Pak. J. Statist.
2009, Vol. 25(4), 529-551

DISPERSION MODELS IN REGRESSION ANALYSIS

Peter X.-K. Song[1]

[1] Department of Biostatistics, University of Michigan, Ann Arbor, MI USA
Email: pxsong@umich.edu

ABSTRACT

This paper presents a review about the theory of regression analysis based on Jørgensen's dispersion models, which extends the classical theory of generalized linear models (GLMs). Dispersion models provide a rich class of parametric models useful to conduct regression analysis of non-normal data, which contain all the error distributions considered in the classical GLMs. One example used to illustrate the extension is the simplex GLM for continuous proportional data. Details concerning the maximum likelihood estimation are presented, accompanied with some numerical results.

KEYWORDS

Exponential dispersion model. generalized linear model. maximum likelihood estimation. simplex distribution. Tweedie class.

2000 Mathematics Subject Classification: 62J12, 62E15.

1 INTRODUCTION

This paper concerns an extended regression analysis theory of generalized linear models (GLMs) with the utility of Jørgensen's (1987; 1997) dispersion models. The dispersion models provide a rich class of one-dimensional parametric distributions for various data types, including those commonly considered in the GLM analysis. In effect, error distributions in the classical GLMs (e.g. McCullagh and Nelder, 1989) form a special subclass of the dispersion models, which are called the *exponential dispersion models*. This implies that the GLMs outlined in the current statistical literature, (e.g. McCullagh and Nelder, 1989) are special cases of the models considered in this paper. Two important models that are not included in the current GLMs but special cases of the dispersion models are the von Mises distribution for directional (circular or angular) data and the simplex distribution for compositional (or proportional) data. Some details concerning the simplex distribution will be discussed throughout this paper.

A classical GLM consists of two components: One is the so-called *random component* specified by an exponential dispersion (ED) family density of the following form:

$$p(y;\theta,\phi) = \exp\left[\frac{\{y\theta - \kappa(\theta)\}}{a(\phi)} + C(y,\phi)\right], y \in \mathcal{C}, \tag{1.1}$$

with parameters $\theta \in \Theta$ and $\phi > 0$, where $\kappa(\cdot)$ is the cumulant generating function and $\mathcal{C}$ is the support of the density. It is known that the first derivative of the cumulant function $\kappa(\cdot)$ gives the expectation of the distribution, namely $\mu = \mathrm{E}(Y) = \dot{\kappa}(\theta)$. Table 1 lists some commonly used ED distributions.

Table 1: Some commonly used exponential dispersion GLMs.

Distribution	Domain	Data type	Canonical link	Model
Normal	$(-\infty,\infty)$	Continuous	Identity	Linear model
Binomial	$\{0,1,\ldots,n\}$	Binary or counts	Logit	Logistic model
Poisson	$\{0,1,\ldots,\}$	Counts	Log	Loglinear model
Gamma	$(0,\infty)$	Positive continuous	Reciprocal	Reciprocal model

The other component of the model is the so-called *systematic component* that is assumed to take the following form:

$$g(\mu) = \mathbf{x}^T\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{1.2}$$

where g is the link function, $\mathbf{x} = (1, x_1, \ldots, x_p)^T$ is a $(p+1)$-dimensional vector of covariates, and $\beta = (\beta_0, \ldots, \beta_p)^T$ is a $(p+1)$-dimensional vector of regression coefficients. The *canonical link* function $g(\cdot)$ is specified such that $g(\mu) = \theta$, where θ is the canonical parameter.

The primary statistical tasks include estimation and inference for β. Checking model assumptions is also an important task of regression analysis. This paper focuses on the estimation and inference in an extended class of regression models. Let us begin with a brief review of Jørgensen's dispersion models.

2 DISPERSION MODELS

The normal distribution $N(\mu, \sigma^2)$ plays the central role in the classical linear regression regression. The density of $N(\mu, \sigma^2)$ is

$$p(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}, \ y \in \mathcal{R},$$

where $(y-\mu)^2$ can be regarded as a Euclidean distance that measures the discrepancy between the observed y (data) and the expected μ (model). And this discrepancy measure is used to develop many regression analysis methods, such as the F-statistic for the assessment of goodness-of-fit for nested models.

Mimicking the normal density, Jørgensen (1987) proposes a class of dispersion models (DM) by extending the Euclidean distance $(y-\mu)^2$ to a general discrepancy function $d(y; \mu)$. It is found that many commonly used parametric distributions, such as those in Table 1, are included as special cases of this extension. Moreover, each of such distributions will be determined uniquely by the discrepancy function d, and the resulting distribution is fully parameterized by two parameters μ and σ^2.

2.1 Definitions

A *(reproductive) dispersion model* $\mathrm{DM}(\mu, \sigma^2)$ with *location parameter* μ and *dispersion parameter* σ^2 is a family of distributions whose probability density functions take the following form:

$$p(y; \mu, \sigma^2) = a(y; \sigma^2) \exp\left\{-\frac{1}{2\sigma^2} d(y; \mu)\right\}, \ y \in \mathcal{C} \tag{2.1}$$

where $\mu \in \Omega$, $\sigma^2 > 0$, and $a \geq 0$ is a suitable normalizing term that is independent of the μ. This normalizing term $a(\cdot)$ is determined in such a way that $\int_{\mathcal{C}} a(y; \sigma^2) \exp\left\{-\frac{1}{2\sigma^2} d(y; \mu)\right\} dy = 1$. Usually, $\Omega \subseteq \mathcal{C} \subseteq \mathcal{R}$. The fact that the normalizing term a does not involve μ will allow us to estimate μ (or β in the regression analysis setting) separately from estimating σ^2, which gives rise to great ease in the parameter estimation. Such a nice property, known as the likelihood (or parameter) orthogonality, holds in the normal distribution, and it will remain in the dispersion models.

A bivariate function $d(\cdot; \cdot)$ is called the *unit deviance* defined on $(y, \mu) \in \mathcal{C} \times \Omega$ if it satisfies the following two properties:

(i) It is zero when the observed y and the expected μ are equal, namely

$$d(y; y) = 0, \quad \forall y \in \Omega;$$

(ii) It is positive when the observed y and the expected μ are different, namely

$$d(y;\mu) > 0, \quad \forall y \neq \mu.$$

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

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

For a regular unit deviance, the variance function is defined as follows. The *unit variance function* $V : \Omega \to (0,\infty)$ is

$$V(\mu) = \frac{2}{\frac{\partial^2 d}{\partial \mu^2}\,(y;\mu)\big|_{y=\mu}}, \ \mu \in \Omega. \tag{2.2}$$

Some popular dispersion models are given in Table 2, in which the unit deviance d and variance function V can be found easily through the definition of dispersion model. Readers may refer to Song (2007, Chapter 2) for more examples and details regarding the derivations of deviance and variance functions.

Table 2: Unit deviance and variance functions of some dispersion models.

Distribution	Deviance d	$\mathcal{C}$	Ω	$V(\mu)$
Normal	$(y-\mu)^2$	$(-\infty,\infty)$	$(-\infty,\infty)$	1
Poisson	$2(y\log\frac{y}{\mu} - y + \mu)$	$\{0,1,\ldots\}$	$(0,\infty)$	μ
Binomial	$2\left\{y\log\frac{y}{\mu} + (n-y)\log\frac{n-y}{n-\mu}\right\}$	$\{0,1,\ldots,n\}$	$(0,1)$	$\mu(1-\mu)$
Negative binomial	$2\left\{y\log\frac{y}{\mu} + (1-y)\log\frac{1-y}{1-\mu}\right\}$	$\{0,1,\ldots\}$	$(0,\infty)$	$\mu(1+\mu)$
Gamma	$2\left(\frac{y}{\mu} - \log\frac{y}{\mu} - 1\right)$	$(0,\infty)$	$(0,\infty)$	μ^2
Inverse Gaussian	$\frac{(y-\mu)^2}{y\mu^2}$	$(0,\infty)$	$(0,\infty)$	μ^3
von Mises	$2\{1-\cos(y-\mu)\}$	$(0,2\pi)$	$(0,2\pi)$	1
Simplex	$\frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}$	$(0,1)$	$(0,1)$	$\mu^3(1-\mu)^3$

2.2 Some Key Properties

Here we list some useful properties of the dispersion models without proofs. Readers may refer to Song (2007, Chapter 2) for detailed proofs.

Proposition 2.1. *If a unit deviance d is regular, then*

$$\frac{\partial^2 d}{\partial y^2}(y;y) = \frac{\partial^2 d}{\partial \mu^2}(y;y) = -\frac{\partial^2 d}{\partial \mu \partial y}(y;y), \quad \forall y \in \Omega. \tag{2.3}$$

Proposition 2.2. *Taylor expansion of a regular unit deviance d near its minimum* (μ_0, μ_0) *is given by*

$$d(\mu_0 + x\delta; \mu_0 + m\delta) = \frac{\delta^2}{V(\mu_0)}(x-m)^2 + o(\delta^2),$$

where $V(\cdot)$ *is the unit variance function.*

In some cases, the normalizing term $a(\cdot)$ has no closed form expression, which gives rise to the difficulty of estimating the dispersion parameter σ^2. Proposition 2.3 presents an approximation to the normalizing term $a(\cdot)$, based on the so-called saddlepoint approximation of the density under small dispersion. Notation $a \simeq b$ exclusively stands for an approximation of a to b when the dispersion $\sigma^2 \to 0$, in the context of small-dispersion asymptotics.

Proposition 2.3 (Saddlepoint approximation). *As the dispersion* $\sigma^2 \to 0$, *the density of a regular DM model can be approximated to be:*

$$p(y;\mu,\sigma^2) \simeq \{2\pi\sigma^2 V(y)\}^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2} d(y;\mu)\right\},$$

which equivalently says that as $\sigma^2 \to 0$, *the normalizing term has a small dispersion approximation,*

$$a(y;\sigma^2) \simeq \{2\pi\sigma^2 V(y)\}^{-1/2}, \tag{2.4}$$

with the unit variance function $V(\cdot)$.

The proof of this proposition is basically an application of the Laplace approximation given in, for example, Barndorff-Nielsen and Cox (1989, page 60). Also see Jørgensen (1997, page 28).

It follows from Propositions 2.2 and 2.3 that the small dispersion asymptotic normality holds, as stated in Proposition 2.4.

Proposition 2.4 (Asymptotic normality). *: Let $Y \sim DM(\mu_0 + \sigma\mu, \sigma^2)$ be a dispersion model with uniformly convergent saddlepoint approximation, namely convergence in (2.4) is uniformly in y. Then*

$$\frac{Y-\mu_0}{\sigma} \stackrel{d}{\to} N(\mu, V(\mu_0)), \quad \text{as } \sigma^2 \to 0.$$

In other words, $DM(\mu_0 + \sigma\mu, \sigma^2) \stackrel{d}{\simeq} N(\mu_0 + \sigma\mu, \sigma^2 V(\mu_0))$ for small dispersion σ^2.

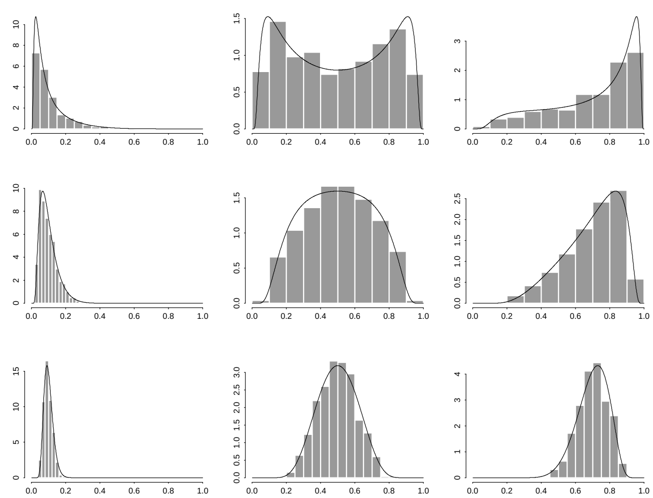

Figure 1: Several simplex density functions.

To illustrate this small-dispersion asymptotic normality, Figure 1 displays the simplex distributions simplex distribution with mean mean $\mu = (0.1, 0.5, 0.7)$ from left to right and dispersion parameter $\sigma^2 = (4^2, 2^2, 1)$ from top to bottom. The solid lines represent the simplex densities with the histograms as the background. These histograms are based on 500 simulated data from respective densities. See the detail of a simplex distribution in Table 2. This figure clearly indicates that the smaller the dispersion is, the less deviation the simplex distribution is from the normality.

2.3 Exponential Dispersion Models

The class of dispersion models contains two important subclasses, namely *the exponential dispersion (ED) models* and *the proper dispersion (PD) models*. Being the family of error distributions in the current theory of GLMs, ED models include continuous distributions such as normal, gamma, and inverse Gaussian, and discrete distributions such as Poisson, binomial, negative binomial, among others.

To establish the connection of the ED model representation (1.1) to the DM, it is sufficient to show that expression (1.1) is a special form of (2.1). An advantage with the DM type of parametrization for the ED models is that both mean μ and dispersion parameters σ^2 are explicitly present in the density, whereas expression (1.1) hides the mean μ in the first order derivative $\mu = \dot{\kappa}(\theta)$. In addition, having a density form similar to the normal enables us to easily borrow the classical normal regression theory to the development of regression analysis for nonnormal data. One example is the analogue of the likelihood ratio test in the GLMs to the F-test for goodness-of-fit in the normal regression model.

To show an ED model, denoted by $\mathrm{ED}(\mu, \sigma^2)$, as a special case of the DM, it suffices to find a unit deviance function d such that the density of the ED model can be expressed in the form of (2.1). First, denote $\lambda = 1/a(\phi)$. Then, the density in (1.1) can be rewritten as of the form:

$$p(y;\theta,\lambda) = c(y;\lambda)\exp[\lambda\{\theta y - \kappa(\theta)\}],\ y \in \mathcal{C} \tag{2.5}$$

where $c(\cdot)$ is a suitable normalizing term. Parameter $\lambda = 1/\sigma^2 \in \Lambda \subset (0,\infty)$ is called the *index parameter* and Λ is called the index set. To reparametrize this density (1.1) by the mean μ and dispersion σ^2, define the *mean value mapping:* $\tau : int(\Theta) \to \Omega$,

$$\tau(\theta) = \dot{\kappa}(\theta) \equiv \mu,$$

where $int(\Theta)$ is the interior of the parameter space Θ.

Proposition 2.5. *The mean mapping function* $\tau(\theta)$ *is strictly increasing.*

It follows that the inverse of the mean mapping function $\tau(\cdot)$ exists, denoted by $\theta = \tau^{-1}(\mu)$, $\mu \in \Omega$. Hence, the density in (2.5) can be reparametrized as follows,

$$p(y;\mu,\sigma^2) = c(y;\sigma^{-2})\exp\left[\frac{1}{\sigma^2}\{y\tau^{-1}(\mu) - \kappa(\tau^{-1}(\mu))\}\right]. \tag{2.6}$$

Proposition 2.6. *The first order derivative of* $\tau^{-1}(\mu)$ *with respect to* μ *is* $1/V(\mu)$, *where* $V(\mu) = \dot{\tau}(\tau^{-1}(\mu))$.

This proposition 2.6 shows that the $V(\mu)$ is indeed the same as the unit variance function $V(\mu)$ given by the definition (2.2). The proof of this result can be found in Song (2007, Chapter 2).

Define the unit deviance function of the ED model as follows:

$$\begin{aligned} d(y;\mu) &= 2\left[\sup_{\mu}\{f(y;\mu)\} - f(y;\mu)\right] \\ &= 2\left[\sup_{\theta\in\Theta}\{\theta y - \kappa(\theta)\} - y\tau^{-1}(\mu) + \kappa(\tau^{-1}(\mu))\right]. \end{aligned} \tag{2.7}$$

Clearly, this d function satisfies (i) $d(y;\mu) \geq 0$ for all $y \in \mathcal{C}$ and $\mu \in \Omega$, and (ii) $d(y;\mu)$ attains the minimum at $\mu = y$ because the supermum term is independent of μ. Thus, (2.7) gives a proper unit deviance function. Moreover, since it is continuously twice differentiable, it is also regular. As a result, the density of an ED model can be expressed as of the DM form:

$$p(y;\mu,\sigma^2) = a(y;\sigma^2)\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\},$$

with the unit deviance function d given in (2.7) and the normalizing term given by

$$a(y;\sigma^2) = c(y;\sigma^{-2})\exp\left[\sigma^{-2}\sup_{\theta\in\Theta}\{y\theta - \kappa(\theta)\}\right].$$

Here are two remarks for the ED models:

(1) Parameter μ is the mean of the distribution, namely $\mathrm{E}(Y) = \mu$.

(2) Variance of the distribution is

$$\mathrm{var}(Y) = \sigma^2 V(\mu). \tag{2.8}$$

This mean-variance relationship is one of the key properties for the ED models, which will play an important role in the development of quasi-likelihood inference.

An important property for the ED models is the closure under convolution operation.

Proposition 2.7 (Convolution for the ED models). *Assume $Y_1,\ldots,Y_n$ are independent and*

$$Y_i \sim ED(\mu, \frac{\sigma^2}{w_i}), i=1,\ldots,n,$$

where w_is are certain positive weights. Let $w_+ = w_1+\cdots+w_n$. Then the weighted average follows still an ED model; that is,

$$\frac{1}{w_+}\sum_{i=1}^n w_i Y_i \sim ED(\mu, \frac{\sigma^2}{w_+}).$$

In particular, with $w_i=1, i=1,\ldots,n$ the sample average

$$\frac{1}{n}\sum_{i=1}^n Y_i \sim ED(\mu, \frac{\sigma^2}{n}).$$

For the example of two *i.i.d.* Poisson random variables with $Y_i \sim \mathrm{ED}(\mu,1)$, $i=1,2$, their average $(Y_1+Y_2)/2 \sim \mathrm{ED}(\mu,\frac{1}{2})$. Note that the resulting $\mathrm{ED}(\mu,\frac{1}{2})$ is no longer a Poisson distribution but it is still an ED distribution.

It is noticeable that although the class of the ED models is closed under the convolution operation, it is in general not closed under scale transformation. That is, cY may not follow an ED model even if $Y \sim \mathrm{ED}(\mu,\sigma^2)$, for a constant c. However, a subclass of the ED models, termed as the *Tweedie class*, is closed under this type of scale transformation. The Tweedie models will be discussed in Section 2.4.

2.4 Tweedie Class

Tweedie class is an important subclass of the ED models, which is closed under the scale transformation. Tweedie models are characterized by the unit variance functions in the form of the power function:

$$V_p(\mu) = \mu^p, \mu \in \Omega_p, \tag{2.9}$$

where $p \in R$ is a *shape* parameter.

It is shown that the ED model with the power unit variance function (2.9) always exists except $0<p<1$. A Tweedie model is denoted by $Y \sim Tw_p(\mu,\sigma^2)$ with mean μ and variance

$$\mathrm{Var}(Y) = \sigma^2\mu^p.$$

The following proposition gives the characterization of the Tweedie models.

Proposition 2.8 (Tweedie Characterization). *Let $ED(\mu,\sigma^2)$ be a reproductive ED model satisfying $V(1) = 1$ and $1 \in \Omega$. If the model is closed with respect to scale transformation, such that there exists a function $f : R_+ \times \Lambda^{-1} \to \Lambda^{-1}$ for which*

$$cED(\mu,\sigma^2) \sim ED[c\mu, f(c,\sigma^2)], \forall c > 0,$$

then

(a) $ED(\mu,\sigma^2)$ *is a Tweedie model for some* $p \in R\backslash(0,1)$*;*

(b) $f(c,\sigma^2) = c^{2-p}\sigma^2$*;*

(c) the main domain $\Omega = R$ *for* $p = 0$ *and* $\Omega = (0,\infty)$ *for* $p \neq 0$*;*

(d) the model is infinitely divisible.

It follows immediately from Proposition 2.8 that

$$cTw_p(\mu,\sigma^2) = Tw_p(c\mu, c^{2-p}\sigma^2).$$

The importance of the Tweedie class is that it serves as a class of limiting distributions of the ED models, as described in the following proposition.

Definition 2.1. The unit variance function V is said to be regular of order p at 0 (or at ∞), if $V(\mu) \sim c_0\mu^p$ as $\mu \to 0$(or $\mu \to \infty$) for certain $p \in \mathcal{R}$ and $c_0 > 0$.

Proposition 2.9. *Suppose the unit variance function V is regular of order p at 0 or at ∞, with $p \notin (0,1)$. For any $\mu > 0$ and $\sigma^2 > 0$,*

$$c^{-1}ED(c\mu, c^{2-p}\sigma^2) \xrightarrow{d} TW_p(\mu, c_0\sigma^2), \text{ as } c \to 0 \text{ or } \infty,$$

where the convergence is through values of c such that $c\mu \in \Omega$ and $c^{p-2}/\sigma^2 \in \Lambda$.

Refer to Jørgensen et al. (1994) for the proof of this result.

3 MAXIMUM LIKELIHOOD ESTIMATION AND INFERENCE

This section is devoted to maximum likelihood estimation and inference in the regression analysis based on the dispersion models. Therefore, the MLE theory given in, for example McCullagh and Nelder (1989), is a special case, because the ED family is a subclass of the DM family.

3.1 General Theory

Consider a cross-sectional dataset, $(y_i, \mathbf{x}_i), i = 1, \ldots, K$, where the y_i's are *i.i.d.* realizations of Y_i's according to $\mathrm{DM}(\mu_i, \sigma^2)$ and $g(\mu_i) = \mathbf{x}_i^T \beta$. Let $\mathbf{y} = (y_1, \ldots, y_K)^T$ and $\mu = (\mu_1, \ldots, \mu_K)^T$. The likelihood for the parameter vector $\theta = (\beta, \sigma^2)$ is given by

$$L(\theta; \mathbf{y}) = \prod_{i=1}^{K} a(y_i; \sigma^2) \exp\left\{ -\frac{1}{2\sigma^2} d(y_i; \mu_i) \right\},\ \beta \in \mathcal{R}^{p+1}, \sigma^2 > 0.$$

The log-likelihood is then

$$\begin{aligned} \ell(\theta; \mathbf{y}) &= \sum_{i=1}^{K} \log a(y_i; \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{K} d(y_i; \mu_i) \\ &= \sum_{i=1}^{K} \log a(y_i; \sigma^2) - \frac{1}{2\sigma^2} D(\mathbf{y}; \mu), \end{aligned} \tag{3.1}$$

where $\mu_i = \mu_i(\beta)$ is a nonlinear function in β and $D(\mathbf{y}; \mu) = \sum_{i=1}^{K} d(y_i; \mu_i)$ is the sum of deviances depending on β only. This D is analogous to the sum of squared residuals in the linear regression model.

The score function for the regression coefficient β is

$$s(\mathbf{y}; \beta) = \frac{\partial \ell(\theta)}{\partial \beta} = -\frac{1}{2\sigma^2} \sum_{i=1}^{K} \frac{\partial d(y_i; \mu_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta}.$$

Denote the i-th linear predictor by $\eta_i = \mathbf{x}_i^T \beta$, and denote the *deviance scores* by

$$\delta(y_i; \mu_i) = -\frac{1}{2} \frac{\partial d(y_i; \mu_i)}{\partial \mu_i},\ i = 1, \ldots, K. \tag{3.2}$$

Note that

$$\frac{\partial \mu_i}{\partial \beta} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} = \{\dot{g}(\mu_i)\}^{-1} \mathbf{x}_i,$$

where $\dot{g}(\mu)$ is the first order derivative of link function g *w.r.t* μ. Table 3 lists some commonly used link functions and their derivatives.

Then the score function for β takes the form

$$s(\mathbf{y}; \beta) = \frac{1}{\sigma^2} \sum_{i=1}^{K} \mathbf{x}_i \frac{1}{\dot{g}(\mu_i)} \delta(y_i; \mu_i). \tag{3.3}$$

Moreover, the score equation leading to the maximum likelihood estimate of the β is

$$\sum_{i=1}^{K} \mathbf{x}_i \frac{1}{\dot{g}(\mu_i)} \delta(y_i; \mu_i) = 0. \tag{3.4}$$

Table 3: Some common link functions and derivatives. NB and IG stand for Negative binomial and Inverse Gaussian, respectively.

Model	Link g	Derivative $\dot{g}$	Domain Ω
Binomial or simplex	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{1}{\mu(1-\mu)}$	$\mu \in (0,1)$
Poisson, NB, gamma, or IG	$\log(\mu)$	$\frac{1}{\mu}$	$\mu \in (0,\infty)$
Gamma	$\frac{1}{\mu}$	$-\frac{1}{\mu^2}$	$\mu \in (0,\infty)$
von Mises	$\tan(\mu/2)$	$\frac{1}{2}\sec^2(\mu/2)$	$\mu \in [-\pi,\pi)$

Note that this equation does not involve the dispersion parameter σ^2. Under some mild regularity conditions, the resulting ML estimator $\widehat{\beta}_K$, which is the solution to the score equation (3.4), is consistent

$$\widehat{\beta}_K \xrightarrow{p} \beta \text{ as } K \to \infty,$$

and asymptotically normal with mean 0 and covariance matrix $\mathbf{i}^{-1}(\theta)$. Here $\mathbf{i}(\theta)$ is the Fisher information matrix given by

$$\begin{aligned} \mathbf{i}(\theta) &= -\mathrm{E}\{\dot{s}(\mathbf{Y};\beta)\} \\ &= \mathbf{X}^T U^{-1} \mathbf{X}/\sigma^2, \end{aligned} \tag{3.5}$$

where $\mathbf{X}$ is a $K \times (p+1)$ matrix with the i-th row being the $\mathbf{x}_i^T$, and U is a diagonal matrix with the i-th diagonal element u_i given by

$$u_i = \frac{\{\dot{g}(\mu_i)\}^2}{\mathrm{E}\{-\dot{\delta}(Y_i;\mu_i)\}},\ i = 1,\ldots,K. \tag{3.6}$$

The asymptotic normality provides the theoretic basis to conduct statistical inference. For example, the Wald statistic can be formed to test for the significance of a subvector of the β, and 95% confidence intervals of the regression coefficients can be constructed based on the asymptotic normal distribution of the MLE.

When the dispersion parameter σ^2 is present in the model, the ML estimation for the dispersion parameter σ^2 can be derived similarly, if the normalizing term $a(y;\sigma^2)$ is simple enough to allow such a derivation, such as the case of the normal distribution. However, in many cases, the term $a(\cdot)$ has no closed form expression and its derivative *w.r.t.* σ^2 may appear too complicated to be numerically solvable. In this case, two methods have been suggested to acquire the estimation for σ^2. The first method is to invoke the small dispersion asymptotic normality (Proposition 2.3), where subject to a constant,

$$\log a(y;\sigma^2) \simeq -\frac{1}{2}\log\sigma^2.$$

Applying this approximation in the log-likelihood (3.1) and differentiating the resulting approximate log-likelihood *w.r.t.* σ^2, one can obtain an equation as follows,

$$-\frac{K}{2\sigma^2} + \frac{1}{2\sigma^4}D(\mathbf{y};\mu) = 0.$$

Solution to this equation gives an estimator of the dispersion parameter σ^2,

$$\widehat{\sigma}^2 = \frac{1}{K}D(\mathbf{y};\widehat{\mu}) = \frac{1}{K}\sum_{i=1}^{K} d(y_i;\hat{\mu}_i). \tag{3.7}$$

Song (2007) refers this estimator to as *the Jørgensen estimator* of the dispersion parameter, which in fact is an average of the estimated unit deviances.

However, the Jørgensen estimator is not, in general, unbiased even if the adjustment on the degrees of freedom, $K-(p+1)$ is made to replace K. Moreover, this formula is recommended when the dispersion parameter σ^2 is small, say less than 5.

To obtain an unbiased estimator of the dispersion parameter σ^2, the second method utilizes a moment property given in the following proposition.

Proposition 3.1. *Let $Y \sim DM(\mu,\sigma^2)$ with a regular unit deviance $d(y;\mu)$. Then,*

$$\begin{aligned} E\{\delta(Y;\mu)\} &= 0, \\ var\{\delta(Y;\mu)\} &= \sigma^2 E\{-\dot{\delta}(Y;\mu)\}, \end{aligned}$$

where $\dot{\delta}$ is the first order derivative of the deviance score given in (3.2) w.r.t. μ.

Based on this result, Song (2007) suggested to consistently estimate the dispersion parameter σ^2 by the method of moments:

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{K}(\delta_i - \overline{\delta})^2}{\sum_{i=1}^{K}(-\dot{\delta}_i)}, \tag{3.8}$$

where $\delta_i = \delta(y_i;\hat{\mu}_i)$, $\dot{\delta}_i = \dot{\delta}(y_i;\hat{\mu}_i)$ and $\overline{\delta} = \frac{1}{K}\sum_i \delta_i$.

3.2 MLE in the ED Models

Now consider the traditional GLMs based on the ED models. For the unit deviance of the ED model given in (2.7), it is easy to see

$$\delta(y;\mu) = \frac{y-\mu}{V(\mu)}. \tag{3.9}$$

It follows that the score equation (3.4) becomes

$$\sum_{i=1}^{K} \mathbf{x}_i \frac{1}{\dot{g}(\mu_i)V(\mu_i)}(y_i-\mu_i) = 0.$$

Let $w_i = \dot{g}(\mu_i)V(\mu_i)$. Then the score equation can be re-expressed as of the form

$$\sum_{i=1}^{K} \mathbf{x}_i w_i^{-1}(y_i-\mu_i) = 0,$$

or in the matrix notation,

$$\mathbf{X}^T W^{-1}(\mathbf{y}-\mu) = 0,$$

where $W = \mathrm{diag}(w_1,\ldots,w_K)$. The following result is useful to calculate the Fisher information.

Proposition 3.2. *Suppose $Y \sim ED(\mu,\sigma^2)$. Then,*

$$E\{-\dot{\delta}(Y;\mu)\} = \frac{1}{V(\mu)},$$

where $\dot{\delta}(y;\mu)$ is the first order derivative of the deviance score $\delta(y;\mu)$ w.r.t. μ.

In the Fisher information matrix $\mathbf{i}(\theta)$ of (3.5), $\mathbf{i}(\theta) = \mathbf{X}^T U^{-1}\mathbf{X}/\sigma^2$, U is a diagonal matrix whose i-th diagonal element can be simplified as

$$u_i = \{\dot{g}(\mu_i)\}^2 V(\mu_i).$$

Furthermore, if the canonical link function $g = \tau^{-1}(\cdot)$ is chosen, then a further simplification leads to $w_i = 1$ and $u_i = 1/V(\mu_i)$ because in this case, $\dot{g}(\mu_i) = 1/V(\mu_i)$. So, the matrix W becomes the identity matrix and the matrix U is determined by the reciprocals of the variance functions.

It is interesting to note that the choice of the canonical link simplifies both score function and Fisher information. In summary, under the canonical link function, the score equation of an ED GLM is

$$\sum_{i=1}^{K} \mathbf{x}_i(y_i-\mu_i) = 0, \text{ or } \mathbf{X}^T(\mathbf{y}-\mu) = 0,$$

and the Fisher information takes the form

$$\mathbf{i}(\theta) = \mathbf{X}^T U^{-1} \mathbf{X} / \sigma^2$$

where $U = \text{diag}(u_1, \ldots, u_K)$, a diagonal matrix with variance function $V(\mu_i)$ as the i-th diagonal element.

Each ED model holds the so-called mean-variance relation, *i.e.* $\text{var}(Y) = \sigma^2 V(\mu)$, which may be used to obtain a consistent estimator of the dispersion parameter σ^2 given as follows:

$$\widehat{\sigma}^2 = \frac{1}{K-p-1} \sum_{i=1}^{K} \hat{r}_{p,i}^2 = \frac{1}{K-p-1} \sum_{i=1}^{K} \left\{ \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \right\}^2,$$

where $\hat{r}_p$ is the Pearson residual. In fact, the relation given in Proposition 3.1 is equivalent to this mean-variance relation for the ED models, simply because of Proposition 3.2.

4 AN APPLICATION: THE SIMPLEX GLM

This section supplies one non-ED GLM based on the simplex distribution for proportional data, which is not available in the classical theory of GLMs (e.g. McCullagh and Nelder, 1989).

4.1 Model Formulation

Continuous proportions or ratios are widely seen in practice as a primary response variable of interest. For example, percent decrease in glomerular filtration rate (GFR) at a follow-up time from the baseline is regarded as a measure that reflects directly the loss of renal function. In cancer studies, percent change of tumors than their actual sizes is of greater interest in disease diagnosis and monitoring. A key feature of this type of data is the confinement with a finite interval, often in the unit interval. An appropriate statistical analysis should take the finite support of the data distribution, say $(0,1)$, into account. The simplex distribution provides a parametric model useful to analyze such proportional data.

From the methodological point of view, in the ED GLMs both score equation and Fisher information can be treated as a special case of weighted least squares estimation, due to the fact that the first order derivative of the unit deviance is $(y-\mu)/V(\mu)$, which is linear in y. However, this linearity no longer holds for a DM GLM outside the class of the ED GLMs. The simplex distribution is one of such examples.

A simplex model $S^{-}(\mu;\sigma^2)$ has the density given by

$$p(y;\mu,\sigma^2) = [2\pi\sigma^2\{y(1-y)\}^3]^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\},\ y\in(0,1),\mu\in(0,1),$$

with the unit deviance function

$$d(y;\mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2},\ y\in(0,1),\mu\in(0,1),$$

where $\mu = \mathrm{E}(Y)$ is the mean. The unit variance function is $V(\mu) = \mu^3(1-\mu)^3$, obtained from (2.2). Figure 1 displays several simplex density functions.

For a non-ED GLM, the canonical link function no longer helps to simplify the weights u_i or the w_i, because the density does not explicitly involve the cumulant generating function $\kappa(\cdot)$ as in the ED GLM. For the simplex distribution, since $\mu\in(0,1)$, one may take the logit as the link function to formulate the systematic component:

$$\log\frac{\mu}{1-\mu} = \mathbf{x}^T\beta.$$

According to Table 3, $\dot{g}(\mu) = \{\mu(1-\mu)\}^{-1}$. It follows from (3.4) that the score equation for the regression parameter β is

$$\sum_{i=1}^{K}\mathbf{x}_i\{\mu_i(1-\mu_i)\}\delta(y_i;\mu_i) = 0, \tag{4.1}$$

where the deviance score is

$$\begin{aligned}\delta(y;\mu) &= -\frac{1}{2}\dot{d}(y;\mu)\\ &= \frac{y-\mu}{\mu(1-\mu)}\left\{d(y;\mu)+\frac{1}{\mu^2(1-\mu)^2}\right\}.\end{aligned} \tag{4.2}$$

It is clear that this δ function is nonlinear in both y and μ. Solving nonlinear equation (4.1) can be done iteratively by the Newton-Raphson algorithm or quasi-Newton algorithm. The calculation of the Fisher information requires the knowledge of $\mathrm{E}\{-\dot{\delta}(Y_i;\mu_i)\}$. It is equivalent to deriving $\frac{1}{2}\mathrm{E}\ddot{d}(Y_i;\mu_i)$.

Differentiating $\dot{d}$ *w.r.t.* μ gives

$$\begin{aligned}\frac{1}{2}\ddot{d}(y;\mu) &= \frac{1}{\mu(1-\mu)}d(y;\mu)+\frac{1-2\mu}{\mu^2(1-\mu)^2}(y-\mu)d(y;\mu)\\ &\quad+\frac{1}{\mu^3(1-\mu)^3}+\frac{1-2\mu}{\mu^4(1-\mu)^4}(y-\mu)\\ &\quad-\frac{1}{\mu(1-\mu)}(y-\mu)\dot{d}(y;\mu)-\frac{2(2\mu-1)}{\mu^4(1-\mu)^4}(y-\mu).\end{aligned} \tag{4.3}$$

Hence,

$$\begin{aligned}\frac{1}{2}\mathrm{E}\{\ddot{d}(Y;\mu)\} &= \frac{1}{\mu(1-\mu)}\left[\mathrm{E}\{d(Y;\mu)\} - \mathrm{E}\{(Y-\mu)\dot{d}(Y;\mu)\}\right] \\ &\quad + \frac{1-2\mu}{\mu^2(1-\mu)^2}\mathrm{E}\{(Y-\mu)d(Y;\mu)\} + \frac{1}{\mu^3(1-\mu)^3} \\ &= \frac{3\sigma^2}{\mu(1-\mu)} + \frac{1}{\mu^3(1-\mu)^3}, \qquad (4.4)\end{aligned}$$

where the last equation holds by applying part (e) of Proposition 4.1 below. Therefore, the Fisher information is

$$\mathbf{i}(\beta) = \frac{1}{\sigma^2}\sum_{i=1}^{K}\mathbf{x}_i u_i^{-1}\mathbf{x}_i^T,$$

where

$$u_i = \frac{\mu_i(1-\mu_i)}{1+3\sigma^2\{\mu_i(1-\mu_i)\}^2},\ i = 1,\ldots,K.$$

As seen in (4.3), the first order derivative of the deviance score $\dot{\delta}$ appears tedious, but its expectation in (4.4) is much simplified. Therefore, it is appealing to implement the Fisher-scoring algorithm in the search for the solution to the score equation (4.1). One complication in the application of Fisher-scoring algorithm is the involvement of the dispersion parameter σ^2. This can be resolved by replacing σ^2 with a $\sqrt{K}$-consistent estimate, $\hat{\sigma}^2$. A consistent estimate of such a type can be obtained by the method of moments. For example, the property (a) in Proposition 4.1 is useful to establish an estimate of σ^2 as follows:

$$\hat{\sigma}^2 = \frac{1}{K-(p+1)}\sum_{i=1}^{K} d(y_i;\hat{\mu}_i). \qquad (4.5)$$

Proposition 4.1. *Suppose* $Y \sim S^-(\mu;\sigma^2)$ *with mean* μ *and dispersion* σ^2. *Then,*

(a) $E\{d(Y;\mu)\} = \sigma^2$;

(b) $E\{(Y-\mu)\dot{d}(Y;\mu)\} = -2\sigma^2$;

(c) $E\{(Y-\mu)d(Y;\mu)\} = 0$;

(d) $E\{\dot{d}(Y;\mu)\} = 0$;

(e) $\frac{1}{2}E\{\ddot{d}(Y;\mu)\} = \frac{3\sigma^2}{\mu(1-\mu)} + \frac{1}{\mu^3(1-\mu)^3}$;

(f) $var\{d(Y;\mu)\} = 2\left(\sigma^2\right)^2$;

(g) $var\{\delta(Y;\mu)\} = \frac{3\sigma^4}{\mu(1-\mu)} + \frac{\sigma^2}{\mu^3(1-\mu)^3}$.

The proof of Proposition 4.1 can be found in Song and Tan (2000) and Song (2007).

4.2 Simulation Experiment

For simplicity, often in practice the normal linear regression model is applied to analyze the proportional data. This analysis apparently ignores the confinement of the support of the data distribution, which can potentially cause problems in prediction. A simple solution to overcome this is to invoke logit transformation on the observations directly. Thus, in the application of the simplex GLM, one issue that deserves some attention is whether there is much difference between the normal linear model based on logit-transformed data, $\log\{y_i/(1-y_i)\}$, and the direct simplex GLM. The difference between the two models is that the former models $\mathrm{E}[\log\{Y_i/(1-Y_i)\}]$ as a linear function of covariates, and that the latter models $\mu_i = \mathrm{E}(Y_i)$ via $\log\{\mu_i/(1-\mu_i)\}$ as a linear function of covariates, so that the mean μ_i can be expressed as an explicit non-linear function of the covariates. As a result, the direct GLM approach gives rise to much ease in interpretation.

The following simulation study suggests that when the dispersion parameter σ^2 is large, the performance of the logit-transformed analysis may be questionable, if the data are really from a simplex distributed population.

The simulation study assumes the proportional data are generated independently from the following simplex distribution,

$$Y_i \sim S^-(\mu_i, \sigma^2),\ i = 1, \ldots, 150,$$

where the mean follows a GLM of the following form:

$$\mathrm{logit}(\mu_i) = \beta_0 + \beta_1 T_i + \beta_2 S_i.$$

Covariates T and S are presumably drug dosage levels indicated by $\{-1, 0, 1\}$ for each 50 subjects and illness severity score ranged in $\{0, 1, 2, 3, 4, 5, 6\}$ that is randomly assumed to each subject by a binomial distribution $\mathrm{B}(7, 0.5)$. The true values of regression coefficients are set as $\beta_0 = 0.5, \beta_1 = -0.5, \beta_2 = 0.5$, and the dispersion parameter $\sigma^2 = 0.5, 50, 200, 400$.

For each combination of parameters, the same simulated data was fit by the simplex GLM for the original responses and the normal linear model for logit-transformed responses. Two hundred replications were done for each case. Results are summarized in Table 4, including the averaged estimates, standard deviations of 200 replicated estimates, and standard errors of estimates calculated from the Fisher information.

This simulation study indicates that *(i)* when the dispersion parameter σ^2 is small, the logit-transformed analysis appears fine, with little bias and little loss of efficiency, because

Table 4: Summary of the simulation results for the comparison between the direct simplex GLM analysis and logit-transformed linear model analysis.

Parameter	Simplex GLM			Logit-Trans LM		
True	Mean	Std Dev	Std Err	Mean	Std Dev	Std Err
$\sigma^2 = 0.5$						
$\beta_0(0.5)$	.4996	.0280	.0254	.5089	.0288	.0263
$\beta_1(-0.5)$	$-.5023$	.0330	.0308	$-.5110$	.0345	.0322
$\beta_2(0.5)$	.5015	.0195	.0205	.5101	.0199	.0222
$\sigma^2 = 50$						
$\beta_0(0.5)$	.5062	.0983	.0960	.8057	.1769	.1752
$\beta_1(-0.5)$	$-.5068$	.1141	.1185	$-.7998$	.2065	.2148
$\beta_2(0.5)$	.5170	.0860	.0835	.8153	.1366	.1483
$\sigma^2 = 200$						
$\beta_0(0.5)$	.5060	.1145	.1021	1.0162	.2741	.2541
$\beta_1(-0.5)$	$-.5262$	.1346	.1263	-1.0479	.3218	.3114
$\beta_2(0.5)$	.5238	.0971	.0899	1.0430	.1919	.2150
$\sigma^2 = 400$						
$\beta_0(0.5)$	.5253	.0963	.1032	1.2306	.2767	.2980
$\beta_1(-0.5)$	$-.5001$	.1486	.1275	-1.1336	.3888	.3652
$\beta_2(0.5)$	.5165	.1000	.0909	1.1686	.2286	.2521

of small-dispersion asymptotic normality; *(ii)* when the dispersion parameter is large, the estimation based on the logit-transformed analysis is unacceptable, in which bias increases and efficiency drops when the σ^2 increases.

One may try to make a similar comparison by simulating data from the normal distribution as well as from the beta distribution. Our simulation study suggested that in the case of normal data, the direct simplex GLM performed nearly as well as the normal model, with only a marginal loss of efficiency; in the case of beta distributed data, the simplex GLM clearly outperformed the normal linear model. Interested readers can verify the findings through their own simulation studies.

4.3 Data Analysis

Penrose et al. (1985) reports a dataset including variables such as percentage of body fat, age, weight, height, and ten body circumference measurements (e.g., abdomen) for 252 men. The data is obtained at http://www.amstat.org/publications/jse/jse_data_archive.html. Body fat, a measure of health, is estimated through an underwater weighing technique. Percentage of body fat may be then calculated by either Brozek's equation or Siri's equation. Fitting body fat to the other measurements using GLM provides a convenient way of estimating body fat for men using only a scale and a measuring tape.

In this example, the simplex GLM is illustrated simply by fitting the the body fat index as a function of covariate `age`. Suppose the percentage of body fat $Y_i \sim S^-(\mu_i, \sigma^2)$, where

$$\log \frac{\mu_i}{1-\mu_i} = \beta_0 + \beta_1 \,\text{age}.$$

The Fisher-scoring algorithm was applied to obtain the estimates of the regression coefficients and the standard errors were calculated from the Fisher information. The results were summarized in Table 5, in which the dispersion parameter is estimated by the method of moments in (4.5). Clearly, from the results given in Table 5, age is an important predictor to the percentage of body fat in both Brozek's and Siri's equations. The dispersion σ^2 is found not small in this study, so according to the conclusions drawn from the above simulation studies, it might be worrisome for the appropriateness of either a direct linear model analysis (with no transformation on the response) or logit-transformed linear model analysis.

Table 5: Results in the regression analysis of body fat percentage using the simplex GLM.

	Parameter		
Body-fat measure	Intercept (Std Err)	Age (Std Err)	σ^2
Brozek's	−2.7929(0.3304)	0.0193(0.0070)	55.9759
Siri's	−2.8258(0.3309)	0.0202(0.0070)	57.0353

5 CONCLUDING REMARKS

This paper has focused on the regression analysis of one-dimensional cross-sectional data using the class of dispersion models. This analysis covers a wider range of models than those considered in the classical GLM theory, and the simplex GLM is illustrated in detail as an important practical example for the analysis of proportional data.

Extension of the univariate dispersion models to multivariate dispersion models has drawn much attention recently in the literature. Jørgensen and Lauritzen (2000) suggested an extension in that the multivariate distributions unfortunately are not marginally closed; that is, the marginal distributions of a multivariate dispersion model are not necessarily the same as the known margins. A more appealing extension was proposed by Song (2000) where the multivariate dispersion models are generated from multivariate Gaussian copulas. In the Song's extension, many desirable properties are established, including the property of marginal closure. Based on Song's extension, one of new developments is the vector generalized linear models (Song et al., 2009). On the other hand, applying dispersion models to the analysis of clustered or longitudinal data has also been studied extensively in the literature. For example, Artes and Jørgensen (2000), Song and Tan (2000), Song et al. (2004), Qiu et al. (2008) and Zhang et al. (2009). Most of the published works have focused on the aspect of estimation, and further works in model diagnostics, Bayesian estimation theory, and software development are worth serious exploration. Readers may refer to Song (2007) for more details about other applications of the dispersion models; for example, the dispersion models in time series data analysis.

Ackownledgements

The author thanks the anonymous referee for helpful suggestions on this paper. The author is grateful to Dr. S. Ejaz Ahmed for the invitation of this paper. The author likes to thank Dr. Z. Qiu for his assistance on the numerical results presented in this paper. This work is support in part by the NSF grant DMS-0904177.

REFERENCES

1. Artes, R. and Jørgensen, B. (2000). Longitudinal data estimating equations for dispersion models. *Scand. J. Statist.*, **27**, 321-334.

2. Barndorff-Nielsen, O.E. and Cox, D.R. (1989). *Asymptotic Techniques for Use in Statistics*. London: Chapman and Hall.

3. Jørgensen, B. (1987). Exponential dispersion models (with discussion). *J. Roy. Statist. Soc. Ser. B*, **49**, 127–162.

4. Jørgensen, B. (1997). *The Theory of Dispersion Models*. London: Chapman and Hall.

5. Jørgensen, B., Martinez, J.R. and Tsao, M. (1994). Asymptotic behaviour of the variance function. *Scand. J. Statist.*, **21**, 223-243.

6. Jørgensen, B. and Lauritzen, S.L. (2000). Multivariate dispersion models. *J. Mult. Analy.*, **74**, 267-281.

7. McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

8. Penrose, K., Nelson, A. and Fisher, A. (1985). Generalized body composition prediction equation for men using simple measurement techniques (abstract). *Med. Sci. Sport. Exer.*, **17**, 189.

9. Qiu, Z., Song, P.X.-K. and Tan, M. (2008). Simplex mixed-effects models for longitudinal proportional data. *Scand. J. Statist.*, **35**, 577-596

10. Song, P.X.-K. (2000). Multivariate Dispersion Models Generated from Gaussian Copula. *Scand. J. Statist.*, **27**, 305–320.

11. Song, P.X.-K. (2007). *Correlated data analysis: Modeling, analytics, and applications.* New York: Springer.

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

13. Song, P.X.-K., Qiu, Z. and Tan, M. (2004). Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biom. J.*, **46**, 540–553.

14. Song, P.X.-K., Li, M. and Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, **65**, 60-68.

15. Zhang, P., Qiu, Z., Fu, Y. and Song, P. X.-K. (2009). Robust transformation mixed-effects models for longitudinal continuous proportional data. *Canad. J. Statist.*, **37**, 266-281.