))^T$, where $\mu(\theta_i)=\frac{\partial b(\theta_i)}{\partial\theta_i}$ is called mean value mapping, and the variance of $\mathbf{Y}$ is equal to $Var(\mathbf{Y})=a(\phi)\mathbf{V}(\boldsymbol{\mu})$, where $\mathbf{V}(\boldsymbol{\mu})$ is an $n\times n$ diagonal matrix with elements $\mathbf{V}(\mu_i)=\frac{\partial^2 b(\theta_i)}{\partial\theta_i^2}$. $\mathbf{V}(.)$ is called the variance function.

3.1.1 Target Space

Since $\mu(.)$ is a one-to-one function from $int(\Theta)$ onto $\tilde{\mathcal{S}}=\mu\{int(\Theta)\}$, $p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta},\phi)$ may be parameterized by $(\boldsymbol{\mu};\phi)$. We consider ϕ as an unknown parameter. Assuming that Θ is open, the set

$$\mathcal{S}=\{p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu}):\boldsymbol{\mu}\in\tilde{\mathcal{S}}\} \tag{3.2}$$

is a minimal and regular exponential family of order n and can be treated as a differential manifold where the parameter vector $\boldsymbol{\mu}$ plays the role of a co-ordinate system (Amari, 1985). The notion of differential manifold is necessary for extending the methods of differential calculus to a space that is more general than $\mathbb{R}^n$. For a rigorous definition of a differential manifold the reader is referred to Spivak (1979) and do Carmo (1992). It is worth noting that the results coming from differential geometry are not related to the chosen co-ordinate system, i.e. the parameterization that is used to specify the probability density function (3.1). This means that we could work with the differential manifold $\mathcal{S}$ using the parameter vector $\boldsymbol{\theta}$ as co-ordinate system. In this paper we prefer to use definition (3.2) only because we believe that this makes the generalization of the LARS algorithm clearer.

As shown in previous section, a GLM is completely specified by the following assumptions:

1. $\mathbf{y}$ is a random observation drawn from the distribution on $\mathbf{Y}$;

2. for each random variable Y_i there is a vector of covariates $\mathbf{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^T\in\mathcal{X}\subseteq\mathbb{R}^p$;

3. $E(Y_i|\mathbf{x}_i)=\mu_i(\boldsymbol{\beta})=g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$, where g is called the *link function.*

In James (2002) an interesting extension of the classical GLM is proposed to handle functional predictors. In the literature this model is known as the generalized functional linear model and was also studied in Müller and Stadtmüller (2005) and Li *et al.* (2010), among others.

To simplify our notation, we denote $\boldsymbol{\mu}(\boldsymbol{\beta})=(g^{-1}(\mathbf{x}_1^T\boldsymbol{\beta}),g^{-1}(\mathbf{x}_2^T\boldsymbol{\beta}),\cdots g^{-1}(\mathbf{x}_n^T\boldsymbol{\beta}))^T$.

Assuming that $\boldsymbol{\beta}\longrightarrow\boldsymbol{\mu}(\boldsymbol{\beta})$ is an embedding, the set

$$\mathcal{M} = \{p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu}(\boldsymbol{\beta})) \in \mathcal{S} : \boldsymbol{\beta} \in \mathbb{R}^p\}$$

is a p-dimensional submanifold of $\mathcal{S}$. To obtain a natural generalization of the equiangularity condition that was proposed by Efron *et al.* (2004), it is necessary to introduce two fundamental notions on which Riemannian geometry is based: the notions of a tangent space and a Riemannian metric. To complete the differential geometric setting for the GLM, we shall assume that the usual regularity conditions hold (Amari (1985), page 16). Throughout this paper we use the convention that the indices i, j and k correspond to the quantities that are related to $\boldsymbol{\mu} \in \tilde{\mathcal{S}}$ whereas the indices l, m and q correspond to the quantities that are related to the coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ of our regression model. Consider a double-differentiable curve, say $\boldsymbol{\mu} : \Gamma \longrightarrow \tilde{\mathcal{S}}$, where Γ is the real interval $(-\delta, \delta)$ with $\delta > 0$. The tangent vector to the one-parametric family $p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu}(\gamma))$ at $\boldsymbol{\mu} = \boldsymbol{\mu}(0)$ is defined as

$$v(\mathbf{Y}) = \left.\frac{dl(\boldsymbol{\mu}(\gamma);\mathbf{Y})}{d\gamma}\right|_{\gamma=0} = \sum_{i=1}^{n} d\mu_i(0)\partial_i\ell(\boldsymbol{\mu};\mathbf{Y}), \tag{3.3}$$

where $d\mu_i(0) = \frac{d\mu_i(\gamma)}{d\gamma} |_{\gamma=0}$ and $\partial_i\ell(\boldsymbol{\mu};\mathbf{Y}) = \frac{\partial \log\{p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu}(\gamma))\}}{\partial\mu_i} |_{\gamma=0}$. Roughly speaking, the tangent space of $\mathcal{S}$ at the point $p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu}(\gamma))$, denoted by $T_{p(\boldsymbol{\mu})}\mathcal{S}$, is the set of all possible tangent vectors at $\boldsymbol{\mu} = \boldsymbol{\mu}(0)$. Formally, $T_{p(\boldsymbol{\mu})}\mathcal{S}$ is the vector space that is spanned by the *n score functions* $\partial_i l(\boldsymbol{\mu};\mathbf{Y})$:

$$T_{p(\boldsymbol{\mu})}\mathcal{S} = span\{\partial_1 l(\boldsymbol{\mu};\mathbf{Y}), \partial_2 l(\boldsymbol{\mu};\mathbf{Y}), \cdots, \partial_p l(\boldsymbol{\mu};\mathbf{Y})\}. \tag{3.4}$$

Under the regularity conditions cited above, $T_{p(\boldsymbol{\mu})}\mathcal{S}$ is a subspace of squared integrable random variables, in which elements $v(\mathbf{Y})$ satisfy the property $E_\mu[v(\mathbf{Y})] = 0$, where the expected value is computed with respect to $p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu})$. As an application of the chain rule, it is easy to see that the definition of a tangent space does not depend on the chosen parameterization; in other words the tangent space can be defined as the vector space that is spanned by the n score functions $\partial_i^* l(\boldsymbol{\theta};\mathbf{Y}) = \frac{\partial \log\{p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}(\gamma))\}}{\partial\theta_i} |_{\gamma=0}$ where $\boldsymbol{\theta}(\gamma) = \theta((\boldsymbol{\mu}(\gamma)))$. Using the terminology that was introduced in Vos (1991), $\partial_i^l(\boldsymbol{\mu};\mathbf{Y})$ are the natural bases of the tangent space when we choose $\boldsymbol{\mu}$ as co-ordinate system, whereas $\partial_i^* l(\boldsymbol{\theta};\mathbf{Y})$ are the natural bases when $\boldsymbol{\theta}$ is used as the co-ordinate system.

Similarly, consider a double-differentiable curve $\boldsymbol{\beta} : \Gamma' \longrightarrow \mathbb{R}^p$, with $\Gamma' = (-\delta', \delta')$ and $\delta' > 0$. The tangent vector to the one-parametric family $p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma)))$ at the point $\boldsymbol{\beta} = \boldsymbol{\beta}(0)$ is defined as

$$\omega(\mathbf{Y}) = \sum_{m=1}^{p} d\beta_m(0)\partial_m\ell(\boldsymbol{\beta};\mathbf{Y}),$$

where $d\beta_m(0) = \frac{d\beta_m(\gamma)}{d\gamma} |_{\gamma=0}=$ and $\partial_m l(\boldsymbol{\beta};\mathbf{Y}) = \frac{\partial \log\{p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma)))\}}{\partial\beta_m} |_{\gamma=0}$. Then, the tangent space of $\mathcal{M}$ at the point $p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma)))$ is

$$T_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}\mathcal{M} = span\{\partial_1 l(\boldsymbol{\beta};\mathbf{Y}), \partial_2 l(\boldsymbol{\beta};\mathbf{Y}), \cdots, \partial_p l(\boldsymbol{\beta};\mathbf{Y})\}.$$

The definition of the inner product on each tangent space allows us to generalize the notion of angle between two curves, say $\boldsymbol{\mu}_1(\gamma)$ and $\boldsymbol{\mu}_2(\gamma)$, intersecting at $\boldsymbol{\mu}_1(0) = \boldsymbol{\mu}_2(0) = \boldsymbol{\mu}$, with tangent vectors belonging to $T_{p(\boldsymbol{\mu})}\mathcal{S}$, denoted by

$$\upsilon_1(\mathbf{Y}) = \sum_{i=1}^{n} d\mu_{1,i}(0)\partial_i \ell(\boldsymbol{\mu};\mathbf{Y})$$

and

$$\upsilon_2(\mathbf{Y}) = \sum_{i=1}^{n} d\mu_{2,i}(0)\partial_i \ell(\boldsymbol{\mu};\mathbf{Y})$$

respectively.When working with a parametric family of distributions, the inner product can be defined in a natural way (Rao, 1945), i.e.

$$\langle \upsilon_1(\mathbf{Y}), \upsilon_2(\mathbf{Y})\rangle_{p(\boldsymbol{\mu})} = E_{\boldsymbol{\mu}}[\upsilon_1(\mathbf{Y})\upsilon_2(\mathbf{Y})] = d\boldsymbol{\mu}_1(0)^T \mathcal{I}(\boldsymbol{\mu}) d\boldsymbol{\mu}_2(0),$$

where $\mathcal{I}(\boldsymbol{\mu})$ is the Fisher information matrix for the mean parameter at point $\boldsymbol{\mu}$. In other words, the Fisher information defines a Riemannian metric by associating with each point of $\mathcal{S}$ an inner product on the tangent space. This Riemannian metric is also called the *information metric* (Burbea and Rao, 1982). Since $T_{p(\boldsymbol{\mu}(\boldsymbol{\beta}))}\mathcal{M}$ is a linear subspace of $T_{p(\boldsymbol{\mu}(\boldsymbol{\beta}))}\mathcal{S}$, the Fisher information also defines an inner product on $T_{p(\boldsymbol{\mu}(\boldsymbol{\beta}))}\mathcal{M}$. Therefore, we can define the inner product between a tangent vector $\omega(\mathbf{Y})$ of $T_{p(\boldsymbol{\mu}(\boldsymbol{\beta}))}\mathcal{M}$ and a tangent vector $\upsilon(\mathbf{Y})$ of $T_{p(\boldsymbol{\mu}(\boldsymbol{\beta}))}\mathcal{S}$, namely

$$\langle \omega(\mathbf{Y}), \upsilon(\mathbf{Y})\rangle_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}} = E_{\boldsymbol{\mu}(\boldsymbol{\beta})}[\omega(\mathbf{Y})\upsilon(\mathbf{Y})] = d\boldsymbol{\beta}(0)^T \frac{\partial \boldsymbol{\mu}(\boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}} \mathcal{I}\{\boldsymbol{\mu}(\boldsymbol{\beta})\} d\boldsymbol{\mu}(0),$$

where $\frac{\partial \boldsymbol{\mu}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ is the Jacobian matrix of the vector function $\boldsymbol{\mu}(\boldsymbol{\beta})$.

Each Riemannian metric defines the notion of a geodesic, i.e. the generalization of a straight line in a differential geometric framework. Roughly speaking, a geodesic can be defined as the shortest path between two given points on a differential manifold. A geodesic is defined as the solution of a system of differential equations, the Euler-Lagrange equations, obtained from defining a connection on a differentiable manifold. In statistical theory a one-parametric family of connections plays a fundamental role, the so-called α-connections, denoted by ∇^α, that generalize the classical notion of a Levi-Civita connection, which is the special case that $\alpha = 0$. In the theory of information geometry, ∇^0 is also called the *information connection* since it is derived from the Fisher information. What is also important for following this paper is that $\mathcal{S}$ is a dually flat space, namely, it is flat with respect to the 1- and -1-connection. In this paper we shall not discuss the details of this dual geometry. For a complete treatment the reader is referred to Amari and Nagaoka (2000). As shown in Vos (1991), associated with the -1-connection and each point $p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu})$ there is a diffeomorphism between a neighbourhood of the origin in $T_{p(\boldsymbol{\mu})}\mathcal{S}$ and a neighbourhood of $p_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\mu})$, called the -1-*exponential map.* The dual nature that exists between ∇^{-1} and ∇^{1} defines the dual of the -1-exponential map, namely the so-called 1-exponential map. Since $\mathcal{S}$ is a dually flat space, the inverses of the two exponential maps are well defined. To complete the geometrical framework that is needed to generalize the LARS algorithm,we consider the inverse of the -1-exponential map,which relates the observed response variable $\mathbf{y}$ to the tangent spaces. Vos (1991) defined what we call the *tangent residual vector*

$$\boldsymbol{r}(\boldsymbol{\mu}(\boldsymbol{\beta}), \mathbf{y}; \mathbf{Y}) = \sum_{i=1}^{n} \{y_i - \mu_i(\boldsymbol{\beta})\} \partial_i \ell(\boldsymbol{\mu}(\boldsymbol{\beta}); \mathbf{Y}), \tag{3.5}$$

where $\partial_i l(\boldsymbol{\mu}(\boldsymbol{\beta}); \mathbf{Y}) = \frac{\partial l(\boldsymbol{\mu}; \mathbf{Y})}{\partial \mu_i} \mid_{\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\beta})}$. We define the tangent residual vector (3.5) with respect to both the fixed observations $\mathbf{y}$ and the random variable $\mathbf{Y}$, in such away that it is a random variable with zero expected value and finite variance, and therefore $\boldsymbol{r}(\boldsymbol{\mu}(\boldsymbol{\beta}), \mathbf{y}; \mathbf{Y}) \in T_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}\mathcal{S}$. Vos (1991) showed that it is possible to give a differential geometric interpretation of the maximum likelihood estimator by using the tangent residual vector and the tangent space $T_{p\{\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})\}}\mathcal{M}$, namely $\hat{\beta}$ is the maximum likelihood estimate of β when the tangent residual vector is orthogonal to the tangent space $T_{p\{\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})\}}\mathcal{M}$. It is worth noting that this statement is well defined even if $\mathbf{y}$ is not an element of the mean value parameter space $\tilde{\mathcal{S}}$. In otherwords, the differential geometric description of the maximum likelihood estimator can be used even if the Kullback-Leibler divergence is not defined (Vos, 1991).

3.1.2 Angles and Equiangularity

A GLM relates a linear combination of covariates via a link function to the distribution of the observations. If it is not known which covariates are actually predictive for the outcome, various procedures have been proposed to zoom in on the most relevant features. A large group of stepwise procedures were the first attempt to 'select' variables (Hocking, 1976). Principal component regression (Jolliffe, 1982) was a recognition of the fact that similar information could be present in several variables. The lasso (Tibshirani, 1996) heralded the era of path algorithms, which often indirectly select variables owing to a *pleasant coincidence* between the geometry of the model and the choice of penalty. Least angle regression (Efron *et al.*, 2004) was originally intended as a computational tool. In this paper, however, we shall present this algorithm as a principled method for directly connecting the geometry of the model to the sparsity of the feature space. In other words, least angle regression is not only 'an important contribution to statistical computing' (Madigan and Ridgeway, 2004) but also a new method in its own right: it can be generalized to any model and its success does not depend on the arbitrary match of the constraint and the objective function.

The original LARS algorithm defines a solution path of a linear regression model by sequentially adding variables to the solution. Starting with only the intercept, the LARS algorithm finds the covariate that is most correlated with the response variable and

Table 3.1: Overview of the dgLARS method to compute the solution curve

Step	Algorithm
1	Start with the intercept-only model
2	Repeat
3	Increase the parameters of the active variables keeping the angles between their scores and residual tangent vector the same
4	If the angle of a not-included variable is the same as the ones currently in the model include that variable in the active set
5	Until a stopping rule is met

proceeds in this 'direction' by changing its associated linear parameter. The algorithm takes the largest step possible in the direction of this covariate until some other covariate has as much correlation with the current residual as the current covariate. At that point the LARS algorithm proceeds in an equiangular direction between the two covariates until a new covariate earns its way into the *equally most correlated set.* Then it proceeds in the direction in which the residual makes an equal angle with the three covariates, and so on. For an extensive review of this method, the reader is referred to Hesterberg *et al.* (2008). In this section we generalize these notions for GLMs by using differential geometry. Table 2 gives an overview of the method.

In the linear regression model, the notion of the angle between the covariates and the residual is independent from the form of the model space simply because the model is defined as the collection of linear combinations of the covariates. The linearity of the models results in the piecewise linearity of the LARS solution paths. For a GLM, the effect of any covariate on the residual is moderated by the link function and the parameterization. In this section we describe how the geometrical setting that was introduced in Subsection 3.1.2 can be used to define a genuine generalization of the LARS algorithm for GLMs. In what follows we shall assume that all models include an intercept.

Let $\hat{\beta}_{a_0}$ be the maximum likelihood estimate of the intercept β_{a_0} within the intercept-only log-likelihood $\ell(\boldsymbol{\mu}(\beta_{a_0}); \mathbf{y})$, which is used as the starting point of the proposed generalization. In our approach the use of the maximum likelihood estimator is limited to the starting point. As noted above, the tangent residual vector $\boldsymbol{r}(\boldsymbol{\mu}(\hat{\beta}_{a_0}), \mathbf{y}; \mathbf{Y})$ is orthogonal to the basis $\partial_{a_0}\ell(\hat{\beta}_{a_0}; \mathbf{Y})$ of the tangent space $T_{p\{\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{\boldsymbol{a_0}})\}}\mathcal{M}$. The tangent residual vector can be used to rank the covariates locally by using the notion of angle defined on the tangent space. As shown in Fig. 1(a), the method proposed finds that covariate, say $\mathbf{x}_{a_1}$, whose basis vector $\partial_{a_1}\ell(\hat{\beta}_{a_0}; \mathbf{Y})$ has the smallest angle with the tangent residual vector.

The method then includes the covariate $\mathbf{x}_{a_1}$ in the active set $\mathcal{A}(\gamma^{(1)}) = \{a_0, a_1\}$. The solution curve $\boldsymbol{\beta}(\gamma) = (\beta_{a_0}(\gamma), \beta_{a_1}(\gamma))^T$ is chosen in such a way that it satisfies the condition that the tangent residual vector is always orthogonal to the basis $\partial_{a_0}\ell(\boldsymbol{\beta}(\gamma); \mathbf{Y})$. The direction of the curve $\boldsymbol{\beta}(\gamma)$ is defined by the projection of the tangent residual vector on the basis vector $\partial_{a_1}\ell(\boldsymbol{\beta}(\gamma); \mathbf{Y})$.

The curve $\boldsymbol{\beta}(\gamma)$ continues as defined above until $\gamma^{(2)}$, for which there is a covariate, say $\mathbf{x}_{a_2}$, that satisfies the equiangularity condition on the tangent space $T_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma^{(2)}))\}}\mathcal{M}$, in other words

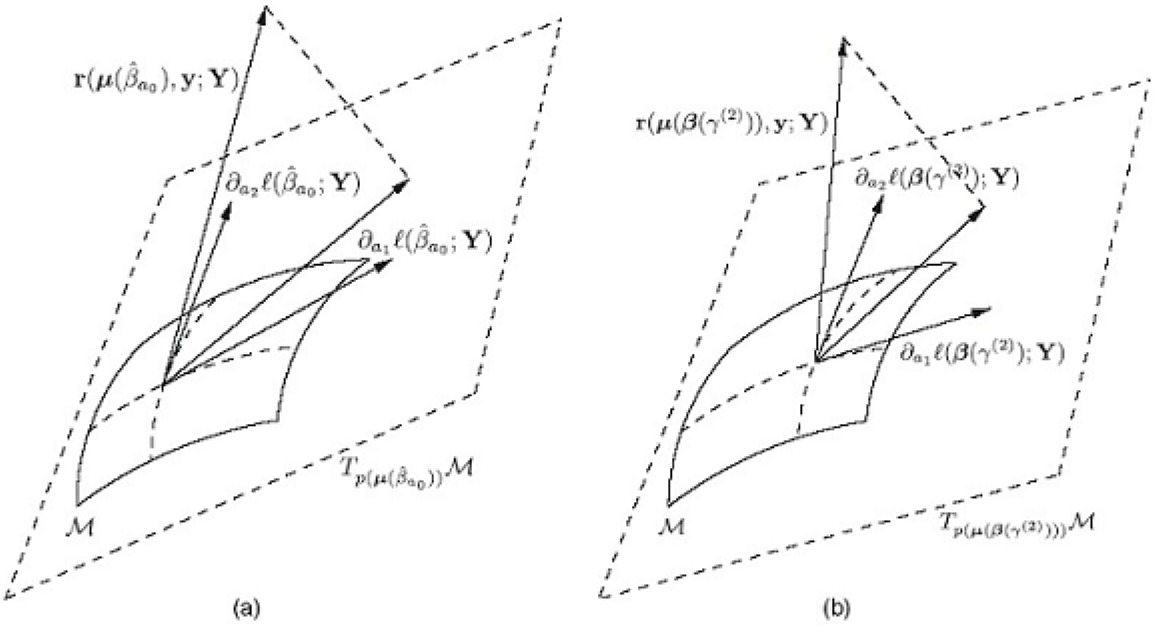
$\mathbf{r}(\boldsymbol{\mu}(\hat{\beta}_{a_0}), \mathbf{y}; \mathbf{Y})$
$\partial_{a_2}\ell(\hat{\beta}_{a_0}; \mathbf{Y})$
$\partial_{a_1}\ell(\hat{\beta}_{a_0}; \mathbf{Y})$
$T_{p(\boldsymbol{\mu}(\hat{\beta}_{a_0}))}\mathcal{M}$
$\mathcal{M}$
(a)
$\mathbf{r}(\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma^{(2)})), \mathbf{y}; \mathbf{Y})$
$\partial_{a_2}\ell(\boldsymbol{\beta}(\gamma^{(2)}); \mathbf{Y})$
$\partial_{a_1}\ell(\boldsymbol{\beta}(\gamma^{(2)}); \mathbf{Y})$
$\mathcal{M}$
$T_{p(\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma^{(2)})))}\mathcal{M}$
(b)

Figure 3.1: Differential geometrical description of the LARS algorithm for a GLM with two covariates: (a) the first covariate x_{a_1} is found and included in the active set; (b) the generalized equiangularity condition (3.10) is satisfied for variables x_{a_1} and x_{a_2}.

$$\rho_{a_1}\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma^{(2)}))\} = \rho_{a_2}\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma^{(2)}))\},$$

where $\rho_m\{\boldsymbol{\mu}(\boldsymbol{\beta})\}$ is the angle between the tangent residual vector and the basis vector $\partial_m\ell(\boldsymbol{\beta}(\gamma); \mathbf{Y})$. At this point $\mathbf{x}_{a_2}$ is included in the active set $\mathcal{A}(\gamma^{(2)})$ and a new curve $\boldsymbol{\beta}(\gamma) = (\beta_{a_0}(\gamma), \beta_{a_1}(\gamma), \beta_{a_2}(\gamma))^T$ is defined, such that the tangent residual vector is always orthogonal to the basis vector $\partial_{a_0}\ell(\boldsymbol{\beta}(\gamma); \mathbf{Y})$ with direction defined by the tangent vector that bisects the angle between the basis vectors $\partial_{a_1}\ell(\boldsymbol{\beta}(\gamma); \mathbf{Y})$ and $\partial_{a_2}\ell(\boldsymbol{\beta}(\gamma); \mathbf{Y})$, as shown in Fig. 1(b).

We note that, in principle, we treat the intercept differently from the other covariates. Unless there are some special reasons to do otherwise, the intercept will always be included. Therefore, we do not 'penalize' the intercept, in the sense that the tangent residual vector is constrained to be always orthogonal to the basis vector $\partial_{a_0}\ell(\boldsymbol{\beta}(\gamma); \mathbf{Y})$. This means that the tangent residual vector contains only information on the covariates. Although the proposed generalization is based on the idea of using $\hat{\beta}_{a_0}$ as the starting point, when $\boldsymbol{\mu}(0) \in \tilde{\mathcal{S}}$, it can be modified to deal with models without the intercept term. In this case $\boldsymbol{r}(\boldsymbol{\mu}(0), \mathbf{y}; \mathbf{Y})$ is used to rank the covariates locally. This modification can be used for several important models such as the logistic regression model and the Poisson regression model, in both cases with and without an intercept term.

3.1.3 Rao Score Statistic and Generalized Equiangularity

The derivative of the log-likelihood $\ell(\boldsymbol{\beta}(\gamma); \mathbf{y})$ with respect to the mth covariate parameter can be written as the inner product between the current tangent residual vector $\boldsymbol{r}(\boldsymbol{\beta}(\gamma), \mathbf{y}; \mathbf{Y})$ and the mth base of the tangent space of $\mathcal{M}$,

$$\partial_m \ell(\boldsymbol{\beta}(\gamma); \mathbf{y}) = \langle\ \partial_m \ell(\boldsymbol{\beta}(\gamma)\ ; \mathbf{Y});\ \boldsymbol{r}(\boldsymbol{\beta}(\gamma), \mathbf{y}; \mathbf{Y})\ \rangle_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}. \tag{3.6}$$

Using the law of cosines, this expression is equivalent to

$$\begin{aligned}\partial_m \ell(\boldsymbol{\beta}(\gamma); \mathbf{y}) &= \cos[\rho_m\{\boldsymbol{\beta}(\gamma)\}]\ \|\boldsymbol{r}(\boldsymbol{\beta}(\gamma), \mathbf{y}; \mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}\ \|\partial_m \ell(\boldsymbol{\beta}(\gamma)\ ; \mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}} \\ &= \cos[\rho_m\{\boldsymbol{\beta}(\gamma)\}]\ \|\boldsymbol{r}(\boldsymbol{\beta}(\gamma), \mathbf{y}; \mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}\ \mathcal{I}_m^{1/2}\{\boldsymbol{\beta}(\gamma)\},\end{aligned} \tag{3.7}$$

where $\rho_m\{\boldsymbol{\beta}(\gamma)\}$ is the angle between the tangent residual vector $\boldsymbol{r}(\boldsymbol{\beta}(\gamma), \mathbf{y}; \mathbf{Y})$ and the mth basis function $\partial_m \ell(\boldsymbol{\beta}(\gamma)\ ; \mathbf{Y})$, $\|\cdot\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}$ is the norm defined on $T_{p\{\boldsymbol{\mu}(\boldsymbol{\beta})\}}\mathcal{M}$ and $\mathcal{I}_m\{\boldsymbol{\beta}(\gamma)\}$ is the *Fisher information* for $\beta_m(\gamma)$. Importantly, equation (3.7) shows that the gradient of the log-likelihood function does not generalize the equiangularity condition that was proposed in Efron *et al.* (2004) to define the LARS algorithm, since the latter does not consider the variation related to the square root of the Fisher information $\mathcal{I}_m^{1/2}\{\boldsymbol{\beta}(\gamma)\}$, which in the case of a GLM is typically not constant. Using equation (3.7), the angle $\rho_m\{\boldsymbol{\beta}(\gamma)\}$ can be written as

$$\rho_m\{\boldsymbol{\beta}(\gamma)\} = \cos^{-1}\left[\frac{\partial_m \ell(\boldsymbol{\beta}(\gamma); \mathbf{y})}{\|\boldsymbol{r}(\boldsymbol{\beta}(\gamma), \mathbf{y}; \mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}\ \mathcal{I}_m^{1/2}\{\boldsymbol{\beta}(\gamma)\}}\right].$$

We can define the equiangularity condition directly on $\rho_m\{\boldsymbol{\beta}(\gamma)\}$ as in the case of LARS, but it is easier and more intuitive to define the same condition on a transformation of the same quantity. Let $r_m(\gamma)$ be the signed *Rao score test statistic*, where

$$\begin{aligned}r_m(\gamma) &= \frac{\partial_m \ell(\boldsymbol{\beta}(\gamma); \mathbf{y})}{\mathcal{I}_m^{1/2}\{\boldsymbol{\beta}(\gamma)\}} && (3.8)\\ &= \cos[\rho_m\{\boldsymbol{\beta}(\gamma)\}]\ \|\boldsymbol{r}(\boldsymbol{\beta}(\gamma), \mathbf{y}; \mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}. && (3.9)\end{aligned}$$

Note that the inverse cosine is a strictly increasing function on its restricted domain and that $\|\boldsymbol{r}(\boldsymbol{\beta}(\gamma), \mathbf{y}; \mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}$ does not depend on m. Therefore, the signed Rao score test statistic contains the same information as the angle $\rho_m\{\boldsymbol{\beta}(\gamma)\}$. As a result, for GLMs we can define dgLARS with respect to the Rao score test statistics, rather than the angles.

Furthermore, we note that the Rao score test statistic as defined by equation (3.9) breaks down into a variable selection part, $\cos[\rho_m\{\boldsymbol{\beta}(\gamma)\}]$, and a measure of global fit of the model, $\|\boldsymbol{r}(\boldsymbol{\beta}(\gamma), \mathbf{y}; \mathbf{Y})\|_{p\{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\}}$. In contrast, in the form (3.8), the Rao score test statistic stresses its invariance to any one-to-one reparameterization of the form $\zeta_m = \zeta_m(\beta_m)$. In Efron *et al.* (2004) this aspect is not treated, because they assumed that for all m the information $\mathcal{I}_m(\beta)$ is equal to 1. In this way they could drop $\mathcal{I}_m^{-1/2}\{\boldsymbol{\beta}(\gamma)\}$ and focus only on the derivative of the log-likelihood function, i.e. the covariance between x_m and the tangent residual vector.

The solution curve, which is denoted by $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma) \subset \mathbb{R}^{k+1}$, with $\gamma \in [0, \gamma^{(1)}]$, whereby

$$0 \leqslant \gamma^{(p)} \leqslant \cdots \leqslant \gamma^{(2)} \leqslant \gamma^{(1)},$$

is defined in the following way: for any $\gamma \in (\gamma^{(k+1)}, \gamma^{(k)}]$, $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)$ is chosen in such a way that

$$\begin{aligned} &\mathcal{A}(\gamma) = \{a_1, a_2, \cdots, a_k\}, && \\ &|r_{a_i}(\gamma)| = |r_{a_j}(\gamma)|, && \forall a_i, a_j \in \mathcal{A}(\gamma), \\ &|r_{a_h^c}(\gamma)| < |r_{a_i}(\gamma)|, && \forall a_h^c \in \mathcal{A}^c(\gamma) \text{ and } \forall a_i \in \mathcal{A}(\gamma). \end{aligned} \tag{3.10}$$

In what follows we shall call expression (3.10) the *generalized equiangularity condition.* When $\gamma = \gamma^{(j)}$, with $j = 2, \cdots, p$, the following condition is satisfied:

$$\exists a_h^c \in \mathcal{A}^c(\gamma): \quad |r_{a_h^c}(\gamma^{(j)})| = |r_{a_i}(\gamma^{(j)})|, \qquad \forall a_i \in \mathcal{A}(\gamma); \tag{3.11}$$

in this case a new covariate is included in the active set.

3.1.4 Predictor-Corrector Algorithm

To compute the solution curve we use the predictor-corrector algorithm (Allgower and Georg, 2003). The basic idea underlying the predictor?corrector algorithm is to trace a curve implicitly defined by a system of non-linear equations. The curve is obtained by generating a sequence of points satisfying a chosen tolerance criterion. A predictor?corrector algorithm was also used in Park and Hastie (2007) to compute the path of the coefficients of a GLM with L_1-penalty function.

Let us suppose that k covariates are included in the active set, $\mathcal{A}(\gamma) = \{a_1, a_2, \cdots, a_k\}$. Using the generalized equiangularity condition (3.10), the solution curve satisfies the relationship

$$|r_{a_1}(\gamma)| = |r_{a_2}(\gamma)| = \cdots = |r_{a_k}(\gamma)|, \tag{3.12}$$

for any $\gamma \in (\gamma^{(k+1)}, \gamma^{(k)}]$. Let $\mathbf{v}_k = \{v_{a_1}, v_{a_2}, \cdots, v_{a_k}\}$ be the vector such that $v_{a_j} = \operatorname{sgn}\{r_{a_j}(\gamma^{(k)})\}$ the solution curve $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)$ is implicitly defined by the following system of $k{+}1$ non-linear equations:

$$\left.\begin{array}{rcl}\partial_{a_0}\ell(\boldsymbol{\beta}(\gamma);\mathbf{y}) & = & 0\ ,\\ r_{a_1}(\gamma) & = & v_{a_1}\gamma\ ,\\ \vdots & & \vdots\\ r_{a_k}(\gamma) & = & v_{a_k}\gamma\ .\end{array}\right\} \tag{3.13}$$

When $\gamma = 0$ we obtain the maximum likelihood estimates of the subset of the parameter vector $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$, of the covariates in the active set. The point $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma^{k+1})$ lies on the solution curve joining $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma^k)$ with $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$. To simplify our notation, we define $\tilde{\boldsymbol{\varphi}}_{\mathcal{A}}(\gamma) = \boldsymbol{\varphi}_{\mathcal{A}}(\gamma) - \mathbf{v}_{\mathcal{A}}\gamma$, where $\boldsymbol{\varphi}_{\mathcal{A}}(\gamma) = (\partial_{a_0}\ell(\boldsymbol{\beta}(\gamma);\mathbf{y}), r_{a_1}(\gamma), \cdots, r_{a_k}(\gamma))^T$ and $\mathbf{v}_{\mathcal{A}} = (0, \mathbf{v}_k)^T$. If the model is with no intercept and satisfies the condition that was seen at the end of Subsection 3.1.2, then $\boldsymbol{\varphi}_{\mathcal{A}}(\gamma) = (r_{a_1}(\gamma), \cdots, r_{a_k}(\gamma))^T$ and $\mathbf{v}_{\mathcal{A}} = \mathbf{v}_k$. By differentiating $\tilde{\boldsymbol{\varphi}}_{\mathcal{A}}(\gamma)$ with respect to γ, we obtain

$$\frac{d\tilde{\boldsymbol{\varphi}}_{\mathcal{A}}(\gamma)}{d\gamma} = \frac{\partial\boldsymbol{\varphi}_{\mathcal{A}}(\gamma)}{\partial\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)}\frac{d\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)}{d\gamma} - \mathbf{v}_{\mathcal{A}} = \mathbf{0}, \tag{3.14}$$

where $\frac{\partial\boldsymbol{\varphi}_{\mathcal{A}}(\gamma)}{\partial\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)}$ is the Jacobian matrix of the vector function $\boldsymbol{\varphi}_{\mathcal{A}}(\gamma)$ evaluated at the point $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)$. Using expression (3.14), we can locally approximate the solution curve by the expression

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma - \Delta\gamma) \approx \tilde{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma - \Delta\gamma) = \hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma) - \left(\frac{\partial\boldsymbol{\varphi}_{\mathcal{A}}(\gamma)}{\partial\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)}\right)^{-1}\mathbf{v}_{\mathcal{A}}\Delta\gamma, \tag{3.15}$$

where $\Delta\gamma \in [0; \gamma - \gamma^{(k+1)}]$. We use expression (3.15) for the predictor step of the proposed algorithm. An efficient implementation of the predictor-corrector method requires a suitable method to compute the step size $\Delta\gamma$. Several methods have been proposed in the literature to solve this problem. For example, we can consider a fixed value of $\Delta\gamma$ or we can relate the step size with a fixed variation in the arc length parameterization of the solution curve (see chapter 6 in Allgower and Georg (2003) for further details). In this paper, we use the method that was proposed in Park and Hastie (2007), namely we consider the step size that changes the active set. Using expression (3.11), we have a change in the active set when

$$\exists a_h^c \in \mathcal{A}^c(\gamma): \quad |r_{a_h^c}(\gamma - \Delta\gamma)| = |r_{a_i}(\gamma - \Delta\gamma)|, \qquad \forall a_i \in \mathcal{A}(\gamma); \tag{3.16}$$

Expanding $r_{a_h^c}(\gamma)$ in a Taylor series around γ, we consider the expression

$$|r_{a_h^c}(\gamma - \Delta\gamma)| \approx \left| r_{a_h^c}(\gamma) - \frac{dr_{a_h^c}(\gamma)}{d\gamma}\Delta\gamma \right|.$$

Then, observing that the solution curve satisfies system (3.13), it is easy to see that the following identity holds:

$$|r_{a_i}(\gamma - \Delta\gamma)| = (\gamma - \Delta\gamma), \qquad\qquad \Delta\gamma \in [0;\gamma].$$

By combining these two results, condition (3.16) can be rewritten in the following way:

$$\exists a_h^c \in \mathcal{A}^c(\gamma): \quad \left| r_{a_h^c}(\gamma) - \frac{dr_{a_h^c}(\gamma)}{d\gamma}\Delta\gamma \right| \approx \gamma - \Delta\gamma, \qquad \forall a_i \in \mathcal{A}(\gamma) \text{ and } \Delta\gamma \in [0;\gamma]$$

then

$$r_{a_h^c}(\gamma) \approx \frac{dr_{a_h^c}(\gamma)}{d\gamma}\Delta\gamma \pm (\gamma - \Delta\gamma),$$

so that it gives us two values for $\Delta\gamma$, namely $\dfrac{\gamma - r_{a_h^c}(\gamma)}{1 - \dfrac{dr_{a_h^c}(\gamma)}{d\gamma}}$ or $\dfrac{\gamma + r_{a_h^c}(\gamma)}{1 + \dfrac{dr_{a_h^c}(\gamma)}{d\gamma}}$. Although the minimum of these two values is desired to be selected it should be positive and also less than γ, because $0 \leq \Delta\gamma \leq \gamma$. Therefore, we have

$$\Delta\gamma^{opt} = \min_{a_h^c \in \mathcal{A}^c(\gamma)}{}^{+} \left\{ \frac{\gamma - r_{a_h^c}(\gamma)}{1 - \dfrac{dr_{a_h^c}(\gamma)}{d\gamma}}; \frac{\gamma + r_{a_h^c}(\gamma)}{1 + \dfrac{dr_{a_h^c}(\gamma)}{d\gamma}} \right\}. \tag{3.17}$$

where

$$\begin{aligned}
\frac{dr_{a_h^c}(\gamma)}{d\gamma} &= \frac{d}{d\gamma}\left(\frac{\partial_{a_h^c}\ell(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma);\mathbf{y})}{\sqrt{\mathcal{I}_{a_h^c}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}} \right) \\
&= \frac{\frac{d\ \partial_{a_h^c}\ell(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma);\mathbf{y})}{d\gamma} \cdot \mathcal{I}_{a_h^c}^{1/2}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) - \partial_{a_h^c}\ell(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma);\mathbf{y}) \cdot \frac{d\ \mathcal{I}_{a_h^c}^{1/2}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{d\gamma}}{\mathcal{I}_{a_h^c}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))} \\
&= \mathcal{I}_{a_h^c}^{-1/2}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) \cdot \langle \partial_{a_i a_h^c}\ell(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma);\mathbf{y}), \frac{d\hat{\beta}_{a_i}(\gamma)}{d\gamma} \rangle \\
&- \frac{1}{2}\ r_{a_h^c}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) \cdot \mathcal{I}_{a_h^c}^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) \cdot \langle \partial_{a_i}\mathcal{I}_{a_h^c}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)), \frac{d\hat{\beta}_{a_i}(\gamma)}{d\gamma} \rangle\ , \quad \forall a_h^c \in \mathcal{A}^c(\gamma) \text{ and } \forall a_i \in \mathcal{A}(\gamma) \\
&= \frac{\sum_{a_i \in \mathcal{A}} \partial_{a_i a_h^c}\ell(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma);\mathbf{y})\ \frac{d\hat{\beta}_{a_i}(\gamma)}{d\gamma}}{\mathcal{I}_{a_h^c}^{1/2}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))} - \frac{1}{2}\ \frac{r_{a_h^c}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) \sum_{a_i \in \mathcal{A}} \partial_{a_i}\mathcal{I}_{a_h^c}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))\ \frac{d\hat{\beta}_{a_i}(\gamma)}{d\gamma}}{\mathcal{I}_{a_h^c}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))} \\
&= \sum_{a_i \in \mathcal{A}} \frac{d\hat{\beta}_{a_i}(\gamma)}{d\gamma} \left(\frac{\partial_{a_i a_h^c}\ell(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma);\mathbf{y})}{\mathcal{I}_{a_h^c}^{1/2}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))} - \frac{1}{2}\ \frac{r_{a_h^c}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)) \cdot \partial_{a_i}\mathcal{I}_{a_h^c}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))}{\mathcal{I}_{a_h^c}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))} \right).
\end{aligned} \tag{3.18}$$

where $\frac{d\hat{\beta}_{a_i}(\gamma)}{d\gamma}$ is an element of $\frac{d\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)}{d\gamma} = \left(\frac{\partial\boldsymbol{\varphi}_{\mathcal{A}}(\gamma)}{\partial\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)}\right)^{-1}\mathbf{v}_{\mathcal{A}}$, and $\partial_{a_i a_h^c}\ell(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma);\mathbf{y})$ and $\partial_{a_i}\mathcal{I}_{a_h^c}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma))$ are given in (3.20) and (3.22), respectively.

Table 3.2: Pseudocode of the developed algorithm to compute the solution curve defined by the dgLARS method for a model with the intercept

Step	Algorithm
1	Compute $\hat{\beta}_{a_0}$
2	$\mathcal{A} = \arg\max_{a_j^c \in \mathcal{A}^c} \lvert r_{a_j^c}(\hat{\beta}_{a_0}) \rvert$ and $\gamma = \lvert r_{a_1}(\hat{\beta}_{a_0}) \rvert$
3	Repeat
4	Use equation (3.17) to compute $\triangle\gamma^{opt}$ and set $\gamma = \gamma - \triangle\gamma^{opt}$
5	Use equation (3.15) to compute $\tilde{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)$ (*predictor step*)
6	Use $\tilde{\boldsymbol{\beta}}_{\mathcal{A}}(\gamma)$ as starting point to solve system (3.13) (*corrector step*)
7	For all $a_h^c \in \mathcal{A}^c$ compute $r_{a_h^c}(\gamma)$
8	If $\exists a_h^c \in \mathcal{A}^c$ such that $\lvert r_{a_h^c}(\gamma) \rvert > \gamma$ then
9	$\gamma = \gamma + \varepsilon$, with ε a small positive constant, and go to step 5
10	If $\exists a_h^c \in \mathcal{A}^c$ such that $\lvert r_{a_h^c}(\gamma) \rvert = \lvert r_{a_i}(\gamma) \rvert$, $\forall a_i \in \mathcal{A}$ then update $\mathcal{A}$
11	Until convergence criterion rule is met

Expression (3.17) generalizes the step size that was proposed in Efron *et al.* (2004).

Since the optimal step size is based on a local approximation,we also include an exclusion step for removing incorrectly included variables in the model.When an incorrect variable is included in the model after the corrector step, we have that there is a non-active variable such that the absolute value of the corresponding Rao score test statistic is greater than γ. Checking this is trivial. To overcome this drawback, the 'optimal' step size from the previous step is reduced by using a small positive constant ε and the inclusion step is redone until the correct variable is joined to the model. A possible choice for ε could be a half of $\Delta\gamma^{opt}$. In Table 3.2 we report the pseudocode of the algorithm that was proposed in this section for a model with the intercept.

From an inspection of the algorithm, it is clear that computationally the most expensive steps are solving the system of equations in expression (3.13) and taking the inverse in equation (3.14). These steps have complexity $O(|\mathcal{A}|^3)$ in a naive implementation, but which can be improved to $O(|\mathcal{A}|^{2.376})$ according to the Coppersmith-Winograd algorithm. Furthermore, iteration across the active set variables results in a total computational complexity of $O(np^{2.376} \min\{n,p\})$, where p is the number of variables and n the number of observations. This compares with a complexity of $O(np \min\{n,p\})$ for the original LARS algorithm.

3.2 Gamma and Inverse Gaussian dgLARS

Consider independent random variables $Y_1, Y_2, ..., Y_n$ satisfying the properties of a generalized linear model. We wish to estimate parameters $\boldsymbol{\beta}$ which are related to the Y_i's through $E(Y_i) = \mu_i$ and $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta} = \eta_i$ where $\mathbf{x}_i$ is a vector with elements x_{ij} , $j = 1, ..., p$.

We know that

$$\ell(\boldsymbol{\beta};\mathbf{y}) = \sum_i \ell(\theta_i,\phi;y_i) = \sum_i \left[\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi)\right]$$

then the Score statistic is

$$\begin{aligned}
\partial_j\ell(\boldsymbol{\beta};\mathbf{y}) &= \frac{\partial\ell(\boldsymbol{\beta};\mathbf{y})}{\partial\beta_j} \\
&= \phi^{-1}\sum_{i=1}^{n}\left[\left(\frac{\partial\theta_i}{\partial\beta_j}\right)(y_i-\mu_i)\right] = \phi^{-1}\sum_{i=1}^{n}\left[x_{ij}\left(\frac{\partial\theta_i}{\partial\mu_i}\right)\left(\frac{\partial\mu_i}{\partial\eta_i}\right)(y_i-\mu_i)\right] \\
&= \phi^{-1}\sum_{i=1}^{n}\left[\frac{(y_i-\mu_i)}{V(\mu_i)}x_{ij}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)\right], \qquad (3.19)
\end{aligned}$$

because $\frac{\partial\mu_i}{\partial\theta_i} = \frac{\partial^2 b(\theta_i)}{\partial\theta_i^2} = V(\mu_i)$. Also, we can obtain $\partial_{mn}\ell(\boldsymbol{\beta};\mathbf{y})$:

$$\begin{aligned}
\partial_{mn}\ell(\boldsymbol{\beta};\mathbf{y}) &= \frac{\partial^2\ell(\boldsymbol{\beta};\mathbf{y})}{\partial\beta_m\partial\beta_n} = \frac{\partial\ \partial_n\ell(\boldsymbol{\beta};\mathbf{y})}{\partial\beta_m} \\
&= \phi^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\beta_m}\left[\left(\frac{\partial\theta_i}{\partial\beta_n}\right)(y_i-\mu_i)\right] \\
&= \phi^{-1}\sum_{i=1}^{n}\left[\left(\frac{\partial\theta_i}{\partial\beta_m\partial\beta_n}\right)(y_i-\mu_i) - \left(\frac{\partial\theta_i}{\partial\beta_n}\right)\cdot\left(\frac{\partial\mu_i}{\partial\beta_m}\right)\right] \\
&= \phi^{-1}\sum_{i=1}^{n}\left[\frac{\partial}{\partial\mu_i}\left(\frac{\partial\theta_i}{\partial\beta_n}\right)\cdot\left(\frac{\partial\mu_i}{\partial\eta_i}\cdot x_{im}\right)\cdot(y_i-\mu_i) - \left(\frac{\partial\theta_i}{\partial\mu_i}\cdot\frac{\partial\mu_i}{\partial\eta_i}\cdot x_{in}\right)\cdot\left(\frac{\partial\mu_i}{\partial\eta_i}\cdot x_{im}\right)\right]
\end{aligned}$$

Since $\frac{\partial\theta_i}{\partial\beta_n} = \frac{\partial\theta_i}{\partial\mu_i}\cdot\frac{\partial\mu_i}{\partial\eta_i}\cdot\frac{\partial\eta_i}{\partial\beta_n}$ and $\frac{\partial\eta_i}{\partial\beta_n} = x_{in}$, so we have

$$\frac{\partial}{\partial\mu_i}\left(\frac{\partial\theta_i}{\partial\beta_n}\right) = \left(\frac{\partial^2\theta_i}{\partial\mu_i^2}\cdot\frac{\partial\mu_i}{\partial\eta_i} + \frac{\partial\theta_i}{\partial\mu_i}\cdot\frac{\partial^2\mu_i}{\partial\eta_i^2}\cdot\frac{\partial\eta_i}{\partial\mu_i}\right)\cdot x_{in}$$

Therefore, we have

$$\partial_{mn}\ell(\boldsymbol{\beta};\mathbf{y}) = \phi^{-1}\sum_i x_{im}\ x_{in}\left\{\left(\frac{\partial^2\theta_i}{\partial\mu_i^2}\cdot\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 + \frac{\partial\theta_i}{\partial\mu_i}\cdot\frac{\partial^2\mu_i}{\partial\eta_i^2}\right)(y_i-\mu_i) - \frac{\partial\theta_i}{\partial\mu_i}\cdot\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right\} \qquad (3.20)$$

where $\frac{\partial\mu_i}{\partial\eta_i}$, $\frac{\partial^2\mu_i}{\partial\eta_i^2}$, $\frac{\partial\theta_i}{\partial\mu_i} = \frac{1}{V(\mu_i)}$ and $\frac{\partial^2\theta_i}{\partial\mu_i^2} = -\frac{\partial V(\mu_i)/\partial\mu_i}{V(\mu_i)^2}$ given in Subsections 2.2.1 and 2.2.2 for Gamma and Inverse Gaussian distributions, respectively.

The Fisher Information Matrix has terms

$$\begin{aligned}
\mathcal{I}_{jk}(\boldsymbol{\beta}) &= E[\partial_j\ell(\boldsymbol{\beta};\mathbf{y})\cdot\partial_k\ell(\boldsymbol{\beta};\mathbf{y})] = E[-\partial_{jk}\ell(\boldsymbol{\beta};\mathbf{y})] \\
&= \phi^{-1}\sum_{i=1}^{n}\left[x_{ij}\ x_{ik}\ \frac{\partial\theta_i}{\partial\mu_i}\cdot\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right] \\
&= \phi^{-1}\sum_{i=1}^{n}\left[\frac{x_{ij}\ x_{ik}}{V(\mu_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right], \qquad (3.21)
\end{aligned}$$

because $E[(Y_i - \mu_i)(Y_l - \mu_l)] = 0$ for $i \neq l$ as the Y_i's are independent, and $E[(Y_i - \mu_i)^2] = Var(Y_i) = \phi\ V(\mu_i) = \phi\ \frac{\partial \mu_i}{\partial \theta_i}$. Also, we can obtain $\partial_m \mathcal{I}_n(\boldsymbol{\beta})$ or $\partial_m \mathcal{I}_{nn}(\boldsymbol{\beta})$ as follows:

$$\begin{aligned}
\partial_m \mathcal{I}_n(\boldsymbol{\beta}) &= \frac{\partial\ \mathcal{I}_n(\boldsymbol{\beta})}{\partial \beta_m} \\
&= \phi^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \beta_m} \left[\frac{\partial \theta_i}{\partial \mu_i} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \cdot x_{in}^2 \right] \\
&= \phi^{-1} \sum_{i=1}^{n} x_{in}^2 \left[\frac{\partial}{\partial \beta_m} \left(\frac{\partial \theta_i}{\partial \mu_i} \right) \cdot \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 + 2\ \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial}{\partial \beta_m} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right]
\end{aligned}$$

Since $\frac{\partial}{\partial \beta_m} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) = \frac{\partial^2 \mu_i}{\partial \eta_i^2} \cdot x_{im}$ and

$$\begin{aligned}
\frac{\partial}{\partial \beta_m} \left(\frac{\partial \theta_i}{\partial \mu_i} \right) &= \frac{\partial}{\partial \mu_i} \left(\frac{\partial \theta_i}{\partial \mu_i} \right) \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_m} \\
&= \frac{\partial^2 \theta_i}{\partial \mu_i^2} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{im}\ ,
\end{aligned}$$

we have

$$\partial_m \mathcal{I}_n(\boldsymbol{\beta}) = \phi^{-1} \sum_i x_{im}\ x_{in}^2 \left\{ \frac{\partial^2 \theta_i}{\partial \mu_i^2} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^3 + 2\ \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial^2 \mu_i}{\partial \eta_i^2} \right\}. \tag{3.22}$$

where $\frac{\partial \mu_i}{\partial \eta_i}$, $\frac{\partial^2 \mu_i}{\partial \eta_i^2}$, $\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{V(\mu_i)}$ and $\frac{\partial^2 \theta_i}{\partial \mu_i^2} = -\frac{\partial V(\mu_i)/\partial \mu_i}{V(\mu_i)^2}$ given in Subsections 2.2.1 and 2.2.2 for Gamma and Inverse Gaussian distributions, respectively.

The Rao Score test statistic is

$$r_m(\boldsymbol{\beta}) = \frac{\partial_m \ell(\boldsymbol{\beta}; \mathbf{y})}{\sqrt{\mathcal{I}_m(\boldsymbol{\beta})}} = \phi^{-1/2}\ \frac{\sum_{i=1}^{n} \frac{(y_i - \mu_i)\ x_{im}}{V(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i}}{\left(\sum_{i=1}^{n} \frac{x_{im}^2}{V(\mu_i)} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right)^{1/2}}. \tag{3.23}$$

In the case of the *canonical link*, $\theta_i = \eta_i$, all of the above equations reduce to:

$$\partial_j \ell(\boldsymbol{\beta}; \mathbf{y}) = \phi^{-1} \sum_{i=1}^{n} (y_i - \mu_i)\ x_{ij}, \tag{3.24}$$

$$\partial_{mn} \ell(\boldsymbol{\beta}; \mathbf{y}) = -\ \phi^{-1} \sum_{i=1}^{n}\ x_{im}\ x_{in}\ \frac{\partial \mu_i}{\partial \eta_i}, \tag{3.25}$$

$$\mathcal{I}_{jk}(\boldsymbol{\beta}) = \phi^{-1} \sum_{i=1}^{n}\ x_{ij}\ x_{ik}\ \frac{\partial \mu_i}{\partial \eta_i}, \tag{3.26}$$

$$\partial_m \mathcal{I}_n(\boldsymbol{\beta}) = \phi^{-1} \sum_{i=1}^{n} x_{im}\ x_{in}^2\ \frac{\partial^2 \mu_i}{\partial \eta_i^2}, \tag{3.27}$$

$$r_m(\boldsymbol{\beta}) = \phi^{-1/2}\ \frac{\sum_{i=1}^{n} (y_i - \mu_i)\ x_{im}}{\left(\sum_{i=1}^{n} x_{im}^2\ \frac{\partial \mu_i}{\partial \eta_i} \right)^{1/2}}. \tag{3.28}$$

3.2.1 Gamma dgLARS

Let us assume that the responses $Y_1, Y_2, ..., Y_n$ are Gamma random variables with $G(\nu, \frac{\mu}{\nu})$, so that

$$E(Y_i) = -\frac{1}{\theta_i} = \mu_i,$$
$$Var(Y_i) = \phi\ V(\mu_i) = \frac{\mu_i^2}{\nu},$$

where $\phi^{-1} = \nu$.

Since we are going to consider three of the most commonly used link functions for Gamma distribution, in the following, we obtain whatever we will need with these link functions. From Subsections 2.2.1 and 2.2.2, and equations (3.19)-(3.23), we have:

- The *canonical inverse* link ($\eta_i = -\frac{1}{\mu_i}$):

$$\begin{aligned}
\partial_j \ell(\boldsymbol{\beta}; \mathbf{y}) &= \nu \sum_{i=1}^{n} (y_i - \mu_i)\ x_{ij} \quad , \\
\partial_{mn} \ell(\boldsymbol{\beta}; \mathbf{y}) &= -\nu \sum_{i=1}^{n} x_{im}\ x_{in}\ \mu_i^2 \quad , \\
\mathcal{I}_{jk}(\boldsymbol{\beta}) &= \nu \sum_{i=1}^{n} x_{ij}\ x_{ik}\ \mu_i^2 \quad , \\
\partial_m \mathcal{I}_n(\boldsymbol{\beta}) &= 2\ \nu \sum_{i=1}^{n} x_{im}\ x_{in}^2\ \mu_i^3 \quad , \\
r_m(\boldsymbol{\beta}) &= \sqrt{\nu}\ \frac{\sum_{i=1}^{n} (y_i - \mu_i)\ x_{im}}{(\sum_{i=1}^{n} x_{im}^2\ \mu_i^2)^{1/2}} \quad .
\end{aligned}$$

- The *log* link:

$$\begin{aligned}
\partial_j \ell(\boldsymbol{\beta}; \mathbf{y}) &= \nu \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\mu_i}\ x_{ij}, \\
\partial_{mn} \ell(\boldsymbol{\beta}; \mathbf{y}) &= -\nu \sum_{i=1}^{n} \frac{y_i}{\mu_i}\ x_{im}\ x_{in} \ , \\
\mathcal{I}_{jk}(\boldsymbol{\beta}) &= \nu \sum_{i=1}^{n} x_{ij}\ x_{ik} \ , \\
\partial_m \mathcal{I}_n(\boldsymbol{\beta}) &= 0 \ , \\
r_m(\boldsymbol{\beta}) &= \sqrt{\nu}\ \frac{\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\mu_i}\ x_{im}}{(\sum_{i=1}^{n} x_{im}^2)^{1/2}} \quad .
\end{aligned}$$

- The *identity* link:

$$
\begin{aligned}
\partial_j \ell(\boldsymbol{\beta};\mathbf{y}) &= \nu \sum_{i=1}^{n} \frac{(y_i-\mu_i)}{\mu_i^2}\, x_{ij}\ ,\\
\partial_{mn} \ell(\boldsymbol{\beta};\mathbf{y}) &= -\nu \sum_{i=1}^{n} \left(\frac{2y_i}{\mu_i^3}-\frac{1}{\mu_i^2}\right)\, x_{im}\, x_{in},\\
\mathcal{I}_{jk}(\boldsymbol{\beta}) &= \nu \sum_{i=1}^{n} \frac{x_{ij}\, x_{ik}}{\mu_i^2},\\
\partial_m \mathcal{I}_n(\boldsymbol{\beta}) &= -2\ \nu \sum_{i=1}^{n} \frac{x_{im}\, x_{in}^2}{\mu_i^3},\\
r_m(\boldsymbol{\beta}) &= \sqrt{\nu}\, \frac{\sum_{i=1}^{n} \frac{(y_i-\mu_i)}{\mu_i^2}\, x_{im}}{\left(\sum_{i=1}^{n} \frac{x_{im}^2}{\mu_i^2}\right)^{1/2}}\ .
\end{aligned}
$$

3.2.2 Inverse Gaussian dgLARS

Let us assume that the responses $Y_1, Y_2, ..., Y_n$ are Inverse Gaussian random variables with $IG(\mu_i, \lambda)$, so that

$$
\begin{aligned}
E(Y_i) &= \frac{1}{\sqrt{-2\theta_i}} = \mu_i,\\
Var(Y_i) &= \phi\ V(\mu_i) = \frac{\mu_i^3}{\lambda},
\end{aligned}
$$

where $\phi^{-1} = \lambda$.

Since we are going to consider four of the most commonly used link functions for Inverse Gaussian distribution, in the following, we obtain whatever we will need with one canonical link and three noncanonical links as follows:

- The *canonical inverse-square* link ($\eta_i = -\frac{1}{2\mu_i^2}$):

$$\begin{aligned}
\partial_j \ell(\boldsymbol{\beta};\mathbf{y}) &= \lambda \sum_{i=1}^{n} (y_i - \mu_i)\ x_{ij}\ , \\
\partial_{mn} \ell(\boldsymbol{\beta};\mathbf{y}) &= -\lambda \sum_{i=1}^{n} x_{im}\ x_{in}\ \mu_i^3\ , \\
\mathcal{I}_{jk}(\boldsymbol{\beta}) &= \lambda \sum_{i=1}^{n} x_{ij}\ x_{ik}\ \mu_i^3\ , \\
\partial_m \mathcal{I}_n(\boldsymbol{\beta}) &= 3\ \lambda \sum_{i=1}^{n} x_{im}\ x_{in}^2\ \mu_i^5\ , \\
r_m(\boldsymbol{\beta}) &= \sqrt{\lambda}\ \frac{\sum_{i=1}^{n} (y_i - \mu_i)\ x_{im}}{\left(\sum_{i=1}^{n} x_{im}^2\ \mu_i^3\right)^{1/2}}\ .
\end{aligned}$$

- The *inverse* link ($\eta_i = \frac{1}{\mu_i}$):

$$\begin{aligned}
\partial_j \ell(\boldsymbol{\beta};\mathbf{y}) &= -\lambda \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\mu_i}\ x_{ij}\ , \\
\partial_{mn} \ell(\boldsymbol{\beta};\mathbf{y}) &= -\lambda \sum_{i=1}^{n} x_{im}\ x_{in}\ y_i\ , \\
\mathcal{I}_{jk}(\boldsymbol{\beta}) &= \lambda \sum_{i=1}^{n} x_{ij}\ x_{ik}\ \mu_i\ , \\
\partial_m \mathcal{I}_n(\boldsymbol{\beta}) &= -\lambda \sum_{i=1}^{n} x_{im}\ x_{in}^2\ \mu_i^2\ , \\
r_m(\boldsymbol{\beta}) &= -\sqrt{\lambda}\ \frac{\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\mu_i}\ x_{im}}{\left(\sum_{i=1}^{n} x_{im}^2\ \mu_i\right)^{1/2}}\ .
\end{aligned}$$

- The *log* link:

$$\partial_j \ell(\boldsymbol{\beta};\mathbf{y}) = \lambda \sum_{i=1}^{n} \frac{(y_i-\mu_i)}{\mu_i^2}\ x_{ij}\ ,$$

$$\partial_{mn} \ell(\boldsymbol{\beta};\mathbf{y}) = -\lambda \sum_{i=1}^{n} \left(\frac{2y_i}{\mu_i^2} - \frac{1}{\mu_i}\right) x_{im}\ x_{in}\ ,$$

$$\mathcal{I}_{jk}(\boldsymbol{\beta}) = \lambda \sum_{i=1}^{n} \frac{x_{ij}\ x_{ik}}{\mu_i}\ ,$$

$$\partial_m \mathcal{I}_n(\boldsymbol{\beta}) = -\lambda \sum_{i=1}^{n} \frac{x_{im}\ x_{in}^2}{\mu_i}\ ,$$

$$r_m(\boldsymbol{\beta}) = \sqrt{\lambda}\ \frac{\sum_{i=1}^{n} \frac{(y_i-\mu_i)}{\mu_i^2}\ x_{im}}{\left(\sum_{i=1}^{n} \frac{x_{im}^2}{\mu_i}\right)^{1/2}}.$$

- The *identity* link:

$$\partial_j \ell(\boldsymbol{\beta};\mathbf{y}) = \lambda \sum_{i=1}^{n} \frac{(y_i-\mu_i)}{\mu_i^3}\ x_{ij}\ ,$$

$$\partial_{mn} \ell(\boldsymbol{\beta};\mathbf{y}) = -\lambda \sum_{i=1}^{n} \left(\frac{3y_i}{\mu_i^4} - \frac{2}{\mu_i^3}\right) x_{im}\ x_{in}\ ,$$

$$\mathcal{I}_{jk}(\boldsymbol{\beta}) = \lambda \sum_{i=1}^{n} \frac{x_{ij}\ x_{ik}}{\mu_i^3}\ ,$$

$$\partial_m \mathcal{I}_n(\boldsymbol{\beta}) = -3\ \lambda \sum_{i=1}^{n} \frac{x_{im}\ x_{in}^2}{\mu_i^4}\ ,$$

$$r_m(\boldsymbol{\beta}) = \sqrt{\lambda}\ \frac{\sum_{i=1}^{n} \frac{(y_i-\mu_i)}{\mu_i^3}\ x_{im}}{\left(\sum_{i=1}^{n} \frac{x_{im}^2}{\mu_i^3}\right)^{1/2}}\ .$$

Chapter 4

Simulation Study and Data Analysis

Our simulation study is based on a Gamma regression model with sample size $n = (10, 100, 1000)$ and three different values of p, namely $p = (10, 100, 1000)$. The large values of p are useful to study the behaviour of the methods in a high dimensional setting. The study is based on four different configurations of the covariance structure of the p predictors, such that $X_1, X_2, \cdots, X_p$ sampled from an $N(\mathbf{0}, \Sigma)$ distribution, where the diagonal element of Σ is 1 and the off-diagonal element is ρ (in our study we use $\rho = (-0.4, 0, 0.4, 0.8)$).

Only the first five predictors are used to simulate the binary response variable. We simulate 50 data sets and choose

$$\boldsymbol{\beta} = (1, \underbrace{2, 3, 4, 5}_{4}, \underbrace{0, \cdots, 0}_{p-4}).$$

Chapter 5

Conclusions

In this paper extend the dgLARS method for a GLM to a larger class of the exponential family, namely the *exponential dispersion family* (when the dispersion parameter, ϕ, is known and unknown). We present the **dglars.G.IG** package that implements both the algorithms to compute the solution curve implicitly defined by dgLARS based on Gamma and Inverse Gaussian models. The object returned by these functions is a S3 class object, for which specific methods and functions have been implemented.

Bibliography

[1] Akaike H., *Information Theory as an Extension of the Maximum Likelihood Principle.* In BN Petrov, F Czaki (eds.), Second International Symposium on Information Theory, pp. 267-281 (1973). Akademiai Kiado, Budapest.

[2] Augugliaro L., Mineo A. M. and Wit E. C., *Differential Geometric Least Angle Regression: A Differential Geometric Approach to Sparse Generalized Linear Models.* J. R. Statist. Soc. B, 75(3), 471-498 (2013).

[3] Augugliaro L., Mineo A. M. and Wit E. C., *dglars: An R Package to Estimate Sparse Generalized Linear Models.* J. Statist. Software , 59(8), 1-40 (2014).

[4] Efron B., Hastie T., Johnstone I. and Tibshirani R., *Least Angle Regression.* Ann. Statist. 32(2), 407-499 (2004).

[5] Fan J. and Li R., *Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties.* J. Am. Statist. Ass. 96(456), 1348-1360 (2001).

[6] R Development Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria (2012). ISBN 3900051-07-0, http://www.R-project.org/.

[7] Schwarz G., *Estimating the Dimension of a Model.* Ann. Statist. 6(2), 461-464 (1978).

[8] Tibshirani R., *Regression Shrinkage and Selection Via the Lasso.* J. R. Statist. Soc. B 58(1), 267-288 (1996).

[9] Zhang C.H., *Nearly Unbiased Variable Selection Under Minimax Concave Penalty.* Ann. Statist. 38(2), 894-942 (2010).