Electronic Journal of Statistics
Vol. 10 (2016) 1630–1676
ISSN: 1935-7524
DOI: 10.1214/16-EJS1152

Joint estimation and variable selection for mean and dispersion in proper dispersion models

Anestis Antoniadis

Laboratoire Jean Kuntzmann
Université de Grenoble
Grenoble, F-38041, France

Department of Statistical Sciences
University of Cape Town
South Africa
e-mail: Anestis.Antoniadis@imag.fr

Irène Gijbels

Department of Mathematics and Leuven Statistics Research Centre (LStat)
KU Leuven
Celestijnenlaan 200B, Box 2400, B-3001 Leuven (Heverlee), Belgium
e-mail: Irene.Gijbels@wis.kuleuven.be

Sophie Lambert-Lacroix

Laboratoire TIMC-IMAG, UMR 5525
Université de Grenoble
Grenoble, F-38041, France
e-mail: Sophie.Lambert@imag.fr

and

Jean-Michel Poggi

Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS
Université Paris-Saclay, 91405 Orsay, France

Université Paris Descartes, Paris, France
e-mail: Jean-Michel.Poggi@math.u-psud.fr

Abstract: When describing adequately complex data structures one is often confronted with the fact that mean as well as variance (or more generally dispersion) is highly influenced by some covariates. Drawbacks of the available methods is that they are often based on approximations and hence a theoretical study should deal with also studying these approximations. This however is often ignored, making the statistical inference incomplete. In the proposed framework of double generalized modelling based on proper dispersion models we avoid this drawback and as such are in a good position to use recent results on Bregman divergence for establishing theoretical results for the proposed estimators in fairly general settings. We also study variable selection when there is a large number of covariates, with this number possibly tending to infinity with the sample size. The proposed estimation and selection procedure is investigated

via a simulation study, that includes also a comparative study with competitors. The use of the methods is illustrated via some real data applications.

MSC 2010 subject classifications: Primary 62Gxx, 62Hxx; secondary 62Jxx.
Keywords and phrases: Bregman divergence, Fisher-orthogonality, penalization, proper dispersion models, variable selection, SCAD.

Received March 2016.

Contents

1. Introduction

1.1. From generalized linear models to quasi-likelihood and pseudo-likelihood

Within the framework of univariate regression modelling techniques, the generalized linear model (GLM) holds a prominent place ([58]). Generalized linear models allow to model responses which are not normally distributed, using methods closely analogous to linear methods for normal data ([53]). They assume an exponential family distribution for the response variable and are more general than normal linear methods in that a mean-variance relationship appropriate for the data can be accommodated and an appropriate scale can be chosen for modelling the mean on which the action of the covariates is approximately linear. Once the mean-variance relationship is specified, the variance is assumed known up to a constant of proportionality, the dispersion parameter.

For many years, the focus has been on modelling and estimating the mean structure of the data while treating the dispersion parameter as a constant since GLMs automatically allow for dependence of the variance on the mean through the distributional assumptions. While there is often strong motivation for using exponential family distributions on the grounds of interpretability, the observed data may however exhibit greater variability than the one which is implied by the mean-variance relationship. Then the loss of efficiency in estimating the mean parameters, using constant dispersion models when the dispersion is varying, may be substantial. This has trigged the development of more flexible models than GLMs, by additional modelling of the dispersion as a function of covariates. When only the mean varies with the covariates, the conditional distribution function, of the response given the covariates, is considered to belong to the exponential family. The usual likelihood theory then allows for showing the basic properties of statistical inference. When the dispersion function however is also varying with the covariates, this road can no longer be taken.

Quasi-likelihood ([72]) provides one simple approach to do statistical inference in the presence of over-dispersion, where the exponential family assumption is dropped and only a model for the mean is given with the response variance a function of the mean up to a multiplicative constant. However, this approach does not allow over-dispersion to be modelled as a function of covariates. An extension of quasi-likelihood which allows this is the extended quasi-likelihood (EQL) of [57], but in general extended quasi-likelihood estimators may not be consistent ([22]). Alternatively, generalized least squares estimating equations

for the mean parameters can be combined with normal score estimating equations for the variance parameters. Such a procedure is referred to as pseudo-likelihood ([21]). [64] considers modelling the mean and variance in a parametric class of models which allows normal, inverse Gaussian and gamma response distributions, and a quasi-likelihood extension is also proposed which uses a similar approach to pseudo-likelihood for estimation of variance parameters. [65, 66] consider extensions of residual maximum likelihood (REML) estimation of variance parameters to double generalized linear models where dispersion parameters are modelled linearly in terms of covariates after transformation by a link function.

An approach related to EQL is the double exponential family of distributions introduced in [26]. Starting from such a double exponential family of distributions [33] and [19] allow the dispersion parameter to vary with the covariates, and consider (robust for the second reference) statistical inference for mean and dispersion functions. Their modelling which is essentially nonparametric in nature, allows for a flexible unknown dispersion pattern (over-dispersion, under-dispersion, ...). This is in contrast with papers that focus (for example) specifically on overdispersion. See [16] and [82] among others. Other recent work includes hierarchical modelling of mean and variance. See [13, 14]. We will shortly refer to such approaches as model-based approaches.

Estimating equation approaches (pseudo-likelihood approaches) and model-based approaches (EQL, double-exponential families, hierarchical modelling...) have both serious drawbacks when it comes to establishing asymptotic properties of proposed estimators in a fairly general setting.

Inference about mean and variance functions using estimating equations has the drawback that there is no fully specified model, making it difficult to deal with characteristics of the predictive distribution for a future response, other than its mean and variance. Model-based approaches for modelling over-dispersion include exponential dispersion models and related approaches ([40], [64]). Moreover, if an exponential dispersion family with a given variance function exists, then the EQL is the log-likelihood function based on a saddle point approximation to that family ([53], [42]). [47] notice that the unnormalized EQL and Efron's [26] unnormalized double-exponential family are equivalent up to some constant terms and therefore both approaches lead to identical inferences. Hence both approaches share the drawback mentioned above.

In contrast, the framework that we will adopt in this work for flexibly modelling of mean and variance/dispersion functions is based on the proper dispersion regression models introduced by [40]. Proper dispersion regression models do not suffer the drawbacks of extended quasi-likelihood which occur because the extended quasi-likelihood is not a proper log-likelihood function. They provide a convenient framework for modelling as they retain the parsimony and interpretability of GLMs, while allowing, if necessary, the flexible dependence of the link transformed mean and variance parameters on predictors (covariates). This framework allows us to get to our first goal: properly establishing statistical inference properties in general flexible regression settings. Such a theoretical underpinning has been lacking largely in the literature so far.

1.2. Variable selection in joint mean and dispersion modelling

A second goal is to address the problem of variable selection, i.e. the selection of a reasonably small subset of informative covariates that are associated with a response in a particular dispersion model. The selection of informative covariates plays a key role in many practical applications and is often required in applications with high-dimensional data, i.e. data sets with a potentially large number of covariates, and remains an important and challenging issue in statistics. For GLMs there is a variety of variable selection techniques available in the literature. The more traditional methods such as Akaike's information criterion (AIC), Mallows C_p and the Bayesian information criterion (BIC) are practically useful and easy to implement as they use a fixed penalty on the size of a model. They are generally adequate for a small number of covariates.

In the context of joint modelling of the mean and the dispersion in GLMs variable selection was dealt with recently in [71]. Nevertheless, these methods follow stepwise and subset selection procedures, making them computationally intensive and often unstable ([11]), when the number of variables is large. For high-dimensional problems with a number of variables much larger that the number of observations, the use of penalization, or regularization, has become a common approach, where an increasingly frequent goal is to simultaneously select important variables and estimate their effects. It has led to the development of several penalized variable selection procedures for GLMs (parametric and nonparametric). Quadratic penalization on the regression coefficients as the one used by [52] is more stable but as it only shrinks coefficients and doesn't reduce them to 0, it fails to produce parsimonious models. See [4] for a study on penalization methods in extended generalized linear models. [48] propose an ℓ_0 like penalized likelihood approach for variable selection and estimation in generalized linear models. [59] discuss a natural GLM extension of the Lasso (GLasso), and [60] propose a Forward-lasso with adaptive shrinkage for GLMs. [38] relies on a scad penalization for variable selection in generalized linear models.

However, in all papers above, the dispersion parameter is treated as constant (i.e. not varying with covariates). With dispersion models, problems related to variable selection become even more serious, since not only the location parameters but also the dispersion parameters of the response distribution are associated with a set of predictor variables. Both joint mean-dispersion modelling and variable selection are therefore areas of great interest. The contribution of the current paper can be summarized as follows:

(i) We provide a unifying framework for estimation of and variable selection in jointly the mean and dispersion in a context of generalized proper dispersion models.
(ii) Subsequently we provide a solid theoretical framework for establishing estimation and variable selection consistency.

In recent years there have been many papers dealing with joint estimation of mean and variance, but a good theoretical treatment was missing. Not only

we fill this gap, but we also deal with variable selection. The main advantage of our approach is that we provide a solid theoretical framework that allows to establish the necessary statistical properties of the estimation and selection procedure. In addition the unifying framework deals with many special cases in one single track.

The paper is further organized as follows. In Section 2 we briefly describe the necessary elements about dispersion models, including how joint modelling of mean and dispersion is dealt with. This section also highlights the main differences with existing approaches. Section 3 then easily passes on to the setting when covariates are observed. Apart from estimating the unknown mean and dispersion we are also concerned with selecting at the same time the (few) significant covariates in a large number of observed covariates. Within the proposed framework we can nicely rely on the notion of Bregman divergence, to establish statistical estimation and variable selection properties, in Section 4. Issues of practical implementations to solve the optimization problems involved, in particular algorithms for doing so, are discussed in Section 5. The finite-sample performance of the proposed estimation and variable selection procedure is investigated in a simulation study in Section 6. We illustrate the use of the methods on some real data applications in Section 7. Finally, in Section 8 we discuss further extensions and research perspectives.

2. Proper dispersion models, and joint modelling of mean and dispersion

In this section we introduce the framework of proper dispersion models mostly in a general non-regression setting (no covariates included), with exception of Section 2.2 which is about the specific case of generalized linear models in a univariate regression setting. In Section 3 we then consider fully the regression setting and the framework of generalized proper dispersion models.

2.1. Brief review on exponential dispersion models

Following [40] a regular reproductive exponential dispersion model, denoted by $\mathrm{ED}(\mu, \phi)$ is defined via the density function

$$f(y;\theta,\phi) = a(y,\phi)\exp\left\{\frac{1}{\phi}\left(y\theta - \kappa(\theta)\right)\right\}, \tag{2.1}$$

where $a(y;\phi)$ and $\kappa(\theta)$ are known functions and θ and ϕ are the parameters of the model. This model generalizes the exponential family model in which θ is the canonical parameter and ϕ is known. The function $\kappa(\cdot)$ is called the cumulant function, and in a regular model this function is twice continuously differentiable.

We say that the model satisfies the Bartlett identities of order 1 and 2, if, for $Y \sim \mathrm{ED}(\mu, \phi)$ with probability density $f(\cdot; \theta, \phi)$, we have respectively

$$\mathbb{E}\left[\frac{\partial}{\partial\theta} \ln f(Y; \theta, \phi)\right] = 0 \tag{2.2}$$

and

$$\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \ln f(Y; \theta, \phi)\right] = -\mathbb{E}\left[\frac{\partial}{\partial\theta} \ln f(Y; \theta, \phi)\right]^2. \tag{2.3}$$

If the model satisfies the Bartlett identity of order 1, then

$$\mu \equiv \mathbb{E}(Y) = \kappa'(\theta), \tag{2.4}$$

where for simplicity of presentation we assume $\theta \in I\!R$, and where $\kappa'(\theta)$ denotes the derivative of $\kappa(\theta)$. If in addition the Bartlett identity of order 2 holds, then it follows that

$$\mathrm{Var}(Y) = \phi\, \kappa''\left((\kappa')^{-1}(\mu)\right).$$

A special case of regular reproductive exponential dispersion models are the family of *Tweedie models* which are exponential dispersion models for which the variance of Y is proportional to

$$V(\mu) = \mu^p \qquad \text{with } p \notin (0, 1). \tag{2.5}$$

Emphasizing the dependence on p we denote a Tweedie model for a given p, as $Y \sim \mathrm{ED}_p(\mu, \phi)$. It is important to note that Tweedie distributions are closed under scale transformations, i.e. if $Y \sim \mathrm{ED}_p(\mu, \phi)$, then $cY \sim \mathrm{ED}_p(c\mu, c^{2-p}\phi)$ for any $c > 0$. The most important examples of elements of the Tweedie family are normal, Gamma and inverse Gaussian distributions, that correspond respectively with the cases $p = 0$, $p = 2$ and $p = 3$. Other interesting subclasses of distributions in the class of Tweedie family are: (i) the Poisson distributions (with $p = 1$); (ii) the compound Poisson distributions (with $1 < p < 2$) for which the domain is $[0, \infty[$ having a nonzero probability mass at the point zero accompanied with a skewed continuous distribution on $]0, \infty[$; (iii) the stable Tweedie distributions with $p \leq 0$ or $p > 2$, of which the inverse Gaussian is an example (case $p = 3$). In this paper we restrict to the case of continuous distributions, but readers are referred to Section 8.3 for extensions to discrete distributions or mixtures of continuous and discrete distributions (such as the subclasses (i) and (ii)).

Exponential dispersion models are on their turn special cases of dispersion models, which we briefly review in Section 2.3.

2.2. Generalized linear models and double generalized linear models

In a regression setting the aim is to explain the impact of a covariate, say X, on the response vector Y. For simplicity of presentation we restrict, in this section, to the simplest case of a univariate covariate. See Section 3 for a more general multivariate setting.

In generalized linear models, the conditional distribution of Y given $X = x$, with $x \in \mathbb{R}$, is assumed to be of the form (2.1) but with now θ depending on x (i.e. a function of x). More precisely, the conditional density function of Y given $X = x$ is of the form

$$f_{Y|X}(y; \theta(x), \phi) = a(y, \phi) \exp\left\{ \frac{1}{\phi} \left(y\theta(x) - \kappa(\theta(x)) \right) \right\}, \tag{2.6}$$

leading to

$$\mu(x) \equiv \mathbb{E}(Y|X = x) = \kappa'(\theta(x)) \quad \text{and} \quad \mathrm{Var}(Y|X = x) = \phi\, \kappa''\left((\kappa')^{-1}(\mu(x)) \right). \tag{2.7}$$

In a generalized linear model it is further assumed that the influence of X on the mean of Y is linear after having applied a link function g, i.e.

$$g(\mu(x)) = \theta(x) = \theta_0 + \theta_1 x, \tag{2.8}$$

and estimation of the mean function $\mu(x)$ resorts in estimation of the parameters θ_0 and θ_1. Obviously, in an exponential model with canonical parameter $g^{-1} = \kappa$ (see (2.4) and (2.7)).

So far, the parameter ϕ in (2.1) and (2.6) is assumed to be a nuisance parameter, but one that describes correctly the scaling properties of the distribution.

[26] introduced an extra parameter ζ into the exponential model (2.1) to allow for more modelling flexibility in the variance (dispersion). He considered the *double exponential family* probability density given by

$$f^{\mathrm{DE}}(y; \theta, \phi, \zeta) = C(\theta, \zeta) \zeta^{-1/2} \left\{ f(y; \theta, \phi) \right\}^{1/\zeta} \left\{ f(y; (\kappa')^{-1}(y), \phi) \right\}^{1-1/\zeta}, \tag{2.9}$$

where $C(\theta, \zeta)$ is a normalizing constant, and $\zeta > 0$. It is typical in double exponential models to approximate this constant by 1 (see [26]). See also Section 2.3.

In a context of double generalized linear models, the essential idea is to allow the scale parameter ϕ to be influenced by the covariate x, i.e. ϕ becomes a function of x. Again, via the interference of a link function h, this effect is assumed to be linear, i.e.

$$h(\phi(x)) = \gamma(x) = \gamma_0 + \gamma_1 x. \tag{2.10}$$

For the rest of this section we will present all material in a simplified (non-regression) setting, but when writing, for example $g(\mu) = \theta$ and $h(\phi) = \gamma$ one should keep the dependence on covariates in mind.

2.3. Proper dispersion models

Dispersion models were originally introduced by [40] and further studied in [41], the book [42] and [43]. We first introduce some notations and definitions. Let

$\Omega \subseteq \mathbb{R}$ be an open interval. A function $d : \Omega \times \Omega \to \mathbb{R}$ is called a *unit deviance* if

$$d(y, y) = d(\mu, \mu) = 0, \quad \forall y,\ \mu \in \Omega, \qquad d(y, \mu) > 0, \quad \forall y \neq \mu.$$

A unit deviance d is called *regular* if $d(y, \mu)$ is twice continuously differentiable with respect to (y, μ) on $\Omega \times \Omega$ and satisfies

$$\left.\frac{\partial^2 d(y,\mu)}{\partial \mu^2}\right|_{y=\mu} > 0, \quad \forall \mu \in \Omega.$$

Then the *unit variance function* $V : \Omega \to \mathbb{R}_+$ is defined by

$$V(\mu) = \frac{2}{\left.\frac{\partial^2 d(y,\mu)}{\partial \mu^2}\right|_{y=\mu}}. \tag{2.11}$$

A *dispersion model* is defined via the probability density function

$$f(y; \theta, \phi) = b(y, \phi) \exp\left\{-\frac{1}{2\phi} d(y, \mu)\right\}, \tag{2.12}$$

where the position parameter μ belongs to Ω (and μ depends on θ), and $\phi > 0$ is the dispersion parameter. If Y has density function $f(\cdot; \theta, \phi)$ in (2.12) we denote $Y \sim \mathrm{DM}(\mu, \phi)$.

One assumes that $b(y, \phi)$ is twice continuously differentiable, with respect to y and $d(y, \mu)$ is twice continuously differentiable with respect to (y, μ). If the dispersion model satisfies the Bartlett identities of order 1 and 2 (see (2.2) and (2.3)), then this implies that $d(\cdot, \cdot)$ should satisfy

$$\mathbb{E}\left[\frac{\partial}{\partial \mu} d(Y, \mu)\right] = 0 \tag{2.13}$$

and

$$\mathbb{E}\left[\frac{\partial^2}{\partial \mu^2} d(Y, \mu)\right] = \frac{1}{2\phi} \mathbb{E}\left[\frac{\partial}{\partial \mu} d(Y, \mu)\right]^2. \tag{2.14}$$

The identities (2.13) and (2.14) in fact hold for any dispersion model for which the above differentiability assumptions are verified, as stated in Proposition 3.1 of [67].

When the factor $b(y, \phi)$ in (2.12) factorizes as $c(\phi)a(y)$ then the corresponding dispersion model is called a *proper dispersion model*. It is crucial in this that the factor $c(\phi)$ does not depend on μ.

If in addition the factor $a(y)$ equals the inverse of the square root of the unit variance function defined in (2.11), i.e. $a(y) = \{V(y)\}^{-1/2}$ then the model defined via the resulting probability density function, namely

$$f(y; \mu, \phi) = c(\phi) \{V(y)\}^{-1/2} \exp\left\{-\frac{1}{2\phi} d(y, \mu)\right\}, \quad y \in \Omega, \tag{2.15}$$

TABLE 1
$PD(\mu, \sigma^2)$ model examples. $I_0(\cdot)$ denotes the modified Bessel function, i.e. $I_0(z) = \sum_{k=0}^{\infty} \frac{(\frac{1}{4}z^2)^k}{(k!)^2}$.

model	Ω	$c(\phi)$ with $\phi = \sigma^2$	$V(y)$	$d(y, \mu)$
Normal $\mathcal{N}(\mu, \sigma^2)$	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi\phi}}$	1	$(y-\mu)^2$
Gamma $\Gamma(\phi, \theta)$, $\theta = \mu\phi$	$\mathbb{R}_+$	$\frac{\frac{1}{\phi}^{\frac{1}{\phi}} e^{-\frac{1}{\phi}}}{\Gamma\left(\frac{1}{\phi}\right)}$	y^2	$2\left\{\frac{y}{\mu} - \ln\left(\frac{y}{\mu}\right) - 1\right\}$
Inverse normal $\mathcal{IN}(\mu, \sigma^2)$	$\mathbb{R}_+$	$\frac{1}{\sqrt{2\pi\phi}}$	y^3	$\frac{(y-\mu)^2}{\mu^2 y}$
Lognormal $\mathcal{LN}(\theta, \sigma^2)$, $\theta = \ln(\mu)$	$\mathbb{R}_+$	$\frac{1}{\sqrt{2\pi\phi}}$	y^2	$(\ln(y) - \ln(\mu))^2$
von Mises $vM(\mu, \sigma^2)$	$[0, 2\pi)$	$\frac{e^{\frac{1}{\phi}}}{2\pi I_0\left(\frac{1}{\phi}\right)}$	1	$2(1 - \cos(y-\mu))$
simplex $S^-(\mu, \sigma^2)$	$(0, 1)$	$\frac{1}{\sqrt{2\pi\phi}}$	$y^3(1-y)^3$	$\frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}$

is called a *regular proper dispersion model*, denoted by $\mathrm{PD}(\mu, \sigma^2)$, where $\phi = \sigma^2$, d is a regular unit deviance and

$$\frac{1}{c(\phi)} = \int_{\Omega} \{V(y)\}^{-1/2} \exp\left\{-\frac{1}{2\phi} d(y, \mu)\right\} dy,$$

to ensure to have a proper density function in (2.15). Several usual distributions belong to the $\mathrm{PD}(\mu, \sigma^2)$ models, namely the normal, Gamma, inverse normal, log-normal, von Mises and simplex distributions. See Table 1 for the details, as well as [41], [7] and [67], among others.

For a *regular proper exponential dispersion model* (see (2.1)), for which

$$d(y, \mu) = -2(y\theta - \kappa(\theta)), \tag{2.16}$$

the quantity $b(y, \phi)$ in (2.12) is of the form (see [42])

$$b(y, \phi) = \{2\pi\phi V(y)\}^{-1/2} \rho(\phi). \tag{2.17}$$

For the normal and the inverse Gaussian families $\rho(\phi) = 1$, while for the gamma families

$$\rho(\phi) = (2\pi)^{1/2} \frac{\alpha^\alpha \exp(-\alpha)}{\Gamma(\alpha)},$$

with $\alpha = 1/\phi$.

The range of applicability of (2.17) can be extended by the saddle point approximation to exponential dispersion models with dispersion parameter ϕ close to zero (i.e. the case of small scale asymptotics), by using that (see e.g. expression (3.44) on page 103 in [42] and Proposition 2.3 in [67])

$$b(y,\phi) = \{2\pi\phi V(y)\}^{-1/2}\{1+O(\phi)\}. \tag{2.18}$$

Recall that for the Tweedie distributions one has: if $Y \sim \mathrm{ED}_p(\mu,\phi)$, then $cY \sim \mathrm{ED}_p(c\mu, c^{2-p}\phi)$ for any $c > 0$. Consequently, for $p < 2$ and $p \notin (0,1)$, by multiplying $Y \sim \mathrm{ED}_p(\mu,\phi)$ with a small number c the dispersion parameter of the resulting Tweedie distribution will tend to zero (since c^{2-p} tends to zero as c tends to zero). Analogously, for Tweedie distributions with $p > 2$, a mulplication of Y with a large c results into a Tweedie distribution with dispersion parameter closer to zero (since c^{2-p} tends to zero as c tends to infinity).

For the Gamma distribution, the saddle point approximation consists in replacing $\Gamma(1/\phi))$ with its Stirling's formula approximation so that

$$\frac{\alpha^\alpha \exp(-\alpha)}{\Gamma(\alpha)} = \frac{1}{\sqrt{2\pi\,\phi}}\{1+O(\phi)\}.$$

A double exponential model (see (2.9)) involves essentially the approximation

$$b(y,\phi) \approx \{2\pi\phi V(y)\}^{-1/2}, \tag{2.19}$$

which then results in not having a proper distribution function to work with, leading to the drawback for statistical inference, mentioned in the introduction.

2.4. *Proper dispersion models and joint modelling of mean and dispersion*

If one is interested in joint modelling of mean and dispersion in the context of a $\mathrm{PD}(\mu,\sigma^2)$ model, then we consider two monotonic differentiable link functions $g(\cdot)$ and $h(\cdot)$ say, where g links the parameter θ with the mean (μ) and where h links the parameter γ, say, with the dispersion (ϕ), i.e.

$$g(\mu) = \theta \qquad \text{and} \qquad h(\phi) = \gamma.$$

The interest is then in joint estimation of θ and γ.

Suppose now that $Y \sim \mathrm{PD}(\mu,\sigma^2)$ where μ depends on θ and ϕ $(= \sigma^2)$ on γ. The logarithm of the probability density function (see (2.15)) is

$$\ln f(y;\mu,\phi) = \ln c(\phi) - \frac{1}{2}\ln V(y) - \frac{1}{2\phi}d(y,\mu). \tag{2.20}$$

Following [64] in his ideas on double generalized linear models, we look upon this as a double dispersion model, in which one considers a first submodel that corresponds to θ (the parameter associated to the mean) for γ fixed; and a

second submodel that corresponds to γ (the parameter associated to the dispersion ϕ) for θ fixed. For the first submodel, the mean submodel, the response variable is Y, and for the second submodel, called the dispersion submodel, the response variable is $D = d(Y, \mu)$ (known as soon as μ is known). For a PD model with $c(\phi) = 1/\sqrt{2\pi\phi}$, it follows that if μ is fixed then the considered form for $\ln f(Y; \mu, \phi)$ coincides with that of a Gamma distribution $\Gamma(2, 2\phi)$ (with the parametrization as in Table 1) for $D = d(Y, \mu)$.

Given $Y_1, \ldots, Y_n$, a sample of independent and identically distributed (i.i.d.) observations from Y, one can, based on (2.20) form the log-likelihood function, which is given by

$$\sum_{i=1}^{n} \left\{ \ln c(\phi) - \frac{1}{2} \ln V(Y_i) - \frac{1}{2\phi} d(Y_i, \mu) \right\}. \tag{2.21}$$

For estimation of θ one then considers (2.21) with γ (and thus ϕ) fixed (given). For estimation of γ, (2.21) is considered with θ (and thus μ) fixed, which results in looking at $D_i = d(Y_i, \mu)$ as the response variables.

Estimation of θ and γ is then done in two separate steps: maximization of (2.21) with respect to μ, with γ fixed, to estimate μ (and hence θ); and maximization of (2.21) with respect to ϕ, with μ fixed, to estimate ϕ (and hence γ). This leads to the maximum likelihood estimators of (μ, ϕ) under this framework of double dispersion models. Of crucial importance however is that it is justified that the two maximization problems can be carried out independent of each other. This is indeed the case in our modelling framework since from the Bartlett identities or order 1 and 2, as stated in (2.13) and (2.14) it follows that

$$\mathbb{E}\left[\frac{\partial^2}{\partial\mu\partial\phi} \ln f(Y; \mu, \phi) \right] = 0, \tag{2.22}$$

which leads to block diagonality of the Fisher information matrix for the joint maximum likelihood estimation of (θ, γ) (or equivalently (μ, ϕ)). The parameters μ and ϕ are thus orthogonal in the sense of [17, 18]. See also [44]. It is this property that allows us to have an iterative algorithm combining the mean and dispersion submodel iteration.

2.5. Quasi-likelihood and extended quasi-likelihood

Let us remark that, contrary to the framework of generalized linear models, the maximum likelihood estimator for PD models does not necessary coincide with the quasi-likelihood maximum estimator.

Let us first briefly review the basics of quasi-likelihood. Suppose that for a response variable Y we know

$$\mathbb{E}(Y) = \mu \qquad \text{and} \qquad \mathrm{Var}(Y) = \sigma^2 V(\mu),$$

where $V(\cdot)$ is a know function of μ. With this knowledge on the first and second (central) moments of Y, it is to be noted that the quantity

$$u(Y;\mu) \equiv \frac{Y-\mu}{\sigma^2 V(\mu)} \tag{2.23}$$

is such that

$$\mathbb{E}\left(u(Y;\mu)\right) = 0$$
$$\mathbb{E}(u^2(Y;\mu)) = \mathrm{Var}\left(u(Y;\mu)\right) = \frac{1}{\sigma^2 V(\mu)} = -\mathbb{E}\left[\frac{\partial u(Y;\mu)}{\partial \mu}\right]$$

which constitutes thus similar properties as the log-likelihood score $\frac{\partial}{\partial\theta}\ln f(Y;\theta,\phi)$. See for example (2.2) and (2.3).

The quasi-likelihood is then obtained by integrating this analogue of the score function:

$$Q(\mu;y) = \int_y^\mu \frac{y-s}{\sigma^2 V(s)} ds,$$

and its corresponding quasi-likelihood score is given in (2.23). Note that, for a proper dispersion model, the associated score quantity is given by (see (2.20))

$$\frac{\partial}{\partial\mu}\ln f(y;\mu,\phi) = -\frac{1}{2\phi}\frac{\partial}{\partial\mu} d(y,\mu), \tag{2.24}$$

which does not necessarily lead to the same estimator as based on the quasi-likelihood score in (2.23). However, in our modelling framework where the Bartlett identities of order 1 and 2 are satisfied (see (2.13) and (2.14)), the dispersion model score function (2.24) replaces the quasi-score (2.23) appearing in the conventional quasi-likelihood estimating equation, and leads to the generalized quasi-score function of [50] (see equation (3) therein). The generalized quasi-likelihood estimator obtained by using such a generalized quasi-score function is consistent. For the Normal, Gamma, and Inverse normal model, quasi-likelihood and generalized quasi-likelihood lead to the same estimator, as is easily seen from Table 2. However for the other distributions this is not the case.

When a dispersion parameter is involved, [57] suggest, in the context of an exponential dispersion model, to use (see (2.18))

$$Q^+(\mu,\phi;y) = -\frac{1}{2}\ln\{2\pi\phi(\gamma)V(y)\} - \frac{1}{2}\frac{d(y,\mu(\theta))}{\phi(\gamma)}, \tag{2.25}$$

with the dependence of μ on θ and of ϕ on γ emphasized, and where

$$d(y,\mu) = -2\int_y^\mu \frac{y-s}{\sigma^2 V(s)} ds,$$

is the deviance function. The quantity $Q^+(\mu,\phi;y)$ is called the *extended quasi-likelihood*.

TABLE 2

Mean and variance for $PD(\mu, \sigma^2)$ model examples with partial derivative of d. $I_0(\cdot)$ and $I_1(\cdot)$ denote modified Bessel functions: $I_0(z) = \sum_{k=0}^{\infty} \frac{(\frac{1}{4}z^2)^k}{(k!)^2}$ and $I_1(z) = \frac{1}{2}z \sum_{k=0}^{\infty} \frac{(\frac{1}{4}z^2)^k}{k!\,\Gamma(k+2)}$.

model	$\frac{\partial d(y,\mu)}{\partial \mu}$	mean	variance
Normal $\mathcal{N}(\mu, \sigma^2)$	$-2(y-\mu)$	μ	ϕ
Gamma $\Gamma(\phi, \theta)$, $\theta = \mu\phi$	$2\left\{\frac{1}{\mu} - \frac{y}{\mu^2}\right\}$	μ	$\mu^2\phi$
Inverse normal $\mathcal{IN}(\mu, \sigma^2)$	$\frac{2}{\mu^2} - \frac{2y}{\mu^3}$	μ	$\mu^3\phi$
Lognormal $\mathcal{LN}(\theta, \sigma^2)$, $\theta = \ln(\mu)$	$-\frac{2}{\mu}(\ln(y) - \ln(\mu))$	$\mu e^{\phi/2}$	$\left(e^{\phi} - 1\right)\mu^2 e^{\phi}$
von Mises $vM(\mu, \sigma^2)$	$-2\sin(y-\mu)$	μ	$1 - \frac{I_1\left(\frac{1}{\phi}\right)}{I_0\left(\frac{1}{\phi}\right)}$
simplex $S^-(\mu, \sigma^2)$	$-2\frac{(y-\mu)(\mu^2+y-2\mu y)}{y(1-y)\mu^3(1-\mu)^3}$	μ	no closed form expression available

Two remarks are important here. Firstly, the expression for Q^+ is inherent to exponential dispersion models, and hence inherently relies on the approximation (2.19). Secondly, as for the quasi-likelihood, the extended quasi-likelihood not necessarily leads to the same score function as obtained via (2.24). However, for proper dispersion models one has to use the generalized quasi-score function of [50] to end up with an equivalent expression as in (2.25).

3. Generalized Proper Dispersion Models: estimation and variable selection

We now fully turn to the regression setting in which the aim is to explain the impact of vectors of covariates $\mathbf{X} = (X_1, \ldots, X_p)^T$, and $\mathbf{Z} = (Z_1, \ldots, Z_q)^T$ (of dimensions p and q respectively) on respectively the mean and the dispersion of the response Y. The main difference with the non-regression setting in Section 2 is that now the modelling concerns the conditional distribution function of Y given the set of covariates vectors $(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}, \mathbf{z})$, with $\mathbf{x} = (x_1, \ldots, x_p)^T$ $\mathbf{z} = (z_1, \ldots, z_q)^T$ instead of unconditional distributions in the previous section. Note that the two covariate vectors $\mathbf{X}$ and $\mathbf{Z}$ may contain common covariates (i.e. have overlap) or may even both contain all measured covariates (a case of $p = q$). In real data analyses, an important question is which covariates should

be placed in the mean-dependence part and which ones should be placed in the dispersion-dependence part. Since in our setting, nothing prevents for the covariate vectors $\mathbf{X}$ and $\mathbf{Z}$ to be the same or to partially overlap, we can put all the covariates in both parts, and as long as we are not in a hyper-high-dimensional setup, we may apply our variable selection procedure to make the appropriate choice. See also Section 8.4 for some further discussion.

Suppose that Y given $(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}, \mathbf{z})$ has a proper dispersion conditional density function given by

$$f_{Y|(\mathbf{X},\mathbf{Z})}(y; \mu(\mathbf{x}), \phi(\mathbf{z})) = c(\phi(\mathbf{z}))\,\{V(y)\}^{-1/2} \exp\left\{-\frac{1}{2\phi(\mathbf{z})} d(y, \mu(\mathbf{x}))\right\}, \quad y \in \Omega. \tag{3.1}$$

We denote this conditional model as $(Y|(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}, \mathbf{z})) \sim \mathrm{PD}(\mu(\mathbf{x}), \sigma^2(\mathbf{z}))$.

Analogously as in generalized linear models, we assume that by considering appropriate link functions the dependence on the covariates is linear, i.e.

$$g(\mu(\mathbf{x})) = \mathbf{x}^T\boldsymbol{\beta} \qquad \text{and} \qquad h(\phi(\mathbf{z})) = \mathbf{z}^T\boldsymbol{\gamma},$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the vectors of unknown regression parameters, of dimensions p and q respectively. In case of inclusion of intercept terms (introducing then $X_0 = 1$ and $Z_0 = 1$) the dimensions become $p + 1$ and $q + 1$. The interest is in joint estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

Consider the triple $(Y, \mathbf{X}, \mathbf{Z})$ of random variables for which $f_{Y|(\mathbf{X},\mathbf{Z})}(y; \mu(\mathbf{x}), \phi(\mathbf{z}))$ is as in (3.1). Given an i.i.d. sample $((Y_1, \mathbf{X}_1, \mathbf{Z}_1), \ldots, (Y_n, \mathbf{X}_n, \mathbf{Z}_n))$ from $(Y, \mathbf{X}, \mathbf{Z})$, with $\mathbf{X}_j = (X_{1j}, \ldots, X_{pj})^T$, and $\mathbf{Z}_j = (Z_{1j}, \ldots, Z_{qj})^T$, the conditional log-likelihood function is then given by

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \left\{\ln c(\phi(\mathbf{Z}_i)) - \frac{1}{2}\ln V(Y_i) - \frac{1}{2\phi(\mathbf{Z}_i)} d(Y_i, \mu(\mathbf{X}_i))\right\}. \tag{3.2}$$

For estimation of $\boldsymbol{\beta}$ the focus is on (3.2) with $\boldsymbol{\gamma}$ (and thus $\phi(\mathbf{Z}_i)$) fixed. For estimation of $\boldsymbol{\gamma}$, one considers (3.2) with $\boldsymbol{\beta}$ (and thus $\mu(\mathbf{X}_i)$) given, which results in looking at $D_i = d(Y_i, \mu(\mathbf{X}_i))$ as the response variables.

Remark 3.1. *Recall that the score function for $\boldsymbol{\beta}$ is the partial derivative of the proper dispersion log-likelihood with respect to $\boldsymbol{\beta}$ holding $\boldsymbol{\gamma}$ fixed. When $\boldsymbol{\gamma}$ is fixed, however, the following remark facilitates the generalized extension of the quasi-likelihood function. For a datum point $(Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ denote $\mu_i = \mu(\mathbf{X}_i)$ and $\phi_i = \phi(\mathbf{Z}_i)$, and define the pseudo response value associated to this ith datum point as $\widetilde{Y}_i = \mu_i - \phi_i \cdot \left(\frac{\dot{d}(Y_i,\mu_i)/2}{var(-\dot{d}(Y_i,\mu_i)/2)}\right)$. Herein $\dot{d}$ and $\ddot{d}$ denote respectively the first and second order partial derivative of the unit deviance d with respect to the second argument μ. Note that the second order Bartlett identity for μ_i gives $\mathbb{E}\left(\ddot{d}(Y_i, \mu_i)/2\right) = var(-\dot{d}(Y_i, \mu_i)/2)/\phi_i$. For exponential dispersion models one has $Y_i = \widetilde{Y}_i$, while it is easy to see that for proper dispersion models the log-likelihood score for $\boldsymbol{\beta}$ with $\boldsymbol{\gamma}$ fixed is a quasi-likelihood score associated to the pseudo response vector $\widetilde{\mathbf{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_n)^T$.*

Apart from estimating the effect of a covariate on the response, we also want to determine at the same time which of the (many) covariates have a significant effect. We therefore aim to develop an efficient penalized log-likelihood to select important explanatory variables, among the sets $\mathbf{X} = (X_1, \ldots, X_p)^T$, and $\mathbf{Z} = (Z_1, \ldots, Z_q)^T$ that make a significant contribution to the joint mean and dispersion of $\mathrm{PD}(\mu(\mathbf{x}), \sigma^2(\mathbf{z}))$ models. Similar in spirit to [30], we define the penalized log-likelihood function by

$$J(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) - \sum_{j=0}^{p} p_{\lambda_1}(t_j|\beta_j|) - \sum_{j=0}^{q} p_{\lambda_2}(u_j|\gamma_j|), \tag{3.3}$$

where $p_{\lambda_k}(\cdot)$, $k = 1, 2$, is a nonconcave penalizing function that is indexed by a regularization parameter $\lambda_k > 0$. These parameters are preferably chosen by a data-driven criterion (such as (generalized) cross-validation). See also Section 5.2. The sequences $(t_j)_{j=0,\ldots,p}$ and $(u_k)_{k=0,\ldots,q}$ are known (given) nonnegative weights allowing a different amount of regularization for each of the covariates involved in the regression. For more details, see Section 5.

In practical settings, many variables are introduced and the number of these can depend on the sample size. As in [30] and [4], we therefore allow for p and q to possibly depend on n, that is $p = p_n$ and $q = q_n$. Maximization of $J(\boldsymbol{\beta}, \boldsymbol{\gamma})$, alternating again between fixing $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ leads to the penalized maximum log-likelihood estimators. Due to the inclusion of penalty functions, the solution of these optimization problems will simultaneously select a significant variable and estimate its associated parameter coefficient.

To select a good penalty function [2] (see also [29]) proposed three principles that a good penalty function should satisfy: unbiasedness, sparsity and continuity. [29] constructed a new penalty function (SCAD) which results in an estimator that achieves an oracle property: that is, the estimator has the same limiting distribution as an estimator which knows the true model a priori. The smoothly clipped absolute deviation (SCAD) penalty, which is defined by

$$p'_\lambda(\beta) = \lambda \left\{ I(\beta \le \lambda) + \frac{a\lambda - \beta}{(a-1)\lambda} I(\beta > \lambda) \right\},$$

for some $a > 2$ and $\beta > 0$, where $I(\cdot)$ is the indicator function, satisfies the above mentioned three properties. See also [4] for a discussion on appropriate choices of penalty functions that yield variable selection procedures.

Variable selection for joint mean and dispersion using (3.3), with SCAD or LASSO penalties has been studied in the literature for an inverse Gaussian distribution in [73], and for a lognormal distribution in [74]. It should be noted though that for fitting the lognormal distribution into the framework of a proper dispersion model the settings as in Table 1 are needed, for example to ensure the Fisher-orthogonality property. The expressions used in [74] are different though.

4. Theoretical results

Penalization methods are characterized by loss functions measuring data fits and penalty terms constraining model parameters. The commonly used quadratic loss minimizing the sum-of-squares error is equivalent to assuming a Gaussian distributed error term and leads to a least squares algorithm for maximum likelihood estimation. However it is not suitable for maximum likelihood estimation in more general generalized linear and nonlinear models.

We study in this section the asymptotic properties of the penalized likelihood estimators introduced previously. We consider the situations where the number of parameters is either fixed or tends to ∞ with increasing sample size n. We show in particular how model selection and estimation in generalized proper dispersion models can be bridged within the unified framework of penalized Bregman divergence obtained by replacing the negative log-likelihood or generalized quasi-likelihood in the conventional penalized likelihood or quasi-likelihood regression with Bregman divergence (see (4.1) for the definition).

We noted before that for proper dispersion models the mean submodel parameter $\boldsymbol{\beta}$ and the dispersion submodel parameter $\boldsymbol{\gamma}$ are orthogonal. Such a property implies in particular that the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be obtained separately. Holding $\boldsymbol{\gamma}$ fixed at the current estimate $\widehat{\boldsymbol{\gamma}}$, the estimate of $\boldsymbol{\beta}$ is obtained by solving the associated generalized quasi-score estimating equation (see Remark 3.1), using $\phi(\widehat{\boldsymbol{\gamma}})$ as the prior weight. Holding $\boldsymbol{\beta}$ fixed at the current estimate $\widehat{\boldsymbol{\beta}}$ the estimate of $\boldsymbol{\gamma}$ is obtained by fitting a gamma exponential model log-likelihood function for the dispersion parameter. Iterating between holding $\boldsymbol{\beta}$ fixed and $\boldsymbol{\gamma}$ fixed leads then to the penalized maximum likelihood estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ which maximize the objective function (3.3).

The above remark can be now used to show in particular how model selection and estimation in generalized proper dispersion models can be bridged within the unified framework of penalized Bregman divergence obtained by replacing the negative log-likelihood or generalized quasi-likelihood in the conventional penalized likelihood or quasi-likelihood regression with Bregman divergence.

There is a rich literature on connection of Bregman divergence and probability distributions. In one study [8] showed the bijection between regular exponential family distributions and the Bregman divergence. The connection of Bregman divergence and Tweedie distributions has been briefly remarked by [15], and has been specifically studied in a recent report by [77].

Bregman divergences are introduced by Bregman in 1967. By definition, for any real valued differentiable concave function q, the Bregman divergence of τ from ν is given by [8]

$$B(\nu, \tau) = q(\nu) - q(\tau) - (\nu - \tau)q'(\tau). \tag{4.1}$$

It is equal to the tail of the first-order Taylor expansion of $q(\nu)$ at τ. Major classes of cost functions can be generated by the Bregman divergence with appropriate functions q (see e.g. [8] and [79]. They enjoy convex duality properties. Conversely, for a given B-loss, [79] derived an explicit formula for solving the

generating q function. The concavity of q ensures that Bregman divergences are non-negative quantities, i.e. $B(\nu,\tau) \geq 0$ and equality holds only for $\nu = \tau$. However, in general they do not possess the symmetry property, nor do they enjoy a triangular inequality. Hence, they cannot be considered as metrics or distances in the strict sense.

As noted in Remark 3.1 for proper dispersion models the log-likelihood score for $\boldsymbol{\beta}$ with $\boldsymbol{\gamma}$ fixed is a quasi-likelihood score associated to the pseudo response vector $\widetilde{\mathbf{Y}}$. [79] verified that this (negative) quasi-likelihood function belongs to the Bregman divergence and derived the generating q-function,

$$q(\mu) = \int_{\mu_0}^{\mu} \frac{t-\mu}{V(t)} dt \tag{4.2}$$

where μ_0 is a fixed constant such that the integral is well-defined and where V is the variance function of a proper regular dispersion model.

To unify the notation, we will denote hereafter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ the generic vector of unknown parameters, of dimension $m_n = p_n + q_n + 2$ and $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ will denote the penalized maximum likelihood estimator. Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$. We will assume that the vector $\boldsymbol{\theta}_0$ may be sparse and has the representation $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_0^{(1)}, \boldsymbol{\theta}_0^{(2)})$ where, without any loss in generality, $\boldsymbol{\theta}_0^{(1)}$ consists of the nonzero components of $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0^{(2)}$ regroups the possibly zero components of $\boldsymbol{\theta}_0$. Let s_n denote the number of nonzero components of $\boldsymbol{\theta}_0$, i.e. the dimension of the vector $\boldsymbol{\theta}_0^{(1)}$. Note that $s_n = s_{1n} + s_{2n}$, where s_{kn}, for $k = 1, 2$, is the number of nonzero components for respectively the mean and the dispersion part. Similarly to [30] define, for $k = 1, 2$,

$$a_{kn} = \max_{1 \leq j \leq s_{kn}} \left\{ |p'_{\lambda_{kn}}(|\theta_{0j}|)| \right\} \qquad b_{kn} = \max_{1 \leq j \leq s_{kn}} \left\{ |p''_{\lambda_{kn}}(|\theta_{0j}|)| \right\},$$

where $\lambda_{kn} > 0$ is a regularization parameter depending on the sample size n. See (3.3).

The following theorem guarantees, with probability tending to 1, the existence of a consistent local maximizer $\widehat{\boldsymbol{\theta}}$ of the penalized log-likelihood function in (3.3) and states that $\widehat{\boldsymbol{\theta}}$ is a $\sqrt{n/m_n}$-consistent estimator of $\boldsymbol{\theta}_0$. The proof of the theorem is given below.

Theorem 4.1. *Assume $a_{kn} = O(n^{-1/2})$ and $b_{kn} \to 0$, for $k = 1, 2$, as $n \to \infty$. Assume further $m_n^4/n \to 0$, $\max(\lambda_{1n}, \lambda_{2n}) \to 0$, $(n/m_n)^{1/2} \min(\lambda_{1n}, \lambda_{2n}) \to \infty$ and $\min_{j=1,\ldots,s_n}(|\theta_{0j}|)/\max(\lambda_{1n}, \lambda_{2n}) \to \infty$ as $n \to \infty$. Then, under the regularity conditions stated in the Appendix, there exists, with probability tending to one, a local maximizer $\widehat{\boldsymbol{\theta}}$ of the penalized log-likelihood function in (3.3) such that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = O_P((m_n/n)^{1/2})$.*

The model selection consistency of the local maximizer $\widehat{\boldsymbol{\theta}}$ is given in Theorem 4.2 below, the proof of which follows.

Theorem 4.2. *Under the same regularity conditions as those required in Theorem 4.1 and if $m_n^5/n \to 0$, $\max(\lambda_{1n}, \lambda_{2n}) \to 0$, $(n/m_n)^{1/2} \min(\lambda_{1n}, \lambda_{2n}) \to \infty$,*

$\min_{j=1,\ldots,s_n}(|\theta_{0j}|)/\max(\lambda_{1n},\lambda_{2n}) \to \infty$ *as* $n \to \infty$ *and for* $k = 1, 2$

$$\liminf_{n\to\infty} \liminf_{t\to 0_+} \frac{p'_{\lambda_{kn}}(t)}{\lambda_{kn}} > 0,$$

then the local maximizer $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}^{(1)}, \widehat{\boldsymbol{\theta}}^{(2)})$ *in Theorem 4.1 satisfies* $\widehat{\boldsymbol{\theta}}^{(2)} = 0$ *with probability tending to 1.*

Remark 4.1. *When* m_n *(and therefore* s_n*) are fixed and do not depend on* n*, the above results can be derived under weaker conditions, using Theorem 1 of [39] on penalized estimating functions and their condition C.1 which is insured by the results of [7] on estimating equations for proper dispersion models. However, since these results do not extend to the case where* m_n *and* s_n *may diverge to infinity, we have not pursued this approach in the paper.*

Proof of Theorem 4.1. We assume that the regularity conditions stated in the Appendix hold. Note that if $\boldsymbol{\gamma}_0$ is the true vector of dispersion parameters, an oracle with the knowledge of $\boldsymbol{\gamma}_0$ would estimate the mean regression coefficients vector $\boldsymbol{\beta}_0$ by maximizing the penalized objective function $\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}_0) - \sum_{j=1}^{p} p_{\lambda_1}(\beta_j)$ with respect to $\boldsymbol{\beta}$. Denote by $\widetilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}_0)$ the corresponding "estimator". Given that the Bartlett identities are true for the proper dispersion models considered in this paper, the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are orthogonal and the estimation of $\boldsymbol{\beta}$ is insensitive to $\boldsymbol{\gamma}$ in the sense of [44]. It follows that the maximum penalized log-likelihood estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is asymptotically independent of $\widehat{\boldsymbol{\gamma}}$ and, when the conclusions of Theorem 4.1 are true is consistent at either a parametric rate (when p_n is fixed) or at a rate $(s_n/n)^{1/2}$ for a particular chosen λ_{1n}. In particular $(s_n/n)^{1/2}\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_\infty \to 0$ as $n \to \infty$. This shows that the marginal profile penalized maximum likelihood estimator of $\boldsymbol{\gamma}$ defined as the maximizer of $\ell(\widetilde{\boldsymbol{\beta}}_{\lambda_{1n}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}) - \sum_{j=1}^{q_n} p_{\lambda_{2n}}(|\gamma_j|)$ is asymptotically equal to the maximum penalized likelihood estimator $\widehat{\boldsymbol{\gamma}}$ and can be reached with the same iterative scheme as the one adopted for the iterative estimation of the parameters in Section 4. According to the above, it is therefore sufficient to prove the conclusion of Theorem 4.1, separately for each vector component of $\widehat{\boldsymbol{\theta}}$, i.e. $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}$.

We have seen that $\widehat{\boldsymbol{\gamma}}$ the marginal profile penalized maximum likelihood estimator of $\boldsymbol{\gamma}$ estimator, is a penalized Bregman divergence estimator with q function, the function q_2 associated to a regular Gamma-like exponential dispersion family and as such it satisfies condition A6 of [80], while the SCAD penalty satisfies condition A8. The conditions stated for the proof of Theorem 1, when restricted to the design and the link function regarding the dispersion part, are the same as those of the remaining items stated in condition A of [80]. The conclusion of our Theorem 4.1 for the estimator $\widehat{\boldsymbol{\gamma}}$ follows then as a direct application of Theorem 1 of [80].

Regarding now the consistency properties of $\widehat{\boldsymbol{\beta}}$ these follow using similar arguments to the above. Indeed when $\boldsymbol{\gamma}$ is fixed to its estimated value $\widehat{\boldsymbol{\gamma}}$, the estimator of $\boldsymbol{\beta}$, based on the pseudo-response vector $\widetilde{\mathbf{Y}}$ appears again as a

penalized Bregman divergence estimator with generating function $q_1(\cdot)$ the one given in (4.2). It is easy to see that for proper dispersion models this q-function also satisfies condition A6 of [80], and that the remaining conditions on the design and link function of the mean submodel are a subset of the conditions sated in condition A of [80]). Therefore our Theorem 4.1 is true as a corollary of Theorem 1 of the above cited authors. □

Proof of Theorem 4.2. We again assume that the regularity conditions stated in the Appendix hold. The proof of Theorem 4.2 is also a corollary of Theorem 2 part (i) of [80] once the conditions $\liminf_{n\to\infty} \liminf_{t\to 0_+} \frac{p'_{\lambda_{kn}}(t)}{\lambda_{kn}} > 0$, for $k = 1, 2$, are verified. For the choices made for the asymptotic behavior of the penalty parameters λ_{kn}, $k = 1, 2$ and the fact that the penalty functions are the same SCAD penalty, the condition is true (see, e.g. [29]) and this completes the proof of Theorem 4.2. □

Remark 4.2. *An important issue is in general the multiplicity of local minimizer of the penalized loss. On this issue, one may follow the same route as in [4] for the* SCAD *penalty, using, instead of the* SCAD*, a δ-approximation to the* SCAD *penalty in the neighborhood of the origin, in a similar spirit to the local linear approximation used by [29]. Such an approximation ensures the existence of an appropriate local maximizer with the same asymptotic behavior.*

5. Computation

To ensure the practicality of the statistical methodology for estimation and variable selection in generalized proper dispersion models discussed in Section 3, we propose in this section a simple algorithm to solve the corresponding penalized maximum log-likelihood problem. In the following we focus on the maximization problem with non convex penalties

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) = \text{argmax}_{(\boldsymbol{\beta},\boldsymbol{\gamma})} J(\boldsymbol{\beta}, \boldsymbol{\gamma}), \tag{5.1}$$

where the penalized log-likelihood function is given by expression (3.3), with the loss $\ell(\boldsymbol{\beta}, \boldsymbol{\gamma})$ the conditional log-likelihood function defined in (3.2). The designs matrices $\mathbf{X}_{\text{D}}$ and $\mathbf{Z}_{\text{D}}$ both contain a column of 1's for the intercepts with corresponding coefficients β_0 and γ_0, but their other columns are assumed to have zero-mean, as it is standard practice. Concerning the weights t_j's and u_k's in (3.3), taking $t_0 = u_0 = 0$ allows to not penalize for the intercepts. Taking the remaining weights proportional to the standard deviations of the corresponding columns of the design matrices ensures that the penalty is applied equally to all covariates in an equivariant manner.

We have seen that optimization of the $p+q+2$ dimensional problem (5.1) reduces to two separate optimization problems of dimensions $p+1$ and $q+1$. As a consequence the unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be respectively estimated by alternating between an extended generalized penalized quasi-likelihood estimation procedure for $\boldsymbol{\beta}$ in the mean submodel and a standard penalized exponential

family log-likelihood estimation procedure for $\boldsymbol{\gamma}$ in the corresponding dispersion submodel. To see this recall from (2.20) that for the parameter $\boldsymbol{\gamma}$ we need to maximize

$$\ln c(\phi) - \frac{1}{2\phi} d(y,\mu),$$

with respect to ϕ. When working with $c(\phi) = (\sqrt{2\pi\phi})^{-1/2}\rho(\phi)$, it is then easily verified, keeping in mind (2.16) and (2.17), that this is equivalent with considering for the random variable $D = d(Y,\mu)$ a regular exponential dispersion model with parameter $\theta = -\dfrac{1}{\phi}$ and cumulant function

$$\kappa_D(\theta) = -\ln(-\theta) - 2\ln\rho\big(-\frac{1}{\theta}\big)$$

with first and second derivative functions

$$\begin{aligned} \kappa'_D(\theta) &= -\frac{1}{\theta} - 2\psi\big(-\frac{1}{\theta}\big)\,\frac{1}{\theta^2} \\ \text{and}\quad \kappa''_D(\theta) &= \frac{1}{\theta^2} - 2\psi'\big(-\frac{1}{\theta}\big)\,\frac{1}{\theta^4} + 4\psi\big(-\frac{1}{\theta}\big)\,\frac{1}{\theta^3}\,, \end{aligned} \tag{5.2}$$

where $\psi = \rho'/\rho$. This leads to (for μ given), $\delta = \mathbb{E}(D) = \kappa'_D\big(-\frac{1}{\phi}\big)$. We obtain the unit variance function for the dispersion model by $V_D(\delta) = \kappa''_D(\delta)$. Further, we denote the unit variance function for the mean submodel by $V_M(\mu)$. In the particular case of the Gamma distribution a similar approach to this can be found in [64].

From a computational point of view, Taylor expansions of the relevant Bregman divergence loss functions (see also [35]), show that the optimizations in the two submodels are equivalent in finding the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ by solving iteratively the following approximate Fisher scoring interlinked weighted penalized least squares problems:

$$\min_{\boldsymbol{\beta}} \left\{ (\mathbf{a}_M - \mathbf{X}_{\text{D}}\boldsymbol{\beta})^T \mathbf{W}_M (\mathbf{a}_M - \mathbf{X}_{\text{D}}\boldsymbol{\beta}) + \sum_{j=1}^{p} p_{\lambda_1}(t_j|\beta_j|) \right\}, \tag{5.3}$$

$$\min_{\boldsymbol{\gamma}} \left\{ (\mathbf{a}_D - \mathbf{Z}_{\text{D}}\boldsymbol{\gamma})^T \mathbf{W}_D (\mathbf{a}_D - \mathbf{Z}_{\text{D}}\boldsymbol{\gamma}) + \sum_{k=1}^{q} p_{\lambda_2}(s_k|\gamma_k|) \right\}, \tag{5.4}$$

where $\mathbf{X}_{\text{D}}$ and $\mathbf{Z}_{\text{D}}$ are the design matrices for the mean and dispersion submodel, and where $\mathbf{W}_M$ and $\mathbf{W}_D$ are the diagonal matrices of working weights, for the mean and the dispersion respectively, defined by:

$$\mathbf{W}_M = \operatorname{diag}\left\{ \left[\frac{\partial g(\mu_i)}{\partial \mu}\right]^{-2} \frac{1}{\phi_i V_M(\mu_i)} \right\}, \quad \mathbf{W}_D = \operatorname{diag}\left\{ \left[\frac{\partial h(\phi_i)}{\partial \phi}\right]^{-2} \frac{V_D(\delta_i)}{2\phi_i^4} \right\},$$

where the ith components of the working vectors $\mathbf{a}_M$ and $\mathbf{a}_D$ are respectively defined by

$$\mathbf{a}_{Mi} = \mathbf{x}_i^T\boldsymbol{\beta} + \frac{\partial g(\mu_i)}{\partial \mu}(\widetilde{y}_i - \mu_i), \quad \mathbf{a}_{Di} = \mathbf{z}_i^T\boldsymbol{\gamma} + \frac{\partial h(\phi_i)}{\partial \phi}\frac{\phi_i^2}{V_D(\delta_i)}(D_i - \delta_i).$$

It is understood that all terms in the above iterations are evaluated at their previous iteration values. In general, the initial values for each of the parameters are set to be the un-penalized maximum likelihood estimates. See also Remark 5.2.

Therefore the proposed penalized approximate Fisher scoring algorithm consists of two layers of loops. For each parameter, the outer layer is the Iteratively Reweighted Least Squares (IRLS) strategy which at each iteration approximates the objective function in (5.1) by the penalized objective function (5.3) (or (5.4)). After obtaining the minimizer of the penalized objective function the next iteration begins by updating the working response vector and weights. The outer loop continuous until convergence. The inner layer is dedicated to obtaining the minimizer of the penalized weighted least squares objective function by a thresholding based iterative procedure that we will now describe.

5.1. A general iterative thresholding algorithm

We will start by describing an efficient iterative shrinkage and thresholding algorithm to solve a class of non-smooth and nonconvex optimization problems, covering the important special case of the penalized weighted least squares minimization problem introduced in this section. Without loss of generality consider a weighted least squares penalized problem of the form

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\frac{1}{2}\|X_{\text{D}}\boldsymbol{\beta}-\mathbf{y}\|^2_{\mathbf{W}}+P_\lambda(\boldsymbol{\beta})\right\} \tag{5.5}$$

where $X_{\text{D}}=[\mathbf{x}_1,\ldots,\mathbf{x}_p]$ is the $n\times p$ design matrix, $\mathbf{W}=\text{diag}(w_1,\ldots,w_n)$ is a diagonal matrix with positive entries, defining the norm $\|\mathbf{u}\|^2_{\mathbf{W}}:=\mathbf{u}^T\mathbf{W}\mathbf{u}$, $\mathbf{y}\in\mathbb{R}^n$ is the response vector and $P_\lambda(\boldsymbol{\beta})$ represents a non-smooth and nonconvex penalty that is separable, i.e. $P_\lambda(\boldsymbol{\beta})=\sum_{j=1}^p r_\lambda(t_j|\beta_j|)$ with t_j, $j=1,\ldots,p$ some given scaling weights, $r_\lambda(\cdot)$ a scalar penalty function with penalty λ as a regularization parameter. Without loss of generality (changing $\mathbf{X}_{\text{D}}\to\mathbf{W}^{1/2}\mathbf{X}_{\text{D}}$ and $\mathbf{y}\to\mathbf{W}^{1/2}\mathbf{y}$ in equation (5.5)) we consider the case of minimizing

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\frac{1}{2}\|\mathbf{X}_{\text{D}}\boldsymbol{\beta}-\mathbf{y}\|^2_2+P_\lambda(\boldsymbol{\beta})\right\}, \tag{5.6}$$

where now $\|\cdot\|_2$ is the standard Euclidian norm of $\mathbb{R}^n$.

However note that, in our context, the above outer loop iterative reweighing depends on both the data and the parameter values at the previous iteration and therefore introduces some difficulties. This explains why, instead of the classical penalization on the components of $\boldsymbol{\beta}$, we propose during the iterations an adaptive rescaling of the penalties by simply replacing λ in P_λ by a component-specific $\widetilde{\boldsymbol{\lambda}}=(\lambda\|w_1^{1/2}\mathbf{x}_1\|_2,\lambda\|w_2^{1/2}\mathbf{x}_2\|_2,\ldots,\lambda\|w_p^{1/2}\mathbf{x}_p\|_2)^T$ to match the continuously changing scale of the covariates. The algorithmic consequences for the optimization procedure are straightforward.

In the simplified case where the design matrix is orthogonal i.e. $\mathbf{X}_{\mathrm{D}}^T\mathbf{X}_{\mathrm{D}} = \mathbf{I}_p$, like for wavelet denoising (see [3]), changing $\mathbf{y} \to \mathbf{z} := \mathbf{X}_{\mathrm{D}}^T\mathbf{y}$ the loss part in the objective function in (5.6) becomes $\|\boldsymbol{\beta} - \mathbf{z}\|_2^2$ and is a strictly convex function of $\boldsymbol{\beta}$ that is also separable. In such a case, to solve the minimization problem (5.6) we only need to deal with the univariate case

$$\min_{\beta} \left\{ \frac{(\beta - z)^2}{2} + r_\lambda(t|\beta|) \right\}. \tag{5.7}$$

When $r_\lambda(\cdot)$ is a (closed) convex function, the solution of (5.7) is unique, and defines the proximal mapping or proximal operator of r_λ given by:

$$\mathrm{Prox}_{r_\lambda}(z) = \arg\min_{u} \left\{ \frac{(u - z)^2}{2} + r_\lambda(t|u|) \right\}. \tag{5.8}$$

While proximal operators are well studied for convex functionals, the nonconvex case has been of interest to researchers recently (see [36]). Very often, even when $r_\lambda(\cdot)$ is nonconvex, problem (5.7) results in a unique solution and allows us to define an appropriate proximal map. The resulting proximal operator allows then to solve the optimization problem (5.6) by a generalized gradient projection algorithm of similar flavor as the one described, in the infinite-dimensional case, by [10] (see the equation (3) therein). In fact the algorithm works by generating a sequence of iterates $\{\boldsymbol{\beta}^{(k)}, k = 0, 1, \dots\}$ and is tailored to the following problem which can be set up and solved efficiently at each iteration

$$\arg\min_{\boldsymbol{\xi}} \left\{ \frac{1}{2}\|\boldsymbol{\xi} - \boldsymbol{\beta}^{(k)} + \mathbf{X}_{\mathrm{D}}^T\mathbf{X}_{\mathrm{D}}\boldsymbol{\beta}^{(k)} - \mathbf{X}_{\mathrm{D}}^T\mathbf{y}\|_2^2 + P_\lambda(\boldsymbol{\xi}) \right\}. \tag{5.9}$$

Under appropriate conditions, we will see later that this iterative algorithm converges to a global minimizer of problem (5.6). With a proper scaling of the design matrix $\mathbf{X}_{\mathrm{D}}$, one may also include a relaxation parameter $\omega > 0$ into the iterations, which is an efficient way, in certain cases, to accelerate the convergence.

Remark 5.1. *Instead of using the generalized projection method described above for solving the optimization problem (5.6), one may be tempted to use a kind of majorization-minimization (MM) approach (see [46]) via surrogate functionals as proposed in [20] in the functional case where one replaces the objective function in (5.6) with*

$$\Phi(\boldsymbol{\xi}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{X}_D\boldsymbol{\beta} - \mathbf{y}\|_2^2 + P_\lambda(\boldsymbol{\beta}) + \frac{C}{2}\|\boldsymbol{\xi} - \boldsymbol{\beta}\|^2 + \frac{1}{2}\|\mathbf{X}_D\boldsymbol{\beta} - \mathbf{X}_D\boldsymbol{\xi}\|_2^2,$$

and defines an iteration through $\boldsymbol{\beta}^{(k+1)} = arg\min_{\boldsymbol{\xi}} \Phi(\boldsymbol{\xi}, \boldsymbol{\beta}^{(k)})$. *Indeed this produces decreasing values of the original objective function but one has to solve "a fixed-point" problem at each iteration (see [61]). When the* ℓ_1 *penalty is used, such an iterative procedure can be shown to converge to a global optimum even if* $\mathbf{X}_D$ *is not of full rank. It has been proposed in different forms ([63], [20]) and is*

strongly advocated for large-data problems. However, in the nonconvex case as is the one we are looking at in this subsection, this fixed-point equation becomes an inclusion with a discontinuous operator, and hence, there is no guarantee of convergence.

A critical step of the iterative algorithm that we propose is the computation of the proximal map for far more flexible penalties in the algorithmic design including nonconvex ones (see [3]). For the SCAD penalty adopted in this paper the proximal mapping has a closed-form expression. To simplify the notation, the proximal mapping we are trying to compute is the one which solves the minimization problem

$$\begin{aligned}\operatorname{Prox}_{r_{\lambda,s}}(z) &= \arg\min_u \left\{ \frac{(u-z)^2}{2} + s \cdot r_\lambda(t|u|) \right\} \\ &= \arg\min_u \left\{ \frac{(u-tz)^2}{2} + s \cdot t^2 r_\lambda(|u|) \right\} \\ &= \operatorname{Prox}_{r_{\lambda,st^2}}(tz),\end{aligned} \tag{5.10}$$

where $r_\lambda(|u|)$ is the SCAD penalty function, t a given nonnegative weight and s a step size for the algorithm. Recall that the SCAD penalty function is given, for some $a > 2$, by

$$r_\lambda(v) = \lambda \int_0^{|v|} \min\left(1, \frac{(a\lambda - x)_+}{(a-1)\lambda}\right) dx.$$

We can now recast problem (5.10) into the following three problems:

$$u_1 = \arg\min_u \left\{ \frac{(u-tz)^2}{2} + \lambda s \cdot t^2 \cdot |u| \right\} \quad \text{s.t. } |u| \le \lambda \tag{5.11}$$

$$u_2 = \arg\min_u \left\{ \frac{(u-tz)^2}{2} - \frac{u^2 - 2a\lambda st^2|u| + (\lambda st^2)^2}{2(a-1)} \right\} \text{s.t.} \lambda < |u| \le a\lambda \tag{5.12}$$

$$u_3 = \arg\min_u \left\{ \frac{(u-tz)^2}{2} + \frac{(a+1)(st^2)^2\lambda^2}{2} \right\} \quad \text{s.t. } a\lambda \le |u|. \tag{5.13}$$

Considering that $a > 2$ we easily obtain

$$u_1 = \operatorname{sign}(z)\min(\lambda, \max(0, t|z| - s \cdot t^2\lambda)) \tag{5.14}$$

$$u_2 = \operatorname{sign}(z)\min(a\lambda, \max(\lambda, \frac{|z|(a-1)/(st) - a\lambda}{(a-2)/(s \cdot t^2)})) \tag{5.15}$$

$$u_3 = \operatorname{sign}(z)\max(a\lambda, t|z|). \tag{5.16}$$

which show that the proximal mapping is just the SCAD thresholding operator. In our case $p < n$ and to avoid unnecessary complications, we assume here that the design matrix $\mathbf{X}_D$ has full column rank. Therefore the matrix $\mathbf{X}_D^T\mathbf{X}_D$ is nonsingular and the SCAD proximal map is a contraction which shows that the algorithm converges to a stationary point of the original optimization problem. Using the results of [10] with the notation adopted in this subsection, whenever

any singular value of $\mathbf{X}_{\text{D}}$, say $\nu(\mathbf{X}_{\text{D}})$, is such that $\frac{1}{a-1} \leq \nu(\mathbf{X}_{\text{D}})^2 \leq \frac{2a-3}{a-1}$ the stationary point is a global minimizer and therefore leads to a unique estimator even when the penalty is nonconvex provided that the proximal map is continuous at $\boldsymbol{\xi}$. When such a condition on $\mathbf{X}_{\text{D}}$ is not satisfied, and assuming that the weight matrix $\mathbf{W}$ remains bounded during the iterations, we can always rescale $\mathbf{X}_{\text{D}}$ and $\mathbf{y}$ by dividing them by $\kappa > \|\mathbf{X}_{\text{D}}\|_2 = \nu_{\max}(\mathbf{X}_{\text{D}})$ and by using λ/κ^2 as a regularization parameter for the SCAD penalty. This doesn't affect the iterative solution.

5.2. Choosing the regularization parameter

As for any regularization problem, choosing proper data-driven regularization parameters in (3.3) is very important. The penalty functions in expressions (5.3) and (5.4) involve two penalty parameters λ_1 and λ_2 that control the amount of regularization. K-fold cross-validation and generalized cross-validation procedures are popular methods for choosing these tuning parameters, but they are rather complicated and computationally intensive. Instead we use, as in [13], a BIC driven Grid Search (BICGS). We first specify a $K_1 \times K_2$ grid $\Lambda_1 \times \Lambda_2 = \{\left(\lambda_1^{(i)}, \lambda_2^{(j)}\right), i = 1, \ldots, K_1; j = 1, \ldots, K_2\}$ within a rectangle $[0, \lambda_{1\,\max}] \times [0, \lambda_{2\,\max}]$ with $\lambda_{1\,\max}$ (respectively $\lambda_{2\,\max}$) large enough to kill all estimated components of $\boldsymbol{\beta}$ (respectively of $\boldsymbol{\gamma}$). We then run the previously described algorithm for every value of (λ_1, λ_2) in the grid to get a solution path $\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}\right)^{(k)}$, $k = 1, \ldots, K_1 \times K_2$, and finally, use BIC as model comparison criterion to find the optimal estimate. The number of grid points K_1 and K_2 are pre-specified. In our implementation the grid points in Λ_1 and Λ_2 are uniform in the log scale. We estimate the prediction accuracy of each estimated model by evaluating the corresponding conditional likelihood loss $\ell\left((\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})^{(k)}\right)$ given by expression (3.2) to which we add a BIC correction term to obtain the criterion

$$\text{BICGS}(\lambda_1^{(k)}, \lambda_2^{(k)}) = -2\ell\left((\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})^{(k)}\right) + \left(\text{DF}(\widehat{\boldsymbol{\beta}}^{(k)}) + \text{DF}(\widehat{\boldsymbol{\gamma}}^{(k)})\right) \log n,$$

where DF is the degrees of freedom function. For the SCAD penalty, DF is approximately the number of nonzero components in the estimate. The optimal estimate is then chosen from the original solution path by minimizing the BICGS criterion.

Remark 5.2. *When the objective function is (Bregman divergence) strictly convex, the estimated coefficients in the iterations vary continuously with the penalty parameters and produce a path of solutions regularized by the couple* (λ_1, λ_2). *In our case this (strict convexity) is not exactly true all the time, but we have noticed that starting with initial values for the parameters-to-be-estimated given by the un-penalized maximum likelihood estimators and assuming the coefficient paths to be continuous (which is true for example for a SCAD penalty with a=3.7) then using the estimates from the previous values of the regularization*

parameters λ_1 and λ_2 as the initial values for the next values of the regularization parameters doesn't affect the results in a significant way. Therefore, for all the numerical results in this paper, we have followed this approach.

6. Simulation study and comparison

The focus of most multiple regression methodologies center around the conditional mean. However, the understanding of the variability of the data, measured by the conditional variance is very important in many applications. With the methodologies presented in this paper we are able to deal with this under *various model settings*. In this simulation study we consider subsequently

- in Section 6.1 a Gaussian heteroscedastic model with a sparsity scenario for both the mean and the variance:
- in Section 6.2 a skewed distribution, namely a two parameter inverse Gaussian distribution which arises in a variety of applications;
- in Section 6.3 joint modelling of mean and dispersion with the simplex distribution, which is often used for modelling outcomes with continuous proportions (percentages, rates or proportions confined in the unitary interval).

For the Gaussian sparsity model setting of Section 6.1 we can compare the performances of the proposed methods with that of the few competitors that are available in the literature.

For all simulation models and settings we consider samples of size $n = 200$ and summary results of the simulations are based on 150 samples drawn from the described models.

6.1. Heteroscedastic Gaussian regression: analysis and comparison with related methods

Recently, regression estimation under the combination of sparsity and heteroscedasticity in Gaussian regression models has been addressed by [23] (the `HHR` algorithm) and [45] (the `HIPPO` algorithm). Because of the inherent non-convexity of the penalized (pseudo-)log-likelihood considered in these works, the methods proposed therein also alternate between the two parameters, estimating one while keeping the other one fixed. `HHR` performs a doubly regularized likelihood estimation with LASSO penalties to attain sparse estimates for the mean and variance parameters. `HIPPO` is closely related to the iterative `HHR` algorithm but differs from `HHR` firstly by the choice of penalty functions used: `HIPPO` uses the LASSO procedure with heteroscedasticity adjusted penalties using the square-root LASSO procedures of [1] or [9] for the mean parameters and the SCAD penalty for the variance. Secondly, `HIPPO` only carries out two iterations as opposed to `HHR` that continues to iterate with the updated mean and variance parameters until convergence.

Unfortunately, we failed to have available the HIPPO code of [45]. We have therefore used the square-root LASSO procedure as implemented in the R-package flare (see [49]) to perform the first stage of the HIPPO algorithm that selects the mean parameters and adopted the convenient part of the HHR code to perform stage 2 for the variance parameter. For HHR we have used the R-code provided by the authors of that paper. [9] have shown that for a Gaussian or sub-Gaussian regression model with constant but unknown variance, the square-root LASSO with a deterministic value of the penalty parameter that depends only on known parameters achieves near-oracle performance for estimation and model selection of the mean. However, experiments with the optimal λ provided by the square-root LASSO as implemented in flare ([49]) seems to be suboptimal and we have used instead in our implementation of HIPPO the value advocated by [9].

In this subsection, we conduct a small scale simulation study to investigate and compare the finite-sample performances of

- the proposed method, named BREG hereafter;
- the HHR procedure;
- our implementation of the HIPPO procedure as described above,

under a Gaussian heteroscedastic high-dimensional (sparsity) regression model.

The hyper parameters for HHR are chosen by AIC and by Belloni's methods, and for HIPPO by BIC. Our choice on the use of AIC for the HHR procedure is justified by the results of [23] who compared AIC to BIC using Monte Carlo simulations and found that AIC in HHR is preferred for prediction accuracy and inclusion of relevant variables for sample sizes as the ones we are using here. For BREG we always use (in this section and the next one) the BIC procedure described in Section 5.2.

It is well known that the LASSO procedure estimates the nonzero coefficients of the regression vector with a bias towards zero. To attenuate these effects, whatever method is used, we hereafter apply a two-steps procedure that applies first the method (HHR, HIPPO or BREG) with the appropriate penalties, but then, after selection, discards from the regression matrix the columns that correspond to vanishing coefficients and applies the methods with the new regression matrix (so only estimation, without model selection). Such bias removing methods have also been used for example by [54].

Data in the heteroscedastic Gaussian setting are simulated according to

$$Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad i = 1, \ldots, n,$$

using for the Y_i's, the μ_i's and the σ_i's the following model, borrowed from [23] and also used in [45]:

$$Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \exp\Big\{\gamma_0 + \sum_{j=1}^{q} X_{ij}\gamma_j\Big\}\,\epsilon_i, \tag{6.1}$$

with $p = q = 100$, $\beta_0 = 2$, $\gamma_0 = 1$ and

$$\beta_{[12]} = (3, 3, 3, 1.5, 1.5, 1.5, 0, 0, 0, 2, 2, 2),$$

TABLE 3
Criteria used for evaluating the estimation and variable selection performances of a method.

Evaluation criteria	
S	the average number of variables selected
FP	the average number of false positives (i.e. truly zero coefficients/variables that were selected)
FN	the average number of false negatives (i.e. truly nonzero coefficients/variables that were not selected)
ME	the average model error (and its standard deviation)

TABLE 4
Heteroscedastic Gaussian regression model. Estimation and variable selection ability, according to the criteria in Table 3. True (or aimed at) values are indicated in the top row.

method	S $\binom{10}{10}$	FP $\binom{0}{0}$	FN $\binom{0}{0}$	ME (as small as possible)
`HHR`				
β	71.83	61.83	0.0	2.21 (1.25)
γ	6.48	1.0	4.52	3.57 (1.09)
`HIPPO`-like				
β	12.66	3.2	0.54	9.88 (4.06)
γ	1.79	0.52	8.73	6.17 (0.43)
`BREG`				
β	9.3	0.03	0.73	2.08 (2.5)
γ	9.62	1.81	2.18	2.77 (.95)

for the 12 next components of $\boldsymbol{\beta}$ following it's first, and

$$\gamma_{[15]} = (1, 1, 1, 0, 0, 0, 0.5, 0.5, 0.5, 0, 0, 0, 0.75, 0.75, 0.75),$$

for the next 15 components of $\boldsymbol{\gamma}$ following it's first. The remaining components of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ up to p are 0. Hence under this model the number of truly nonzero coefficients in both the mean and the variance is ten. The covariates $\mathbf{X}_j$ are jointly normal with covariance matrix the identity matrix, and the errors ϵ_i are independent and have a standard normal distribution. For each method and each parameter (mean parameter $\boldsymbol{\beta}$ and variance parameter $\boldsymbol{\gamma}$), we report on the evaluation criteria listed in Table 3.

A summary of the simulation results, over the 150 independent runs, is presented in Table 4. Figures 1 and 2 illustrate graphically the results displayed in Table 4 concerning the overall runs performances of `HIPPO`, `HHR` and `BREG`.

From Table 4 and Figures 1 and 2 we observe that from the selection point of view, the `HIPPO`-like procedure has a satisfactory performance, at least for $\boldsymbol{\beta}$. In addition, it outperforms `HHR` which gives very poor results, especially with respect to the false positives. But the model error for the mean is very large, probably due to the under estimation of the dispersion part. These results are consistent with those of Table 2 in [45]. Both methods, `HHR` and `HIPPO`, are outperformed by `BREG` for all criteria.

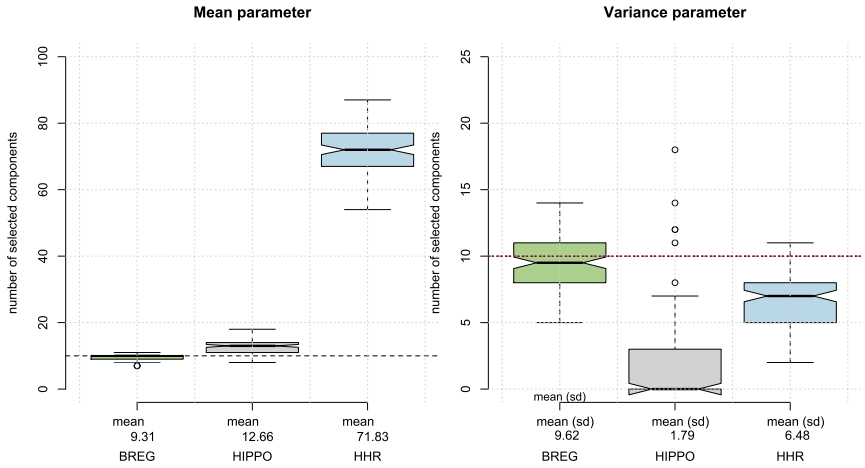

FIG 1. *Boxplots concerning the number of variables selected for the mean (left panel) and for the variance (right panel), for each method. True values are indicated by the dashed horizontal lines.*

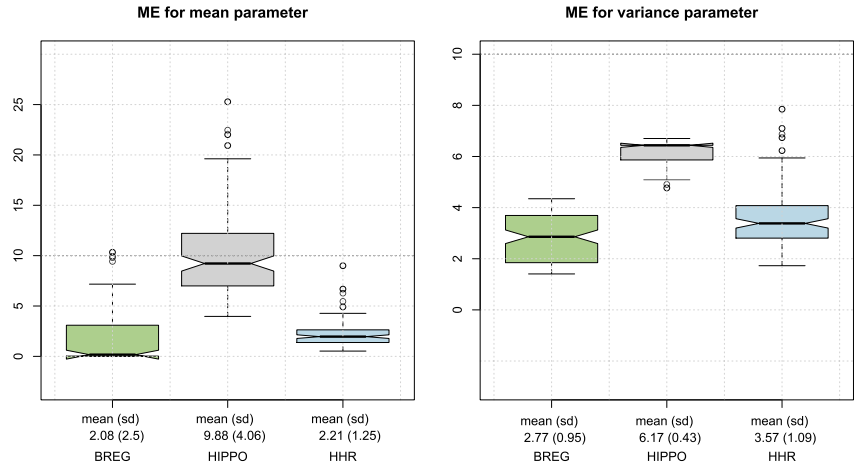

FIG 2. *Boxplots reporting on the estimation error (ME) for the mean (left panel) and for the variance (right panel), for each method.*

6.2. Simulation study with the two parameter inverse Gaussian distribution

In this section we investigate the performance of the proposed `BREG` method in the setting of the inverse Gaussian distribution, where now the data are simulated from the inverse normal regression model (see Table 1)

$$Y_i \sim \mathcal{IN}(\mu_i, \sigma_i^2), \quad i = 1, \ldots, n,$$

TABLE 5

Inverse Gaussian regression model. Estimation and variable selection ability, according to the criteria in Table 3. True (or aimed at) values are indicated in the top row.

parameter	S $\binom{4}{4}$	FP $\binom{0}{0}$	FN $\binom{0}{0}$	ME (as small as possible)
β	3.84	0.83	0.24	0.59 (1.05)
γ	3.91	0.042	0.13	0.27 (0.49)

TABLE 6

Inverse Gaussian regression model. Average over simulation runs of the estimated (and selection) mean and dispersion parameters with their standard deviations in parenthesis.

coefficients	true values	averaged estimates
β_0	2	2.043 (0.426)
β_1	1	0.923 (0.364)
β_2	1	0.922 (0.355)
β_5	1	0.954 (0.362)
γ_0	1	1.009 (0.118)
γ_1	1	0.970 (0.252)
γ_2	1	0.919 (0.307)
γ_5	1	0.973 (0.278)

with

$$\log(\mu_i) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j \quad \text{and} \quad \log(\phi_i) = \gamma_0 + \sum_{j=1}^{q} Z_{ij}\gamma_j \tag{6.2}$$

where $\mathbb{E}(Y_i|\mathbf{X}_i) = \mu_i$, $\mathrm{Var}(Y_i|\mathbf{Z}_i) = \sigma_i^2\mu_i^3$, $\phi_i = \sigma_i^2$, $p = q = 15$, $\beta_0 = 2$, $\gamma_0 = 1$ and

$$\beta_{[5]} = (1, 1, 0, 0, 1)$$

for the 5 next components of $\boldsymbol{\beta}$ following it's first, and

$$\gamma_{[5]} = (1, 1, 0, 0, 1)$$

for the next 5 components of $\boldsymbol{\gamma}$ following the first component. The remaining components of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ up to $p = q = 15$ are 0. Note that the number if nonzero coefficients in both $\log(\mu_i)$ and $\log(\phi_i)$ is 4. The covariates (components of $\mathbf{X}_i$ and $\mathbf{Z}_i$) are jointly uniform on $[-1, 1]$ with covariance matrix equal to the identity matrix. A summary of the simulation results obtained over the 150 simulation runs is displayed in Table 5. Table 6 displays the averages of the estimated mean and dispersion parameters over the 150 simulation runs for the corresponding selections in Table 5.

From Tables 5 and 6, we can make the following observations. For the inverse Gaussian dispersion model the performance of the `BREG` variable selection is slightly better for the dispersion part than for the mean part, when looking at the model error criterion (ME) and model complexity. However, in terms of

TABLE 7
Simplex model. Estimation and variable selection ability, according to the criteria in Table 3. True (or aimed at) values are indicated in the top row.

parameter	S $\binom{4}{4}$	FP $\binom{0}{0}$	FN $\binom{0}{0}$	ME (as small as possible)
β	4.05	0.05	0	0.0018 (0.002)
γ	4.02	0.03	0.01	0.14 (0.15)

TABLE 8
Simplex model. Average over simulation runs of the estimated (and selected) mean and dispersion parameters with their standard deviations in parenthesis.

coefficients	true values	averaged estimates
β_0	2	1.997 (0.026)
β_1	1	0.998 (0.019)
β_2	1	1.003 (0.018)
β_5	1	1.000 (0.018)
γ_0	1	0.971 (0.092)
γ_1	1	0.999 (0.211)
γ_2	1	1.048 (0.202)
γ_5	1	1.019 (0.191)

selecting the right coefficients the behavior is similar for both the mean and the dispersion. As far as estimation of the parameters is concerned the bias is negligible, as can be seen from Table 6.

6.3. Simulation study with the simplex distribution

In this subsection we conduct a simulation study where the proportional data Y_i, $i = 1, \ldots, n$, were generated independently according to a simplex distribution

$$Y_i \sim S^-(\mu_i, \sigma_i^2), \quad i = 1, \ldots, n,$$

with $\log(\mu_i)$ and $\log(\phi_i)$ with $\phi_i = \sigma_i^2$ as in (6.2). Furthermore, the vectors of coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, as well as the vectors of covariates $\mathbf{X}_i$ and $\mathbf{Z}_i$ are as in Section 6.2. The simulation results, obtained over the 150 simulation runs, are summarized in Table 7. Table 8 reports on the summary statistics of the parameter estimates from the proposed `BREG` method.

From Table 7, we learn that the performances in selection and estimation of both mean and dispersion structures are relatively similar and that the selection of the true nonzero population coefficients is efficient. Table 8 indicates that the estimated coefficients of each structure are very close in average to their population values. In summary, our procedure seems to be efficient even for proportional data with heterogeneous dispersion.

7. Real Data Applications

7.1. Ophthalmological data

In this subsection we analyse the ophthalmological data on the use of intra-ocular gas in retinal repair surgeries ([55]), with a special focus on heterogeneous dispersion. A primary analysis of the data assuming homogeneous dispersion was done by [68] and heterogenous dispersion in [69]. Briefly, the study was to investigate the decay course of the intra-ocular gas in retinal repair surgeries prospectively in 31 patients. The gas was injected into the eye before surgery and patients were followed three to eight (average of 5) times over a three-month period. The response variable Y is the percent of gas left in the eye recorded as proportion (a percent). The question is if the disappearance of the gas is related not only to time but also to other covariates such as the concentration of the gas used. The sample size $n = 31$ is relatively small here, but the data have a longitudinal structure (several measurements per patient). To fit these data with our procedure we have adopted for hereafter an independence correlation for the longitudinal structure.

Following the original analysis of [68] we model the mean effects model as

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 \log(t_{ij}) + \beta_2 \log^2(t_{ij}) + \beta_3 x_{ij}, , \quad i = 1, \ldots, m;\ j = 1, \ldots, n_i,$$

where t_{ij} is the time covariate of days after the gas injection for individual i, and x_{ij} is the covariate of gas concentration levels equal to -1, 0 or 1, corresponding to the concentration levels of 15%, 20% and 25%, respectively. For each individual i, $i = 1, \ldots, m$, let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$ and let $\mathbf{X}_i$ be the matrix of dimension $n_i \times 4$ whose jth row is $(1, \log(t_{ij}, \log^2(t_{ij}), x_{ij})$. Let then $\mathbf{y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_m^T)^T$ be the vector obtained by stacking the vectors $\mathbf{Y}_i$ together and let $\mathbf{X}_{\text{D}}$ be the $n \times 4$ matrix given by $\mathbf{X} = [\mathbf{X}_1^T \mathbf{X}_2^T \cdots \mathbf{X}_m^T]^T$ where $n = \sum_{i=1}^m n_i$. The mean predictor is then given by $\mathbf{X}\boldsymbol{\beta}$, with $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2, \beta_3)$. To address the heterogeneity of proportions in the two covariates of time and concentration level we use a heterogeneous simplex dispersion model with the following dispersion structure

$$\log(\sigma_{ij}^2) = \gamma_0 + \gamma_1 \log(t_{ij}) + \gamma_2 \log^2(t_{ij}) + \gamma_3 x_{ij}, , \quad i = 1, \ldots, m;\ j = 1, \ldots, n_i,$$

which leads to a dispersion predictor of the form $\mathbf{X}\boldsymbol{\gamma}$, with $\boldsymbol{\gamma}^T = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)$.

We ran our model selection and estimation algorithm `BREG` and found the selected coefficients and their estimates as listed in Table 9.

In Figure 3 we present the estimated standardised Pearson residuals $\widehat{\epsilon}_{ij} := (Y_{ij} - \widehat{\mu}_{ij})/\widehat{\sigma}_{ij}$ against the estimated mean values. The residuals seem to behave reasonably well as expected. The plot seems to be in agreement with the residual plot in panel A of Figure 3 given in the paper by [69].

For the mean structure our selection results are similar to those of [68] who found that the quadratic time term $\log^2(t_{ij})$ is significant and that the linear time $\log(t_{ij})$ is not significant. However we differ in that the gas concentration covariate is found in [68] to be marginally unsignificant, at the significance level

TABLE 9
Ophthalmological data. Estimates for the selected coefficients in the heterogeneous simplex dispersion model.

parameters and estimations	
β_0	2.7003
β_2	−0.3171
β_3	0.4057
γ_0	6.1546
γ_1	−0.4580
γ_3	−0.4932

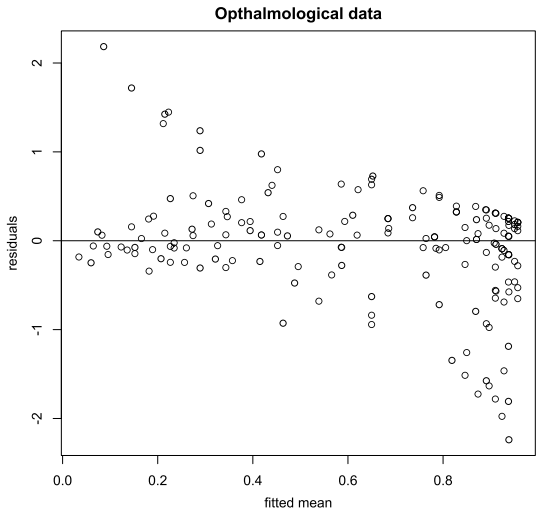

FIG 3. *Scatterplot of scaled residuals versus the fitted mean values in the selected simplex regression model for the Ophthalmological data.*

0.05, while it is selected by our procedure. Notice also that our retained coefficient estimations slightly differ which may be due to the fact that they are using an $AR(1)$ dependence in their longitudinal structure, which is possible due to the fact that their method is based on generalized estimating equations.

There is a better agreement for the dispersion coefficients: the selected coefficients not only coincide but the estimated values are also close, when compared to the results reported in [69].

7.2. Diabetes data

In this section we illustrate our procedure on a real data example modelled with a heteroscedastic Gaussian regression model. We consider the diabetes

TABLE 10
Estimates for the selected coefficients in the heteroscedastic Gaussian dispersion model for the diabetes data.

mean structure parameters and estimation		dispersion structure parameters and estimation	
β_0	151.660	γ_0	3.940
β_1	0.000	γ_1	0.000
β_2	−241.870	γ_2	−2.400
β_3	471.530	γ_3	0.000
β_4	346.780	γ_4	0.000
β_5	0.000	γ_5	0.000
β_6	0.000	γ_6	0.000
β_7	0.000	γ_7	0.000
β_8	0.000	γ_8	0.000
β_9	640.840	γ_9	0.000
β_{10}	0.000	γ_{10}	0.000

data reported on in [27]. The data consist of a response variable Y which is a quantitative measure of diabetes progression one year after baseline and of ten covariates (age, sex, body mass index, average blood pressure and six blood serum measurements). This results in $p = 11$ covariates (including an intercept) while there are $n = 442$ observations. Furthermore we assume that the data are Gaussian, independent with mean μ_i, $i = 1, \ldots, n$ and variance ϕ_i, $i = 1, \ldots, n$, modelled as $\mu_i = \beta_0 + \sum_{j=1}^{10} X_{ij}\beta_j$ and $\log(\phi_i) = \gamma_0 + \sum_{j=1}^{10} X_{ij}\gamma_j$.

We applied our `BREG` procedure for analyzing these data. The unknown tuning parameters are chosen, as before, using the BIC principle. The estimated regression coefficients are presented in Table 10.

From Table 10, we see that our procedure identified five nonzero regression coefficients in the mean model. The covariates 1, 5, 6, 7, 8 and 10 have no impact on the mean of the concentration Y. Further, only the covariate 2 affects the variance.

In Figure 4 we present the rescaled residuals $\widehat{\epsilon}_i := (Y_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})/\sqrt{\widehat{\phi}_i}$ against the estimated mean values. They look homoscedastic and similar to those displayed in Figure 2 of the paper by [70] who applied an adaptive LASSO procedure on the same data.

8. Further extensions and discussion

Although in the paper we restrict to a parametric setting and focus on continuous distributions, the methodology presented can be extended to more flexible settings, which could be needed when dealing with more complex data. In Sections 8.1, 8.2 and 8.3 we briefly describe some extensions. A detailed study of these are part of future research. Subsection 8.4 contains some further discussions on the proposed methodology.

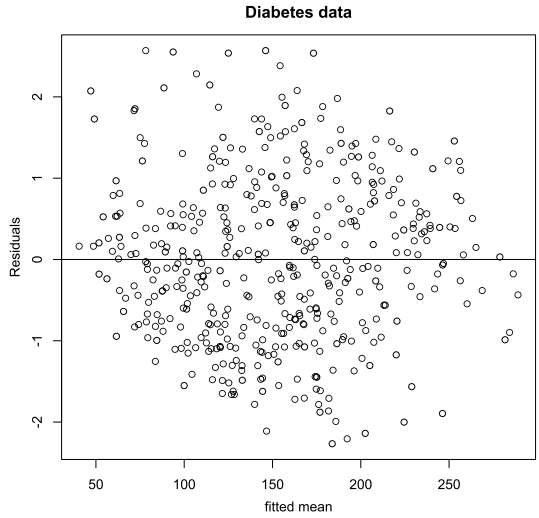

FIG 4. *Scatterplot of scaled residuals versus the fitted mean values in the selected heteroscedastic Gaussian model for the Diabetes data set.*

8.1. Nonparametric and semiparametric inference

Suppose that we are in a univariate setting, with a response variable Y and one covariate X, but are no longer assuming a parametric linear structure as in (2.8) and (2.10). In a nonparametric flexible modelling setting the mean predictor function $\eta(x)$ will be assumed to be well approximated by a linear combination of m_μ given basis functions $\{B_{\mu,1}(\cdot), \ldots, B_{\mu,m_\mu}(\cdot)\}$:

$$\eta(x) \approx \sum_{\ell=1}^{m_\mu} \alpha_{\mu,\ell} B_{\mu,\ell}(x). \tag{8.1}$$

The unknown function $\eta(\cdot)$ does not need to belong to the space spanned by the basis functions, but for m_μ large enough the function should be well approximated by the linear combination above. A common setting to deal with this modelling bias is to take a B-splines basis with a large enough number of knots (i.e. dimension m_μ). Often the number of knots grows with the sample size n, and possible overfitting is dealt with by introducing a penalty function that watches over the regularity of the estimated function (avoiding the extreme case of interpolation of the data). Estimation of $\eta(\cdot)$ then translates into estimation of the vector of unknown coefficients $\boldsymbol{\alpha}_\mu = \left(\alpha_{\mu,1}, \ldots, \alpha_{\mu,m_\mu}\right)^T$.

Similarly, in a nonparametric setting the unknown dispersion function $\gamma(\cdot)$ in (2.10) is approximated by, a possible other set of (m_γ in number) basis functions, denoted by $\{B_{\gamma,1}(\cdot), \ldots, B_{\gamma,m_\gamma}(\cdot)\}$:

$$\gamma(x) \approx \sum_{\ell=1}^{m_\gamma} \alpha_{\gamma,\ell} B_{\gamma,\ell}(x), \tag{8.2}$$

and the estimation task with respect to the dispersion function is equivalent to estimating the vector of unknown coefficients $\boldsymbol{\alpha}_\gamma = \left(\alpha_{\gamma,1}, \ldots, \alpha_{\gamma,m_\gamma}\right)^T$.

One can then proceed similarly as in Sections 3–5, but now for the vectors of coefficients $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_\gamma$.

8.2. Generalized Additive Dispersion Models (GADM)

Passing to a multivariate setting as in Section 3, with covariates $\mathbf{X} = (X_1, \ldots, X_p)^T$, and $\mathbf{Z} = (Z_1, \ldots, Z_q)^T$, an approach towards a fully nonparametric setting is to model the multivariate mean and dispersion functions as

$$\eta(\mathbf{x}) = g(\mu(\mathbf{x})) = \eta_1(x_1) + \ldots + \eta_p(x_p) \tag{8.3}$$

$$\gamma(\mathbf{z}) = h(\phi(\mathbf{z})) = \gamma_1(z_1) + \ldots + \gamma_q(z_q). \tag{8.4}$$

Each of the $p+q$ univariate functions (η_ℓ and γ_ℓ) would then be approximated as in (8.1) and (8.2):

$$\eta_j(x) \approx \sum_{\ell=1}^{m_\mu^{(j)}} \alpha_{\mu,\ell}^{(j)} B_{\mu,\ell}^{(j)}(x) \qquad \text{for } j = 1, \ldots, p$$

$$\gamma_j(x) \approx \sum_{\ell=1}^{m_\gamma^{(j)}} \alpha_{\gamma,\ell}^{(j)} B_{\gamma,\ell}^{(j)}(x) \qquad \text{for } j = 1, \ldots, q.$$

The parameters to be estimated are now

$$\begin{aligned}
\boldsymbol{\alpha}_\mu &= \left(\left(\alpha_{\mu,1}^{(1)}, \ldots, \alpha_{\mu,m_\mu^{(1)}}\right)^T, \ldots, \left(\alpha_{\mu,1}^{(p)}, \ldots, \alpha_{\mu,m_\mu^{(p)}}\right)^T \right)^T \\
\boldsymbol{\alpha}_\gamma &= \left(\left(\alpha_{\gamma,1}^{(1)}, \ldots, \alpha_{\gamma,m_\gamma^{(1)}}\right)^T, \ldots, \left(\alpha_{\gamma,1}^{(p)}, \ldots, \alpha_{\gamma,m_\gamma^{(q)}}\right)^T \right)^T .
\end{aligned} \tag{8.5}$$

By stacking all involved known basis functions in a big design matrix, the problem to be studied can be written equivalently as in the linear setting of Section 3. For the variable selection task, it is important to mention that the inclusion or not of a covariate, in the mean and/or in the dispersion model, is to be brought back to a vector of coefficients, instead of to a real-valued parameter. This then leads to utilization of grouped penalization methods. Then approaches similar in spirit to these in, for example, [6] and [5] are to be followed. Therefore, when analyzing such Generalized Additive Dispersion Models (GA*D*M), the penalized extended generalized quasi-likelihood function to be minimized is similar to expression (3.3) but with a group structure for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (or $\boldsymbol{\alpha}_\mu$ and $\boldsymbol{\alpha}_\gamma$ in the setting of this section), following the structure of subsequences in (8.5), and with absolute values in the penalties $p_{\lambda_{1j}}$, for $j = 0, \ldots, p$, and $p_{\lambda_{2j}}$, for $j = 0, \ldots, q$ (respectively), replaced by the L_2 norm of the group coefficients (for example, for the jth covariate in the mean part, the group

coefficients $\left(\alpha_{\mu,1}^{(j)},\ldots,\alpha_{\mu,m_\mu^{(j)}}\right)^T$). The penalty per group is indexed by a regularization parameter λ_{1j} (j is the group number) for the mean part and λ_{2j} for the dispersion part. The framework and algorithms used in our paper can thus easily be extended to include this setup.

The minimization problem defining the proximal operator in the scalar case is now replaced by its vector version

$$\text{argmin}_{\boldsymbol{\alpha}}(\|\boldsymbol{\alpha}-\mathbf{w}\|^2/2)+p_\lambda(s\|\boldsymbol{\alpha}\|),$$

with $\mathbf{w}$ a given vector of the same dimension as $\boldsymbol{\alpha}$. This extends the definition of the proximal operator to vector-valued arguments as in [10]. The resulting multivariate proximal (SCAD thresholding) operator is the solution to the single-group group LASSO, just as the univariate SCAD-thresholding operator is the solution to the single-variable SCAD problem considered before. This leads to a similar algorithm with multivariate thresholding replacing univariate thresholding.

Note that the methodology described briefly here could even be applied in a context of the so-called Generalized Additive Models for Location Scale and Shape (GAMLSS), of [62]. This would, as a byproduct, provide an approach towards variable selection in such models.

Finally, a semiparametric approach would consist of modelling some of the covariates, in the mean and/or the dispersion, via a parametric function and others via an additive part. The same methodology continues to hold. See for example [32].

8.3. Discrete distributions

Among the distributions that lead to proper dispersion models are also the Binomial distribution, the Poisson distribution, and the Tweedie compound Poisson distributions, among others. Compound Poisson distributions are of special interest to actuarial sciences and insurance, specifically because there is a nonzero probability mass at zero. Although the methodology described in the paper can also in theory deal with these cases, specific issues such as dealing with the exact zero values requires a special treatment. Typically one adds a small positive number to the zero value (see e.g. [57] and [24, 25]). The normalizing constant in the compound Poisson distribution also has a non attractable form (an infinite sum). See also [81].

8.4. Further discussions

In this subsection we provide further discussions on some issues related to our proposed methodology.

8.4.1. Misspecified or unknown link function for the dispersion

In the statistical models discussed in this paper two link functions are appearing: the link function $g(\cdot)$ that links the mean function $\mu(\cdot)$ to the function that is

assumed to be linear in the covariates, and the link function $h(\cdot)$ that links the scale function $\phi(\cdot)$ to obtain the target function $\gamma(\cdot)$. The link function for the mean is often dictated by the nature of the response variable (for example, logit link for binary response data, probit link function for multinomial data, and logarithmic link for count data).

The link function for dispersion, denoted by h, is generally a known monotonic differentiable function. The log link $h(v) = \log(v)$ is often used following what was done for the dispersion for double generalized linear models (see Section 2.2). In our simulation study and in both real data applications this is also the link function that we used for the dispersion. Possible alternatives for the link function for the dispersion are the power link functions $h(v) = v^p$ where p is assumed to be known (see `dlink` in `dglm`).

A first interesting question is what happens when the link function h for the scale parameter ϕ is misspecified. Using a misspecified link function for dispersion will probably lead to a loss in estimation efficiency for the mean regression parameters but the question on asymptotic consistency for the mean parameter β is legitimate. One may follow an approach similar to Fahrmeir's maximum likelihood estimation in misspecified generalized linear models (see [28]) or the one used by [75], to define the corresponding generalized quasi-likelihood score function (2.24) with misspecified variance function obtained by replacing the unknown dispersion function with a known function. A proper analysis of the effect of a misspecified link for the dispersion on the estimates of the regression parameters is an interesting topic for future research. However, we believe that under mild regularity conditions on the regularity and uniform boundedness of the resulting dispersion ϕ, and given that the asymptotic independence of the mean and dispersion parameters still remains under such a misspecification, the existence and the asymptotic consistency of the maximum penalized likelihood or quasi-likelihood estimator of $\boldsymbol{\beta}$ are still true. To illustrate this we have conducted a limited simulation in the Gaussian case (see Section 6.1) comparing the performance of the `BREG` estimator based on the correct log-link dispersion that was used for simulating the data to a `BREG` estimator based on a misspecified square-root like link (see [37]) (called `BREGl`). A summary of the simulation results, over the 150 independent runs, is presented in Table 11. We have used the same random seed as the one used for the simulation results reported in Table 4, so the results for the `BREG` estimator are the same. As one can see from Table 11 and Figures 5 and 6 (left-most and right-most boxplots) the results regarding the mean parameter (in terms of estimation as well as variable selection) are still very good. The biased estimation in the variance parameters might however lead to incorrect inference about the mean in small to moderate samples.

A second follow-up question with respect to the link function for the dispersion, is what to do in case of unknown link function h. Indeed, as correctly noted by a referee, in some situations, complete specification of the link function for the dispersion ϕ function may not be realistic. Inspired by similar approaches for estimating nonparametrically an unknown smooth variance function in heteroscedastic regression (see e.g. [12], [56], [76]), a possible way to deal with this

TABLE 11
Heteroscedastic Gaussian regression model. Estimation and variable selection ability, according to the criteria in Table 3. True (or aimed at) values are indicated in the top row.

method	S $\binom{10}{10}$	FP $\binom{0}{0}$	FN $\binom{0}{0}$	ME (as small as possible)
`BREG`				
β	9.3	0.03	0.73	2.08 (2.5)
γ	9.62	1.81	2.18	2.77 (.95)
`BREGc`				
β	6.93	0.05	3.12	13.87 (9.04)
γ	6.23	1.84	5.61	4.65 (0.85)
`BREGl`				
β	10.02	0.02	0	0.28 (0.14)
γ	6.62	3.45	6.83	5.17 (0.27)

problem would be to fit the data with a possible misspecified dispersion link function, and use the resulting deviance residuals for the mean submodel to estimate nonparametrically under positivity and monotonicity constraints the unknown dispersion function (see e.g. [51]). This then may serve as a guide for adopting an appropriate parametric dispersion link function. Pursuing such an approach is outside the scope of this paper.

8.4.2. Correlated covariates in the vectors $\mathbf{X}$ *and/or* $\mathbf{Z}$

Note that in Section 3 the covariate vectors $\mathbf{X}$ and $\mathbf{Z}$ may have overlap, or may even completely coïncide (for example all possible covariates are included in both vectors). The conditions for estimation consistency and for variable selection consistency of our algorithms depend only on the full ranks of the design matrices $\mathbf{X}_{\mathrm{D}}$ and $\mathbf{Z}_{\mathrm{D}}$, but their degree of correlation conditions the global convexity of the penalized Bregman divergence and may affect the convergence of our algorithm and the identification of active and non active components through the penalty cutoff. Note for example Condition 5 in the Appendix. To illustrate the possible impact of a correlation structure between the covariates in the vectors $\mathbf{X}$ and/or $\mathbf{Z}$ we conducted, following the suggestion of a referee, a limited simulation where the covariates of $\mathbf{X}$ and of $\mathbf{Z}$ are correlated with an AR(1) correlation structure with correlation coëfficient $\rho = 0.5$. We compared the results, referred to as `BREGc`, to those obtained for `BREG` when there is no correlation between any of the entries of $\mathbf{X}$ and of $\mathbf{Z}$. The results are displayed in Table 11 and Figures 5 and 6 (see the left-most and the middle boxplots). As expected, since the penalty parameter is connected to the degree of correlation between the covariates there is a degradation in the number of selected components especially in the variance part which is the more sensitive of the two.

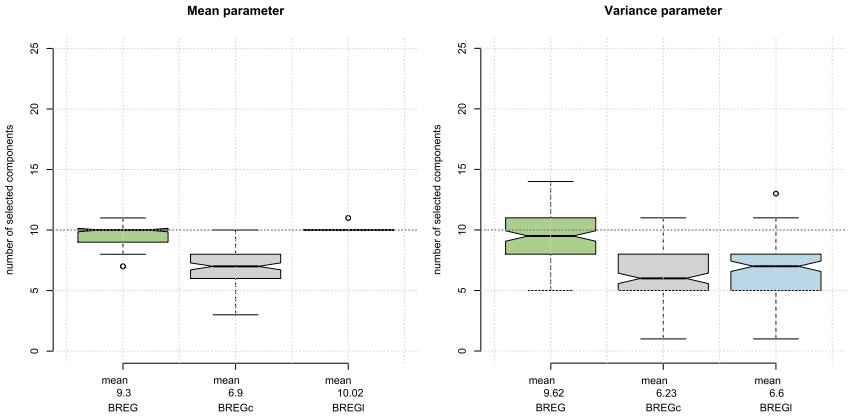

FIG 5. *Boxplots concerning the number of variables selected for the mean (left panel) and for the variance (right panel), for the* `BREG` *estimator with the correct link function (boxplot on the left) and an incorrect power-link function, denoted by the estimator* `BREGl` *(boxplot on the right) for the dispersion. True values are indicated by the dashed horizontal lines.*

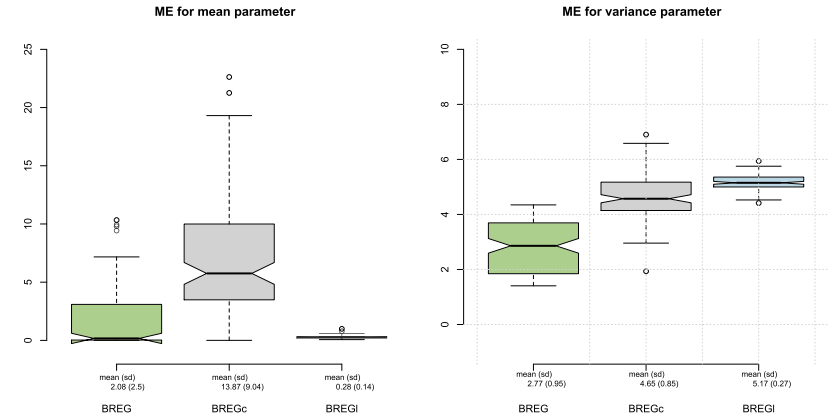

FIG 6. *Boxplots reporting on the estimation error (ME) for the mean (left panel) and for the variance (right panel), for the* `BREG` *estimator with the correct link function (boxplot on the left) and with an incorrect power-link function, denoted by the estimator* `BREGl` *(boxplot on the right) for the dispersion.*

8.4.3. Discussion on alternative penalty functions

In theory it is possible to use different penalty functions. Note that the conditions in Theorems 4.1 and 4.2 allow for general penalty functions, as long as they satisfy the conditions stated in the theorems. In this work we favor the SCAD penalty for variable selection to filter non-relevant predictors both in the mean and in the dispersion. This penalty satisfies the theoretical conditions and moreover we manage to deal with two important implementation issues: (i)

the choices of parameters in the definition of the penalty function; and (ii) the occurence of multiplicity of local minimizers of the penalized loss.

As suggested by a referee, it would be interesting to investigate the use of the minimax concave penalty (MCP) introduced by [78] for variable selection for both mean and dispersion. However there are several important details that must be worked out before implementing MCP. It is true that, with a shape parameter appropriately chosen for the MCP penalty, condition A8 of [80] remains satisfied, and together with the regularity conditions stated in the Appendix this ensures that penalized Bregman divergence with an MCP penalty is a viable variable selection procedure for which Theorems 4.1 and 4.2 hold. However, the issue of multiplicity of local minimizers of the penalized loss is more involved that this issue commented on for a SCAD penalty in Remark 4.2. To guarantee that a local linear approximation-like algorithm converges to a global minimum is more involved with MCP and requires some extra conditions on its shape parameter that depend on the design matrices of both the mean and the dispersion. Moreover, it is known that in the univariate case, MCP is equivalent to the firm-thresholding rule of [31] and therefore suffers from the drawback in that it requires two threshold values (one for 'keep' or 'shrink' and another for 'shrink' or 'kill'), thus making the data-driven procedure for the selection of regularization parameters more computationally expensive, especially when used twice, for the mean and for the dispersion. If a way is found to overcome this drawback, the use of MCP will be feasible since the MCP penalty leads also to a proximal map that has a closed-form expression (see [34]) and that could be easily used in the iterative algorithm exposed in Section 5. Implementation of penalized Bregman divergence procedures using the MCP is thus of interest, but requires additional future research work.

Appendix: Regularity conditions

We first introduce some extra notation. Let $q_1(\cdot)$ be the generating concave function of the Bregman divergence associated to the generalized quasi-likelihood score associated to pseudo response vector $\widetilde{\mathbf{Y}}$ described in Remark 3.1 and let $q_2(\cdot)$ the generating concave function associated to the log-likelihood of the regular exponential model associated to the distribution of the unit deviance when the mean vector of the proper dispersion model is known. We impose some technical regularity conditions which may not be the weakest possible but which allow us to derive the desired results.

Conditions

1. The true value $\boldsymbol{\theta}_0$ is in the interior of the parameter space and $\sup_{s_n} \|\boldsymbol{\theta}_0\|_1 < \infty$.
2. The random vectors $\mathbf{X}$ and $\mathbf{Z}$ are uniformly bounded and the corresponding design densities are bounded below by a positive constant.
3. $\mathbb{E}(\mathbf{X}_{\mathrm{D}}^T \mathbf{X}_{\mathrm{D}})$ and $\mathbb{E}(\mathbf{Z}_{\mathrm{D}}^T \mathbf{Z}_{\mathrm{D}})$ exist and are non singular.

4. The link functions g and h are continuously bi-differentiable, and such that $g^{(1)}(\cdot) \neq 0$ and $h^{(1)}(\cdot) \neq 0$.
5. The eigenvalues of the matrices $-\mathbb{E}(q_1^{(2)}(\mu(\mathbf{X}))/[g^{(1)}(\mu(\mathbf{X}))]^2\mathbf{X}_{\text{D}}^T\mathbf{X}_{\text{D}})$ and $-\mathbb{E}(q_2^{(2)}(\phi(\mathbf{Z}))/[h^{(1)}(\phi(\mathbf{Z}))]^2\mathbf{Z}_{\text{D}}^T\mathbf{Z}_{\text{D}})$ are uniformly bounded away from 0.
6. There exists some $\delta > 0$ such that $\mathbb{E}(|Y|^{\delta+2}|(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}, \mathbf{z}))$ is uniformly bounded.

Acknowledgements

The authors thank the Editor, an Associate Editor and two referees for their valuable comments which led to a considerable improvement of the presentation. This research was supported by the Interuniversity Attraction Poles Programme (IAP-network P7/06), Belgian Science Policy Office. The second author also gratefully acknowledges financial support by the GOA/12/014-project of the Research Fund KU Leuven and the FWO-project G.0B26.15N of the Flemish Science Foundation.

References

[1] A. Antoniadis. (2010). Comments on "ℓ_1-penalization for mixture regression models", by N. Städler, P. Bühlmann, S. van de Geer, *TEST*, **19** (2010): 209–256. Comments: *TEST*, **19** (2010): 257–258. MR2677723

[2] A. Antoniadis and J. Fan. (2001). Regularization of Wavelet Approximations. *Journal of the American Statistical Association*, **96** (455): 939–967. MR1946364

[3] A. Antoniadis. (2007). Wavelet methods in Statistics: some recent developments and their applications. *Statistics Surveys*, **1**: 16–55. MR2520413

[4] A. Antoniadis, I. Gijbels and M. Nikolova. (2011). Penalized likelihood regression for generalized linear models with nonquadratic penalties. *The Annals of the Institute of Statistical Mathematics*, **63** (3): 585–615. MR2786949

[5] A. Antoniadis, I. Gijbels and S. Lambert-Lacroix. (2014). Penalized estimation in additive varying coefficient models using grouped regularization. *Statistical Papers*, **55** (3): 727–750. MR3227549

[6] A. Antoniadis, I. Gijbels and A. Verhasselt. (2012). Variable selection in additive models using P-splines. *Technometrics*, **54** (4): 425–438. MR3006391

[7] R. Artes and B. Jørgensen. (2000). Longitudinal data estimating equations for dispersion models. *Scandinavian Journal of Statistics*, **27**: 321–334. MR1777507

[8] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. (2005). Clustering with Bregman divergences. *Journal of Machine Learning*, **6**: 1705–1749. MR2249870

[9] A. Belloni, V. Chernozhukov, and L. Wang. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, **98** (4): 791–806. MR2860324

[10] K. Bredies, D. A. Lorenz and St. Reiterer. (2015). Minimization of non-smooth, non-convex functionals by iterative thresholding. *Journal of Optimization Theory and Applications*, **165**(1):78–112. MR3327417
[11] L. Breiman. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, **24** (6): 2350–2383. MR1425957
[12] R.J. Carroll (1982). Adapting for heteroscedasticity in linear models. *The Annals of Statistics*, **10**: 1224–1233. MR0673657
[13] C. Charalambous, J. Pan and M. Tranmer. (2014). Variable Selection in Joint Mean and Dispersion Models via Double Penalized Likelihood. *Sankhya B*, **76** (2): 276–304. MR3302274
[14] C. Charalambous, J. Pan and M. Tranmer. (2015). Variable selection in joint modelling of the mean and variance for hierarchical data. *Statistical Modelling*, **15**: 24–50. MR3306576
[15] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. (2009). *Nonnegative Matrix and Tensor Factorization*. Wiley.
[16] R. Cottet, R.J. Kohn and D.J. Nott. (2008). Variable Selection and Model Averaging in Semiparametric Overdispersed Generalized Linear Models. *Journal of the American Statistical Association*, **103** (482): 661–671. MR2524000
[17] D.R. Cox and N. Reid. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B*, **49** (1): 1–39. MR0893334
[18] D.R. Cox and N. Reid. (1989). On the stability of maximum-likelihood estimators of orthogonal parameters. *The Canadian Journal of Statistics*, **17** (2): 229–233. MR1033105
[19] C. Croux, I. Gijbels and I. Prosdocimi. (2012). Robust estimation of mean and dispersion functions in extended generalized additive models. *Biometrics*, **68**: 31–44. MR2909851
[20] I. Daubechies, M. Defrise, and C. De Mo. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications in Pure and Applied Mathematics*, **57** (11):1413–1457. MR2077704
[21] M. Davidian and R.J. Carroll. (1987). Variance function estimation. *Journal of the American Statistical Association*, **82** (400): 1079–1091. MR0922172
[22] M. Davidian and R.J. Carroll. (1988). A note on extended quasi-likelihood. *Journal of the Royal Statistical Society, Series B*, **50**:74–82. MR0954733
[23] Z. J. Daye, J. Chen and H. Li. (2012). High-Dimensional Heteroscedastic Regression with an Application to eQTL Data Analysis. *Biometrics*, **68**:316–326. MR2909888
[24] P.K. Dunn and G.K. Smyth. (2005). Series evaluation of Tweedie exponential dispersion meodels densities. *Statistics and Computing*, **15**: 267–280. MR2205390
[25] P.K. Dunn and G.K. Smyth. (2007). Evaluation of Tweedie exponential dispersion meodels densities by Fourier inversion. *Statistics and Computing*, **18**: 73–86. MR2416440
[26] B. Efron. (1986). Double Exponential Families and their Use in General-

ized Linear Regression. *Journal of the American Statistical Association*, **81**: 809–721. MR0860505

[27] B. Efron, T. Hastie and R. Tibshirani. (2004). Least Angle Regression (with discussion). *The Annals of Statistics*, **32**: 407–451. MR2060166

[28] L. Fahrmeir (1990). Maximum likelihood estimation in misspecified generalized linear models. *Statistics*, **21**: 487–502. MR1087280

[29] J. Fan and R. Li. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, **96** (456): 1348–1360. MR1946581

[30] J. Fan and H. Peng. (2004). Nonconcave Penalized Likelihood with A Diverging Number of Parameters. *The Annals of Statistics*, **32** (3): 928–961. MR2065194

[31] H.Y. Gao and A.G. Bruce. (1997). WaveShrink with Firm Shrinkage. *Statistica Sinica*, **7**: 855–874. MR1488646

[32] I. Gijbels and I. Prosdocimi. (2012). Flexible mean and dispersion function estimation in extended Generalized Additive Models. *Communications in Statistics – Theory and Methods*, Special Issue on *Statistics for Complex Problems: Permutation Testing Methods and Related Topics*, **41** (16 & 17): 3259–3277. MR3003866

[33] I. Gijbels, I. Prosdocimi and G. Claeskens. (2010). Nonparametric estimation of mean and dispersion functions in extended generalized linear models. *Test*, **19** (3): 580–608. MR2746003

[34] P.H. Gong, C.S. Zhang, Z.C. Zhao, J.Z. Huang and J.P. Ye (2013). A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *Proceedings of 30th International Conference on Machine Learning*, Atlanta, Georgia, USA. http://arxiv.org/abs/1303.4434

[35] P. J. Green. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society, Series B*, **46**: 149–192. MR0781879

[36] W. Hare and C. Sagastizábal. (2009). Computing proximal points of non-convex functions. *Mathematical Programming, Series B*, **116** (1):221–258. MR2421280

[37] J. Jia, K. Rohe, and B. Yu. (2013). The lasso under Poisson-like heteroscedasticity. *Statistica Sinica*, **23**: 99–118. MR3076160

[38] D. Jiang. (2012). *Concave selection in generalized linear models.* Doctoral dissertation, University of Iowa. http://ir.uiowa.edu/etd/2902

[39] B. Johnson, D.Y. Lin, and D. Zeng. (2008). Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models. *Journal of the American Statistical Association*, **103**: 672–680. MR2435469

[40] B. Jørgensen. (1987). Exponential Dispersion Models. *Journal of Royal Statistical Society, Series B*, **49**: 127–162. MR0905186

[41] B. Jørgensen. (1992). Exponential dispersion models and extensions: a review. *International Statistical Review*, **60** (1): 5–20.

[42] B. Jørgensen. (1997). The theory of dispersion models. New York, Chapman & Hall. MR1462891

[43] B. Jørgensen. (2014). Dispersion Models. In *International Encyclopedia of Statistical Science*, pp 392–397.
[44] B. Jørgensen and S.J. Knudsen. (2004). Parameter orthogonality and bias adjustement for estimating functions. *Scandinavian Journal of Statistics*, **31**:93–114. MR2042601
[45] M. Kolar and J. Sharpnack. (2012). Variance function estimation in high-dimensions. In *Proceedings of the 29 th International Conference on Machine Learning, Edinburgh, Scotland, UK*. Editors, J. Langford and J. Pineau, 1447–1454.
[46] K. Lange, D.R. Hunter and I. Yang. (2000). Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics* **9**:1–59. MR1819865
[47] Y. Lee and J.A. Nelder. (2000). The relationship between double-exponential familes and extended quasi-likelihood familes, with application to modelling Geissler's human sex ratio. *Applied Statistics*, **49** (3): 413–419. MR1824549
[48] Z. Li, S. Wang and X. Lin. (2012). Variable selection and estimation in generalized linear models with the seamless penalty. *The Canadian Journal of Statistics*, **40**: 745–769. MR2998860
[49] X. Li, T. Zhao, X. Yuan and H. Liu. (2012). An R Package `flare` for High Dimensional Linear Regression and Precision Matrix Estimation. *R Package Vignette*.
[50] L. Lin. (2004). Generalized quasi-likelihood. *Statistical Papers*, **45**: 529–544. MR2088121
[51] E. Mammen (1991). Estimating a smooth monotone regression function. *The Annals of Statistics*, **19** (2):724–740. MR1105841
[52] B.D. Marx and P.H.C. Eilers. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, **28**: 193–209.
[53] P. McCullagh and J.A. Nelder. (1989). *Generalized Linear Models*. Chapman and Hall: London. MR3223057
[54] N. Meinshausen. (2007). Relaxed lasso. *Computational Statistics and Data Analysis*, **52**: 374–393. MR2409990
[55] S.M. Meyers, J.S. Ambler, M. Tan, J.C. Werner and S.S. Huang. (1992). Variation of perfluorpropane disappearance after vitrectomy. *Retina*, **12**, 359–363.
[56] H.-G. Müller and U. Stadtmüller (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics*, **15**: 610–625. MR0888429
[57] J.A. Nelder and D. Pregibon. (1987). An extended quasi likelihood function. *Biometrika*, **74**:221–232. MR0903123
[58] J.A. Nelder and R.W.M. Wedderburn. (1972). Generalized Linear Models. *Journal of Royal Statistical Society, Series A*, **135**: 370–384.
[59] M. Park and T. Hastie. (2007). An L1 regularization-path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B*, **69**: 659–677. MR2370074
[60] P. Radchenko and G.M. James. (2011). Improved variable selection with

forward-Lasso adaptive shrinkage. *The Annals of Applied Statistics*, **5** (1): 427–448. MR2810404

[61] R. Ramlau and G. Teschke. (2006). A projection iteration for nonlinear operator equations with sparsity con- straints. *Numerische Mathematik*, **104**: 177–203. MR2242613

[62] R. A. Rigby and D.M. Stasinopoulos. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, **54**: 507–554. MR2137253

[63] Y. She, J. Wang , H. Li and D. Wu. (2013). Group Iterative Spectrum Thresholding for Super-Resolution Sparse Spectral Selection *IEEE Transactions on signal Processing*, **61**(24): 6371–6386. MR3148325

[64] G.K. Smyth. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society, Series B*, **51** (1): 47–60. MR0984992

[65] G. Smyth and A.P. Verbyla. (1999a). Double generalized linear models: approximate REML and diagnostics. In *Statistical Modelling: Proceedings of the 14th International Workshop on Statistical Modelling (IWSM14), Graz, Austria*. Editors, H. Friedl, A. Berghold, G. Kauermann, pages 66–88.

[66] G. Smyth and A.P. Verbyla. (1999b). Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics*, **10**, pages 695–709.

[67] Peter X.-K. Song. (2007). Dispersion models in regression analysis. *Pakistan Journal of Statistics*, **25**, 529–551. MR2654465

[68] Peter X.-K. Song and M. Tan. (2000). Marginal models for longitudinal continuous proportional data. *Biometrics*, **56**, 496–502.

[69] Peter X.-K. Song, Z. Qiu and M. Tan. (2004). Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal*, **46** (5), 540–553. MR2101142

[70] J. Wagener and H. Dette. (2012). The adaptive Lasso in high-dimensional sparse heteroscedastic models. *Mathematical Methods of Statistics*, **22** (2), 137–154. MR3071959

[71] D. Wang and Z.Z. Zhang. (2009). Variable selection in joint generalized linear models. *Journal of Applied Probability and Statistics*, **25** (3): 245–256. MR2641680

[72] R. Wedderburn. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**: 439–447. MR0375592

[73] L. Wu and H. Li. (2012). Variable selection for joint mean and dispersion models of the inverse Gaussian distribution. *Metrika*, **75** (6): 795–808. MR2956276

[74] L.-C. Wu, Z.Z. Zhang and D.-K. Xu. (2012). Variable selection for joint mean and dispersion models of the lognormal distribution. *Hacettepe Journal of Mathematics and Statistics*, **41** (2): 307–320. MR3012185

[75] T. Xia, X.-R. Wang and X.-J. Jiang (2014). Asymptotic properties of maximum quasi-likelihood estimator in quasi-likelihood nonlinear models with misspecified variance function. *Statistics*, **48** (4): 778–786. MR3234061

[76] J. Yin, Z. Geng, R. Li and H. Zhang. (2010). Nonparametric covariance model. *Statistica Sinica*, **20**: 469–479. MR2640671

[77] Y. K. Yýlmazand A. T. Cemgil. (2012). Alpha/beta divergences and Tweedie models. Technical Report arXiv:1209.4280.

[78] C.H. Zhang (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**: 894–942. MR2604701

[79] C.M. Zhang, Y. Jiang and Z. Shang. (2009). New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Canadian Journal of Statistics*, **37**: 119–139. MR2509465

[80] C. Zhang, Y. Jiang and Y. Chai. (2010). Penalized Bregman divergence for large-dimensional regression and classification. *Biometrika*, **97**: 551–566. MR2672483

[81] Y. Zhang. (2013). Likelihood-based and Bayesian methods for Tweedie compound Poisson linear mixed models. *Statistics and Computing*, **23**:743–757. MR3247830

[82] W. Zhao, R. Zhang, Y. Lv and J. Liu. (2014). Variable selection for varying dispersion beta regression model. *Journal of Applied Statistics*, **41** (1): 95–108. MR3291202