RC24863 (W0909-085) September 21, 2009
Mathematics

IBM Research Report

Gradient Boosting for Joint Regression Modeling of Mean and Dispersion

Ramesh Natarajan
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

GRADIENT BOOSTING FOR JOINT REGRESSION MODELING OF MEAN AND DISPERSION

Ramesh Natarajan
IBM T. J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY, 10598.

October 12, 2009

Abstract

We consider the joint regression modeling of the mean and dispersion for a conditional response from the Normal, Gamma and Inverse Gaussian distributions, which are continuous distributions from the exponential dispersion family for which the likelihood function takes a specific simple form. The regression methodology is based on extending Gradient Boosting (Friedman [8]) to incorporate dispersion modeling. When compared to a similar extension of Generalized Linear Models for dispersion modeling, this proposed new approach offers certain advantages, which include for example, the easy incorporation of relevant nonlinear and low-order covariate interaction effects in the regression function; robust and computationally-efficient modeling procedures suitable for large, high-dimensional data sets; and the ability to use high-cardinality categorical covariates directly in the regression without any *ad hoc* preprocessing or grouping of the feature levels that is often necessary for computational tractability in other modeling procedures. We provide the motivation, background theory and algorithmic details of the proposed methodology along with illustrative computational examples.

1 Introduction

In regression applications involving a conditional response distribution from the exponential dispersion family, the primary interest is typically in the regression model for the mean parameter. The corresponding dispersion parameter is either a fixed constant (e.g., unity for the Poisson and Bernoulli distributions), or is an unspecified constant that is estimated from the residual deviance values of the mean regression model.

However, the assumption of constant dispersion implies that the mean and variance of the conditional response distribution have some fixed and unchanging relationship as a function of the covariates, although in many cases these two quantities can plausibly vary quite independently. For example, Heller et al. [12] describe an auto insurance data set where the mean and variance for the loss severity distribution have an independent systematic variation across the levels of a rating factor corresponding to the driver age group (these variations were presumably the consequence of the details of the underwriting policies used in the individual levels for that rating factor, such as the specific values for the deductible and/or claim ceiling limits in each level).

Although the modeling of the dispersion parameter may not be the primary goal of the regression application, one consequence of any dispersion parameter variability is that it leads to non-constant case weights in the deviance-based loss functions used for the mean regression (irrespective of whether the dispersion variability is specified or to be estimated). In particular, accurate estimates for the variability of the dispersion parameter lead to tighter confidence bounds for the parameter estimates in mean regression, since the data associated with the larger dispersion values is appropriately down-weighted in the deviance-based loss function. However, the individual regression models for the response mean and dispersion cannot be obtained independently, or even sequentially, and their joint estimation is intrinsically coupled in the likelihood-based formulation for regression modeling.

The GLM methodology, which is widely used for mean regression modeling in the case of conditional response distributions from the exponential dispersion family (see [25] for a review focused on insurance applications), has been extended to incorporate the joint modeling of mean and dispersion by Smyth [23], Smyth et al. [24]. The use of non-parametric regression functions rather than linear models, has been considered the joint modeling case by Nott [21] and Chan et al. [4], who describe a semi-parametric Bayesian framework. Similarly, Gijbels et al. [9] consider the use of B-spline regression basis functions for the mean and dispersion in the univariate case, with model selection being performed using the Generalized Cross-Validation (GCV) criterion applied to a penalized likelihood function for the double exponential family ([6]).

An alternative to the GLM methodology for the mean regression case, is the well-known Gradient Boosting methodology of Friedman [8] (see also Hastie et al. [11]). For the mean regression modeling, this methodology has some specific advantages over GLM's as reviewed below, as well as over other comparable non-parametric extensions of GLM's, and these advantages extend to the joint modeling case as well.

First, even though Gradient Boosting may use the same link and loss function as the equivalent GLM formulation, it often provides a much better model fit by virtue of being able to directly and flexibly incorporate the most relevant low-order nonlinear and variable-interaction effects into the regression function.

Second, compared to other methods, Gradient Boosting is relatively unaffected by collinear features or by irrelevant (noise) features, both of which are invariably present in high-dimensional data sets. For these data sets, the equivalent GLM formulation would require careful feature selection and regularization in order to obtain robust models with stable parameter estimates.

Third, high-cardinality categorical covariates, particularly covariates whose levels cannot be ordered or grouped in any obvious way, can be directly incorporated into Gradient Boosting models. The equivalent GLM formulation, however, would require significant preprocessing effort, not only to encode the category levels for analysis, but also to group and aggregate the individual levels so as to reduce the feature cardinality for computational tractability in the modeling procedure. In particular, the impact of grouping the levels of a high-cardinality covariate on the quality of the eventual regression models is often unclear. This issue is of concern in insurance loss data sets, which typically include several high-cardinality categorical covariates of this kind, such as territory/zip code, occupation code, or car model type. For example, Mano and Rasa [16] describe the claims modeling procedure for an auto insurance data set, in which a sequence of preprocessing steps is used for the high-cardinality territory code covariate, with preliminary decision tree models or agglomerative clustering models in the first-stage to aggregate the levels of this covariate, and with the groupings from these first stage aggregation then being used as reduced-level factors, for a subsequent second-stage GLM model.

The outline of this paper is as follows. Section 2 summarizes the relevant properties of the exponential dispersion family for modeling of the Normal, Gamma and Inverse Gaussian conditional response distribution, as described in Section 3. Section 4 provides a brief summary of joint mean-dispersion modeling using GLM's, followed by the details for extending the Gradient Boosting methodology to the joint modeling case. Section 5 contains a description of some numerical studies with simulated data sets. Section 6 provides the summary discussion.

2 Properties of Exponential Dispersion Family Distributions

We review the properties of the exponential dispersion family that are relevant for regression modeling applications in this paper (a more extensive review with a specific focus on GLM's can be found in Jorgensen [13]).

A random variable $Y \sim ED^*(\theta, \phi)$ from the univariate exponential dispersion family with canonical parameter θ and dispersion parameter $\phi > 0$ has a probability density function

$$f(y; \theta, \phi) = \exp\{(y\theta - b(\theta) - a(y))/\phi - c(y, \phi)\}, \tag{1}$$

where $y \in Range(Y)$, $b(\theta)$ is the cumulant function, and $a(y)$, $c(y, \theta)$ are certain functions that are specified in (3) for individual distributions.

The cumulant generating function of (1) is given by

$$K_Y(t) \equiv \log E(e^{tY}) = (b(\theta + \phi t) - b(\theta))/\phi, \tag{2}$$

so that, from the identities $K'_Y(0) = E(y); K''_Y(0) = \text{Var}(Y)$, we have

$$E(Y) = b'(\theta),\ \text{Var}(Y) = \phi b''(\theta). \tag{3}$$

Thus, if μ denotes the mean parameter, we have from (3) that $\mu = b'(\theta)$, where $b'(\cdot)$ is the invertible mean-value mapping, and its inverse $b'^{-1}(\cdot)$ is termed the canonical link funtion. Similarly, from (3) we have $\text{Var}(Y) = \phi V(\mu)$, where $V(\mu) = b''(\theta) = b''(b'^{-1}(\mu)) \geq 0$ is the positive variance function.

The inverse mean-value mapping function $\theta = b'^{-1}(\mu)$ can be used to write $ED^*(\theta, \phi)$ (1) in the equivalent form $ED(\mu, \phi)$, which is explicit in the two parameters μ and ϕ of primary interest.

2.1 Convolution property for the $ED(\mu, \phi)$ Family

The convolution property of the exponential dispersion family yields the relationship between the statistical parameters of the distribution of homogeneous sample aggregates and that of the underlying distribution $ED(\mu, \phi)$.

Consider n independent, identically distributed random variables $Y_i \sim ED(\mu, \phi_i)$, where $\phi_i = \phi/\omega_i$ for $\omega_i > 0$, so that the Y_i have the same mean, but different inverse weights for the dispersion. Then from (2), the random variable $Y = (1/\omega)\sum_{i=1}^n \omega_i Y_i$ with $\omega = \sum_{i=1}^n \omega_i$ has the cumulant generating function

$$K_Y(t) = \sum_{i=1}^n K_{Y_i}(\omega_i t/\omega) = \frac{b(\theta + (\phi/\omega)t) - b(\theta)}{\phi/\omega}, \tag{4}$$

so that the weighted sum has the same distribution $Y \sim ED(\mu, \phi/\omega)$ as the individual Y_i, with $E(Y) = \mu$, $\mathrm{Var}(Y) = (\phi/\omega)V(\mu)$.

The case $\omega_i = 1$ in (4) with $\mathrm{Var}(Y) = \phi V(\mu)/n$ is of special interest, and the μ and ϕ parameters for n identically distributed $Y_i \sim ED^*(\mu, \phi)$ can be estimated from the distribution of the aggregate $Y = (1/n)\sum_{i=1}^n Y_i$, since from (4) we have $Y \sim ED(\mu, \phi/n)$. This result is frequently used to fit exponential dispersion models to data in contingency tables, where the the group averages in each cell k correspond to the aggregate distribution with mean μ_k and dispersion ϕ_k/ω_k, in which the weights ω_k are the known sample sizes of the aggregated data in cell k, and μ_k and ϕ_k are the respective mean and dispersion of the underlying distribution. However, for notational simplicity in this paper, just ϕ_k is used to denote the dispersion parameter in cell k, although when explicit case weights w_k for cell k are known, as described above, then ϕ_k must be replaced by ϕ_k/w_k wherever it appears.

2.2 Deviance and Negative Log-Likelihood functions

For the random variable $Y \sim ED^*(\theta, \phi)$ with mean μ and variance function $V(\mu)$, the deviance function $d(y, \mu)$ is defined by

$$d(y, \mu) = -2\int_y^\mu \frac{y - s}{V(s)} ds. \tag{5}$$

Thus, noting that $s = b'(t), V(s) = b''(t)$, we have

$$d(y, \mu) = -2\int_{b'^{-1}(y)}^{b'^{-1}(\mu)} (y - b'(t))dt = -2[y\theta - b(\theta) - a(y)]. \tag{6}$$

Comparing (6) with (1), we have

$$a(y) = yb'^{-1}(y) - b(b'^{-1}(y)), \tag{7}$$

from which we obtain the identity $a'(\cdot) \equiv b'^{-1}(\cdot)$. From (6), we obtain

$$\frac{\partial^2 d}{\partial y^2} = \frac{2}{V(\mu)} > 0, \tag{8}$$

whenever $V(\mu)$ is strictly positive, in which case $d(y, \mu)$ is convex as a function of y.

Similarly, we have

$$\frac{\partial^2 d}{\partial \mu^2} = 2\left\{1 + (y - \mu)\frac{V'(\mu)}{V(\mu)}\right\}\frac{1}{V(\mu)}, \tag{9}$$

so that while $d(y, \mu)$ is always monotonic, it is non-convex as a function of μ for $\mu - V(\mu)/V'(\mu) > y$ (in particular, for the Tweedie distributions, which are the subset of exponential dispersion family distributions with the variance function $V(\mu) = \mu^p$, the deviance is non-convex for $p > 1$

if $\mu(p-1)/p > y$). However, the Expected Value $E\left(\partial^2 d/\partial\mu^2\right) = 2/V(\mu) > 0$ of (9) is always convex whenever $V(\mu)$ is strictly positive.

From (1) and (6), the negative log-likelihood function $l(y;\mu,\phi)$ for the random variable $Y \sim ED(\mu,\phi)$ can be written in the form

$$l(y;\mu,\phi) = \frac{d(y,\mu)}{2\phi} + c(y,\phi). \tag{10}$$

Since the principal minors of the Hessian of (10) must be positive in order for $l(y;\mu,\phi)$ to be convex, from the discussion following (9) we have that $l(y;\mu,\phi)$ is non-convex as a function of μ and ϕ if either $d(y,\mu)$ is non-convex, or if $\phi > 2d(y,\mu)$. The second condition, in particular, indicates that $l(y;\mu,\phi)$ can be non-convex even when the deviance function $d(y,\mu)$ is convex. For example, $l(y;\mu,\phi)$ is non-convex for the Normal distribution, even though its deviance function $d(y,\mu)$ is convex. These regions of non-convex geometry in the likelihood surface $l(y;\mu,\phi)$ must be carefully considered in order to obtain robust, convergent optimization algorithms for joint modeling of mean and dispersion in (4).

3 The Normal, Gamma and Inverse Gaussian distributions

The Normal, Gamma and Inverse Gaussian distributions are the only members of the exponential distribution family for which $c(y,\phi)$ can be written as

$$c(y,\phi) = \frac{1}{2}[\hat{t}(y) - \hat{s}(\phi^{-1})], \tag{11}$$

(see Jorgenson [13]), and the corresponding negative log-likelihood function (10) takes the special form

$$l(y;\mu,\phi) = \frac{1}{2}[\phi^{-1}d(y,\mu) - \hat{s}(\phi^{-1}) - \hat{t}(y)]. \tag{12}$$

This special form can be used to derive iterative schemes for the maximum-likelihood estimation of the mean and dispersion parameters as described in Section (4.2) below (see also, Smyth [23]) .

Furthermore, these distributions are also special cases of the Tweedie family for which the variance function has the special form $V(\mu) = \mu^p$, with $p = 0$ (Normal), $p = 2$ (Gamma) and $p = 3$ (Inverse Gaussian) respectively. Therefore, the corresponding cumulant functions $b_p(\theta)$ for $p = 0, 2, 3$, can be obtained by solving $b_p''(\theta) = (b_p'(\theta))^p$ (with the arbitrary constants of integration being set to convenient simplifying values). Similarly, the corresponding deviance functions $d_p(y,\mu)$ for $p = 0, 2, 3$ can also be obtained by substituting the corresponding variance function in the integral (5).

These three cases are summarized below.

3.1 Normal distribution, $\mathcal{N}(\mu,\sigma)$

The Normal distribution is a $ED(\mu,\phi)$ distribution with $\phi = \sigma^2$ for which

$$b(\theta) = \theta^2/2, \tag{13}$$

$$d(y,\mu) = (y-\mu)^2, \tag{14}$$

so that the canonical link function is the identity link. Then comparing the likelihood with (10), we have

$$a(y) = y^2/2, \qquad c(y, \phi) = \frac{1}{2} \log 2\pi\phi, \tag{15}$$

so that $\hat{s}(\phi^{-1}) = \log \phi^{-1}$, and from (15), we have $\partial c/\partial\phi = (1/2)\phi^{-1}$, $\partial^2 c/\partial\phi^2 = -(1/2)\phi^{-2}$.

3.2 Gamma distribution, $Ga(\alpha, \beta)$

The Gamma distribution is a $ED(\mu, \phi)$ distribution with $\mu = \alpha/\beta$, $\phi = 1/\alpha$, for which

$$\begin{aligned} b(\theta) &= -\log(-\theta), & (16)\\ d(y, \mu) &= 2\left[\frac{y}{\mu} - \log\frac{y}{\mu} - 1\right], & (17) \end{aligned}$$

so that the canonical link function is the negative reciprocal link. From (10), we then obtain

$$\begin{aligned} a(y) &= -(1 + \log y) & (18)\\ c(y, \phi) &= \phi^{-1} - \phi^{-1}\log\phi^{-1} + \log y + \log\Gamma\left(\phi^{-1}\right). & (19) \end{aligned}$$

so that $\hat{s}(\phi^{-1}) = 2\left[\phi^{-1}\log\phi^{-1} - \phi^{-1} - \log\Gamma\left(\phi^{-1}\right)\right]$. In this case, to leading order for $\phi \longrightarrow 0$, we have $c(y, \phi) \approx (1/2)\log(2\pi\phi y^2) + \phi/12 + O(\phi^3)$. However, the derivatives of $c(y, \phi)$ (19) can be evaluated for all values of ϕ in terms of well-known special functions,

$$\begin{aligned} \frac{\partial c}{\partial\phi} &= \phi^{-2}\left[\log(\phi^{-1}) - \Psi(\phi^{-1})\right], & (20)\\ \frac{\partial^2 c}{\partial\phi^2} &= -\phi^{-3}\left[1 + 2\log(\phi^{-1}) - 2\Psi(\phi^{-1}) - \phi^{-1}\Psi'(\phi^{-1})\right], & (21) \end{aligned}$$

where $\Psi(.)$ and $\Psi'(.)$ denote the Digamma and Trigamma functions respectively (Abramovitz and Stegun, [1]).

3.3 Inverse Gaussian distribution

The Inverse Gaussian distribution $IG(\mu, \sigma)$ is a $ED(\mu, \phi)$ distribution with $\phi = \sigma^2$, so that

$$\begin{aligned} b(\theta) &= 2\theta^{1/2}, & (22)\\ d(y, \mu) &= \left[\frac{y}{\mu^2} - \frac{2}{\mu} + \frac{1}{y}\right], & (23) \end{aligned}$$

so that the canonical link is the square reciprocal link. From (10), we then have

$$\begin{aligned} a(y) &= -y^{-1}, & (24)\\ c(y, \phi) &= \frac{1}{2}\log 2\pi\phi y^3. & (25) \end{aligned}$$

Thus $\hat{s}(\phi^{-1}) = \log\phi^{-1}$, and from (25), we have $\partial c/\partial\phi = (1/2)\phi^{-1}$, $\partial^2 c/\partial\phi^2 = -(1/2)\phi^{-2}$.

3.4 Relevance to Other distributions

The form $c(y,\phi)$ in (11), which is exact only for the Normal, Gamma and Inverse Gaussian distributions, also results as the leading-order term for $\phi \longrightarrow 0$ for the saddlepoint density approximation to other conditional response distributions from the exponential dispersion family (Jorgenson [13]).

Following Goustis and Casella [10], an informal derivation of the saddlepoint likelihood approximation for a exponential dispersion family distribution, is obtained as follows. Since the density function $f(y;\theta,\phi)$ of a random variable $Y \sim ED^*(\theta,\phi)$ can be represented as the inverse Fourier transform of the characteristic function with imaginary argument $\exp\{K_Y(iz)\}$ (Feller [7], Ch. XV), we have

$$f(y;\theta,\phi) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{K_Y(iz)}e^{-izy}dz. \tag{26}$$

For a density function from the exponential dispersion family, we have from (2), after setting $\tilde{\theta} = b'^{-1}(y)$, noting from (7) that $a(y) = y\tilde{\theta} - b(\tilde{\theta})$, and comparing with (1) that

$$c(y;\phi) = -\log\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{(b(\theta+i\phi z)-b(\tilde{\theta})-y(\theta+i\phi z)+y\tilde{\theta})/\phi}dz. \tag{27}$$

The change of variable $iz = s$, then transfers the line integral in 27, to the imaginary axis in complex plane for s. However, since the integrand in (27) is analytic and vanishes for $iz \longrightarrow \pm i\infty$, the value of the inversion integral is unaffected by a translation of the line of integration parallel to the imaginary axis, and in particular, so that it passes through the saddlepoint of the argument of the exponential in (27). Thus, using the change of variables $iz = s - (\theta - \tilde{\theta})/\phi$, we have

$$c(y;\phi) = -\log\frac{1}{2\pi i}\int_{-i\infty}^{+i\infty} e^{(b(\tilde{\theta}+\phi s)-b(\tilde{\theta})-\phi sy)/\phi}ds. \tag{28}$$

The integrand in 28, for leading order as $\phi \longrightarrow 0$, is dominated by the contribution in the neighborhood of the saddlepoint. Thus

$$\begin{aligned} c(y;\phi) &\approx -\log\frac{1}{2\pi i}\int_{-i\infty}^{+i\infty} e^{\frac{1}{2}\phi b''(\tilde{\theta})s^2}ds \\ &= \log\sqrt{2\pi\phi V(y)} \\ &\equiv \tilde{c}(y;\phi), \end{aligned}$$

where $\tilde{c}(y;\phi)$ denotes the saddlepoint approximation to $c(y;\phi)$. The various steps in (29) follow from the fact that $b(\theta)$ is at least twice-differentiable, and at the unique saddlepoint $s = 0$ of $b(\tilde{\theta} + i\phi s) - b(\tilde{\theta}) - i\phi s$, we have $b''(\tilde{\theta}) = V(y) > 0$ (the special case when $V(y) = 0$ is addressed below).

From (29), the negative log-likelihood (10) with the saddlepoint approximation to the exact density, can be written as

$$l(y;\mu,\phi) \approx \frac{d(y,\mu)}{2\phi} + \log\sqrt{2\pi\phi V(y)}. \tag{29}$$

Thus (29) which is valid to leading order for $\phi \longrightarrow 0$, has the same form as (12), with $\hat{s}(\phi^{-1}) \approx -\log\phi$.

In fact, this approximation (29) is also the exact negative log-likelihood function for the Normal and Inverse Gaussian distributions, and Goustis and Casella [10] have shown that a renormalization of (29) leads to the exact negative log-likelihood for the Gamma distribution as well.

The saddlepoint likelihood (29) also provides a computationally simpler alternative to the exact negative log-likelihood for joint estimation of μ and ϕ in the small dispersion ($\phi \longrightarrow 0$) limit, and is equivalent to the extended quasi-likelihood function (Nelder and Pregibon [19]), and the double exponential family likelihood (Efron [6]) which have been widely used in GLM modeling.

For Tweedie family distributions $Tw_p(\mu, \phi)$ with $1 \leq p < 2$, which correspond to the Poisson and Compound-Poisson distributions respectively, the value $y = 0$ is in the range of the distribution with $V(0) = 0$. As a result the assumptions leading to the saddlepoint approximation are not uniformly valid for all values of y as $\phi \longrightarrow 0$, even though the corresponding $c(0; \phi$ is always well defined. However, for the distributions considered in this paper, the saddlepoint likelihood (12)) is either exact, or a well-defined approximation uniformly over the range of y.

4 Joint Regression Modeling

The overall framework for joint regression modeling is first described, followed by the description of the extension of the GLM and Gradient Boosting approaches within this framework.

4.1 Loss function for Joint Modeling

The loss function for joint regression modeling is the empirical negative log-likelihood for a conditional response variable from the $ED(\mu, \phi)$ family over the training data records $\{y_i, \boldsymbol{x}_i, \boldsymbol{z}_i\}_{i=1}^n$. From (10), this loss function is given by

$$\mathcal{L}(\mu, \phi) = \sum_{i=1}^{n} l(y_i; \mu_i, \phi_i) = \sum_{i=1}^{n} \left[\frac{d(y_i, \mu_i)}{2\phi_i} + c(y_i, \phi_i) \right]. \tag{30}$$

The regression functions for the mean and dispersion are given by $\eta(\boldsymbol{x})$ and $\xi(\boldsymbol{z})$ respectively, where $\boldsymbol{x}$ and $\boldsymbol{z}$ denote the respective vector of covariates (these two sets of individual covariates need not be mutually exclusive). Then, given suitable invertible link functions. $g : \text{Range}(Y) \longrightarrow \mathbb{R}$ and $h : \mathbb{R}^+ \longrightarrow \mathbb{R}$, we have

$$g(\mu) = \eta(\boldsymbol{x}), \qquad h(\phi) = \xi(\boldsymbol{z}). \tag{31}$$

For twice-differentiable loss (30) and the link (31) functions, the required parameter estimation can be carried out using Newton-type methods, which require the following derivatives

$$\frac{\partial l_i}{\partial \eta} = -\frac{(y_i - \mu_i)}{\phi_i V(\mu_i) g'(\mu_i)}, \qquad \frac{\partial l_i}{\partial \xi} = -\left(\frac{d_i}{2\phi_i^2} - \frac{\partial c_i}{\partial \phi_i} \right) \frac{1}{h'(\phi_i)}, \tag{32}$$

where we denote $d_i = d(y_i, \mu_i)$ and $c_i = c(y_i, \phi_i)$. Similarly, we have

$$\frac{\partial^2 l_i}{\partial \eta^2} = \left[1 + (y_i - \mu_i) \left\{ \frac{V'(\mu_i)}{V(\mu_i)} + \frac{g''(\mu_i)}{g'(\mu_i)} \right\} \right] \frac{1}{\phi_i V(\mu_i)(g'(\mu_))^2}, \tag{33}$$

$$\frac{\partial^2 l_i}{\partial\eta\partial\xi} = \frac{(y_i - \mu_i)}{\phi_i^2 V(\mu_i) g'(\mu_i) h'(\phi_i)}. \tag{34}$$

$$\frac{\partial^2 l_i}{\partial\xi^2} = \left[\left(\frac{d_i}{\phi_i^3} + \frac{\partial^2 c_i}{\partial{\phi_i}^2}\right) + \left(\frac{d_i}{2\phi_i^2} - \frac{\partial c_i}{\partial\phi_i}\right)\frac{h''(\phi_i)}{h'(\phi_i)}\right]\frac{1}{(h'(\phi_i))^2}. \tag{35}$$

From (34), it can be seen the μ and ϕ parameters for an exponential dispersion family distribution are statistically orthogonal since $E(\partial^2 l/\partial\mu\partial\phi) = 0$.

The use of the canonical link $g(\mu) = b'^{-1}(\mu)$ in mean regression model leads to considerable simplification in (32)-(35), since $g'(\mu)V(\mu) = 1$, $g''(\mu)V(\mu) + g'(\mu)V'(\mu) = 0$ in that case.

4.2 Joint Modeling using GLM's

4.2.1 General case

The usual GLM methodology for mean regression modeling (McCullagh and Nelder [17], [18]) can be extended to the joint modeling case, by taking $\eta(\boldsymbol{x})$ and $\xi(\boldsymbol{z})$ to be linear functions of the covariates

$$\eta(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}, \quad \xi(\boldsymbol{z}) = \boldsymbol{\gamma}^T\boldsymbol{z}, \tag{36}$$

Intercept terms can be incorporated in (36) in the usual way by adding a constant-valued dummy covariate to $\boldsymbol{x}$ and $\boldsymbol{z}$ respectively.

The gradient and Hessian of (30) with respect to the model coefficients in (36) is given by

$$(\nabla_{\boldsymbol{\beta}}\mathcal{L})_k = \sum_{i=1}^n \left(\frac{\partial l_i}{\partial\eta}\right)_k \boldsymbol{x}_i, \qquad (\nabla_{\boldsymbol{\gamma}}\mathcal{L})_k = \sum_{i=1}^n \left(\frac{\partial l_i}{\partial\xi}\right)_k \boldsymbol{z}_i, \tag{37}$$

and

$$(\mathcal{H}_{\boldsymbol{\beta}\boldsymbol{\beta}})_k = \sum_{i=1}^n \left(\frac{\partial^2 l_i}{\partial\eta^2}\right)_k \boldsymbol{x}_i\boldsymbol{x}_i^T, \quad (\mathcal{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}})_k = \sum_{i=1}^n \left(\frac{\partial^2 l_i}{\partial\xi^2}\right)_k \boldsymbol{z}_i\boldsymbol{z}_i^T, \tag{38}$$

$$(\mathcal{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}})_k = \sum_{i=1}^n \left(\frac{\partial^2 l_i}{\partial\eta\partial\xi}\right)_k \boldsymbol{x}_i\boldsymbol{z}_i^T, \tag{39}$$

These derivatives, when taken in conjunction with (32)-(35), lead to the following Newton iteration for estimating $\{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$ in the form

$$\begin{bmatrix} \boldsymbol{\beta}_{k+1} \\ \gamma_{k+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_k \\ \gamma_k \end{bmatrix} - \begin{bmatrix} (\mathcal{H}_{\boldsymbol{\beta}\boldsymbol{\beta}})_k & (\mathcal{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}})_k \\ (\mathcal{H}_{\boldsymbol{\gamma}\boldsymbol{\beta}})_k & (\mathcal{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}})_k \end{bmatrix}^{-1} \begin{bmatrix} (\nabla_{\boldsymbol{\beta}} l)_k \\ (\nabla_{\boldsymbol{\gamma}} l)_k \end{bmatrix} \tag{40}$$

As noted in (2.2), since $\mathcal{L}(\mu, \phi)$ is non-convex, the Hessian matrix (40) can be indefinite, so that this overall iteration is not globally convergent. Although various modifications of this basic Newton iteration, such as quasi-Newton methods and trust region methods [20], can be used to provide a globally convergent iteration, these techniques can be difficult to program from scratch for a high-dimensional optimization problem.

An alternative approach is the joint modeling analog of the Fisher scoring algorithm for GLM's, in which the Hessian matrix (40) is replaced by its Expected Value (an approximation that is accurate near the optimum solution). From (34), it can be seen that the cross-derivative term in (39) has has an Expected Value of 0, so that the Fisher scoring algorithm effectively decouples the updating of the parameters for the mean and dispersion, as follows,

$$\begin{aligned}\boldsymbol{\beta}_{k+1} &= \boldsymbol{\beta}_k - [E((\mathcal{H}_{\boldsymbol{\beta}\boldsymbol{\beta}}))_k]^{-1}(\nabla_{\boldsymbol{\beta}} l)_k, & (41)\\ \boldsymbol{\gamma}_{k+1} &= \boldsymbol{\gamma}_k - [E((\mathcal{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}))_k]^{-1}(\nabla_{\boldsymbol{\gamma}} l)_k. & (42)\end{aligned}$$

Since $E((\mathcal{H}_{\boldsymbol{\beta}\boldsymbol{\beta}}))$ in (41) is always positive-definite from (2.2), the iteration (41)-(42) will be globally convergent when $E((\mathcal{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}))$ is also positive-definite.

4.2.2 Special Case

We now consider (41)-(42) for the special case of the Normal, Gamma and Inverse Gaussian distributions, for which the negative log-likelihood takes the form (12). From (32) and (33), in this case we have

$$E(y_i) = \mu_i, \quad E(d_i) \equiv \delta_i = \hat{s}'(\phi_i^{-1}). \tag{43}$$

If we now denote the training data by

$$\boldsymbol{y} = [y_1, \ldots, y_n]^T, \quad \boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T, \quad \boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]^T, \tag{44}$$

we have from (36) that

$$g(\boldsymbol{\mu}) = \boldsymbol{\beta}^T\boldsymbol{X}, \quad h(\boldsymbol{\phi}) = \boldsymbol{\gamma}^T\boldsymbol{Z}. \tag{45}$$

From (33)-(35), we define the diagonal matrices

$$\begin{aligned}\boldsymbol{W} &= \text{Diag}\left\{\phi_i^{-1}(V(\mu_i))^{-1}(g'(\mu_i))^{-2}\right\}, & (46)\\ \boldsymbol{V} &= \text{Diag}\left\{-\frac{1}{2}\phi_i^{-4}\hat{s}''(\phi_i^{-1}))(h'(\phi_i))^{-2}\right\}, & (47)\\ \boldsymbol{D} &= \text{Diag}\left\{-\phi_i^{-2}\hat{s}''(\phi_i^{-1}))\right\}, & (48)\end{aligned}$$

and let

$$\boldsymbol{d} = [d((\mu_1, y_1), \ldots, d((\mu_n, y_n)]^T \qquad \boldsymbol{\delta} = [\hat{s}'(\phi_1^{-1}), \ldots, \hat{s}'(\phi_n^{-1}))]^T. \tag{49}$$

Thus, denoting the working responses at iteration k by

$$\begin{aligned}\hat{\boldsymbol{\mu}}_k = g'(\boldsymbol{\mu}_k)(\boldsymbol{y} - \boldsymbol{\mu}_k) + g(\boldsymbol{\mu}_k), & \quad (50)\\ \hat{\boldsymbol{\phi}}_k = h'(\phi_k)\boldsymbol{D}_k^{-1}(\boldsymbol{d}_k - \boldsymbol{\delta}_k) + h(\boldsymbol{\phi}_k), & \quad (51)\end{aligned}$$

the iterations in (41) and (42) can be written in the form,

$$\begin{aligned}\boldsymbol{\beta}_{k+1} &= (\boldsymbol{X}^T\boldsymbol{W}_k\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}_k\hat{\boldsymbol{\mu}}_k, & (52)\\ \boldsymbol{\gamma}_{k+1} &= (\boldsymbol{Z}^T\boldsymbol{V}_k\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{V}_k\hat{\boldsymbol{\phi}}_k. & (53)\end{aligned}$$

Each iteration step in (52)-(53) then corresponds to a pair of weighted least squares problems, in which the working responses (50)-(51), and the weight vectors $\boldsymbol{W}_k, \boldsymbol{V}_k$ in (48) are updated at each iteration.

Smyth [23] has suggested that the overall convergence of this iteration can be improved by using the most recently updated values $\boldsymbol{\mu}_{k+1}$ from the half-step (52) to obtain the working response and weight vector in order to solve for $\boldsymbol{\phi}_{k+1}$ from the half-step (53). In this way, the iteration alternates between (52) and (53), always using the most recent updates in each half-step of the iteration.

The specific form of (53) used to update the dispersion parameter submodel can be interpretated as GLM model in its own right, since the negative log-likelihood in (12) can be regarded as a "pseudo-deviance," which has the same form as (6), after interpreting the deviance residual d as a response, ϕ^{-1} as the canonical parameter, and $\hat{s}(\phi^{-1})$ as the cumulant function respectively.

In the case of the Normal and Inverse Gaussian distributions, we have $\hat{s}(\phi_i^{-1}) = \log \phi_i^{-1}$. Then from (43), we have $E(d_i) = \phi_i$ and $-\phi_i^{-2}\hat{s}''(\phi^{-1}) = 1$, so that in this case $\boldsymbol{D}_k$ in (51) can be set to the identity matrix, with other simplifications in $\boldsymbol{V}_k$ as well. In these two cases, therefore, the iteration for the dispersion submodel is identical to the Fisher scoring algorithm for a GLM with response d_i and $E(d_i) = \phi_i$, using the Gamma variance function $V(\phi_i) = \phi_i^2$ with fixed dispersion parameter equal to 2.

The Gamma distribution (3.2) is the only case where (12) holds with a non-trivial form for $\hat{s}(\phi^{-1})$, with (19)

$$\hat{s}'(\phi_i^{-1}) = 2\left[\log \phi_i^{-1} - \Psi(\phi_i^{-1})\right], \qquad \hat{s}''(\phi_i^{-1}) = 2\left[\phi_i - \Psi'(\phi_i^{-1})\right]. \tag{54}$$

As noted by Aitken [2], the special form of the likelihood function for the Normal and Inverse Gaussian distributions is such that the structure of the mean and dispersion sub-models in (52)-(53) correspond to Fisher scoring iterations for known loss functions in GLM's. This suggests that instead of programming the iteration in (52)-(53), one can use existing GLM routines as black-box programs to handle the joint modeling case. Therefore, starting from an initial estimate of ϕ, Aitken's procedure alternates between solving for μ with fixed ϕ and response y_i using the GLM model for the Normal or Inverse Gaussian as the case may be. This is followed by solving for ϕ with fixed μ and response d_i using the GLM model for the Gamma case. Although this approach can simplify the programming considerably, by making use of existing GLM routines in statistical packages, the convergence of this decoupled approach is likely to be slower than the tightly-coupled iterations used in (52) and (53).

4.3 Joint Modeling of Mean using Gradient Boosting

The Gradient Boosting algorithm [8] can be extended for joint modeling as follows. Let η_0 and ξ_0 denote constant offsets, and denoting the sample response mean by $\bar{y}$, on possibility is to set these constant offsets to $\eta_0 = g(\bar{y})$, $\xi_0 = h((1/(n-1))\sum_i d_i(y_i, \bar{y}))$, which are obtained from the maximum saddlepoint-likelihood estimators for the unconditional response.

In the Gradient Boosting formulation, the regression functions obtained (31) as stagewise expansions of the form

$$g(\mu_k) = \eta_k(\boldsymbol{x}), \qquad h(\phi_k) = \xi_k(\boldsymbol{z}), \tag{55}$$

where, starting with the offsets η_0 and ξ_0, we compute the terms at the next stage in the form

$$\eta_{k+1}(\boldsymbol{x}) = \eta_k(\boldsymbol{x}) + a_{k+1}R(\boldsymbol{x}, \boldsymbol{\beta}_{k+1}), \qquad \xi_{k+1}(\boldsymbol{z}) = \xi_k(\boldsymbol{z}) + b_{k+1}S(\boldsymbol{z}, \boldsymbol{\gamma}_{k+1}). \tag{56}$$

For each such stage, the positive scalar coefficients $\{a_{k+1},b_{k+1}\}$, and the parameterized basis functions $R(\boldsymbol{x},\boldsymbol{\beta})$,$S(\boldsymbol{z},\boldsymbol{\gamma})$ are chosen to obtain a significant reduction in the loss function (30).

Since the simultaneous optimization of all the variables in (56) is computationally difficult, the two-step algorithm described below is used to compute the terms at each stage.

4.4 Determination of the Stage Basis Functions

In the first step, the parameters $\{\boldsymbol{\beta}_{k+1},\boldsymbol{\gamma}_{k+1}\}$ for the basis functions in (56) are estimated from

$$\boldsymbol{\beta}_{k+1} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left[\left(-\frac{\partial l_i}{\partial \eta} \right)_k - R(\boldsymbol{x}_i,\boldsymbol{\beta}) \right]^2 , \tag{57}$$

$$\boldsymbol{\gamma}_{k+1} = \arg\min_{\boldsymbol{\gamma}} \sum_{i=1}^{n} \left[\left(-\frac{\partial l_i}{\partial \xi} \right)_k - S(\boldsymbol{z}_i,\boldsymbol{\gamma}) \right]^2 . \tag{58}$$

The resulting basis functions $\{R(\boldsymbol{x},\boldsymbol{\beta}_{k+1}),S(\boldsymbol{z},\boldsymbol{\beta}_{k+1})\}$ in (56) are optimal in the sense of being maximally correlated with the corresponding steepest-descent gradient direction of the loss function (30) evaluated at the current solution.

In Gradient Boosting, a suitable choice for these basis functions are regression trees of fixed depth, which are computed using a least-squares splitting criterion (57) and (58 as described in Breiman et al. [3]. For this choice, the parameters $\{\boldsymbol{\beta}_{k+1},\boldsymbol{\gamma}_{k+1}\}$ then represent the covariate split conditions and leaf-node response estimates in the respective regression stumps.

The use of regression trees with a splitting rule based the least squares criterion, as the stage basis functions in (56), is responsible for many of the attractive computational properties of the Gradient Boosting methodology as described further below. These regression trees, which are typically of small depth (either 1 or 2), are termed "weak learners" in the terminology of gradient boosting (Friedman [8]). In particular, regression trees of depth 1 (termed "stumps") lead to piecewise-constant models for $\eta(\boldsymbol{x})$ and $\xi(\boldsymbol{z})$ that have an additive structure in the covariates. Similarly, regression trees of depth 2 lead to piecewise-constant models $\eta(\boldsymbol{x})$ and $\xi(\boldsymbol{z})$ that incorporate first-order local interaction terms. Furthermore, different regression tree depths can be used in the regression functions $R(\boldsymbol{x},\boldsymbol{\beta})$ and $S(\boldsymbol{z},\boldsymbol{\beta})$ respectively, which provides an additional modeling flexibility. For example, the constant dispersion case is ensured by using a regression tree of depth zero as the stage basis function for $S(\boldsymbol{z},\boldsymbol{\beta})$).

4.5 Determination of the Expansion Coefficients

In the second step, after the basis functions $R(\boldsymbol{x},\boldsymbol{\beta}_{k+1})$ and $S(\boldsymbol{z},\boldsymbol{\gamma}_{k+1})$ have been computed from (57) and (58) as outlined above, the scalar coefficients $\{a_{k+1},b_{k+1}\}$ are obtained from

$$(a_{k+1},b_{k+1}) = \arg\min_{(a,b)} \hat{\mathcal{L}}_k(a,b), \tag{59}$$

where

$$\hat{\mathcal{L}}_k(a,b) = \mathcal{L}(g^{-1}(\eta_k + aR(\boldsymbol{x},\boldsymbol{\beta}_{k+1})), h^{-1}(\xi_k + bS(\boldsymbol{z},\boldsymbol{\gamma}_{k+1}))). \tag{60}$$

Since this is a bivariate optimization problem, this step differs from the equivalent step in Friedman [8], where only an univariate optimization problem was solved using a line search optimization algorithm. In the present case, the use of Newton's method seems preferable since it is unaffected by the local geometry of the optimization surface (59), which may quite poorly scaled in the initial stages of the stagewise procedure, and gradient line-search methods are known to be sensitive to the relative scaling of the optimization variables [20].

However, an important aspect of using Newton's method is that starting from an initial guess $(a, b) = (0, 0)$, only a single Newton step is taken, rather than iterating the optimization (59) to convergence. This "early stopping" may help in avoiding overfitting to the specific basis functions appearing in the early stages of the stagewise procedure. However, there is no loss in the overall accuracy of the stagewise procedure, since any sub-optimal solutions for (59) at a given stage, can be corrected since the same stage basis functions may be re-introduced in the subsequent stages of the overall stagewise procedure (56).

This single-step Newton iteration to solve (59) only requires the derivatives of (59) at $(a, b) = (0, 0)$, which are given by

$$(\nabla_a \hat{\mathcal{L}})_k = \sum_{i=1}^{n} \left(\frac{\partial l_i}{\partial \eta}\right)_k R(\boldsymbol{x}_i, \boldsymbol{\beta}_{k+1}), \qquad (\nabla_b \hat{\mathcal{L}})_k = \sum_{i=1}^{n} \left(\frac{\partial l_i}{\partial \xi}\right)_k S(\boldsymbol{z}_i, \boldsymbol{\gamma}_{k+1}), \tag{61}$$

$$(\hat{\mathcal{H}}_{aa})_k = \sum_{i=1}^{n} \left(\frac{\partial^2 l_i}{\partial \eta^2}\right)_k R(\boldsymbol{x}_i, \boldsymbol{\beta}_{k+1})^2, \quad (\hat{\mathcal{H}}_{bb})_k = \sum_{i=1}^{n} \left(\frac{\partial^2 l_i}{\partial \xi^2}\right)_k S(\boldsymbol{z}_i, \boldsymbol{\gamma}_{k+1})^2, \tag{62}$$

$$(\hat{\mathcal{H}}_{ab})_k = \sum_{i=1}^{n} \left(\frac{\partial^2 l_i}{\partial \eta \partial \xi}\right)_k R(\boldsymbol{x}_i, \boldsymbol{\beta}_{k+1}) S(\boldsymbol{z}_i, \boldsymbol{\gamma}_{k+1}), \tag{63}$$

and used in conjunction with (32)-(35), we have

$$\begin{bmatrix} a_{k+1} \\ b_{k+1} \end{bmatrix} = -\hat{\alpha} \begin{bmatrix} (\hat{\mathcal{H}}_{aa})_k + \hat{\lambda} & (\hat{\mathcal{H}}_{ab})_k \\ (\hat{\mathcal{H}}_{ba})_k & (\hat{\mathcal{H}}_{bb})_k + \hat{\lambda} \end{bmatrix}^{-1} \begin{bmatrix} (\nabla_a \hat{\mathcal{L}})_k \\ (\nabla_b \hat{\mathcal{L}})_k \end{bmatrix}. \tag{64}$$

Setting the parameters $\hat{\alpha} = 1$ and $\hat{\lambda} = 0$ in (64) yields the usual single-step Newton iteration for (59). However, as noted earlier, particularly in the initial stages of the stagewise procedure, the optimization geometry may be non-convex and the Hessian in (64) may not be positive-definite. Furthermore, even when the Hessian is positive-definite, the unmodified Newton step can occasionally lead to an increase rather than a desired decrease in the objective function (59).

Therefore, the two parameters $\hat{\alpha}$ and $\hat{\lambda}$ are introduced to modify the usual Newton step, so as to ensure a decrease in the objective function, using ideas from nonlinear optimization (Nocedal and Wright [20], page 74).

The first modification requires computing the the eigenvalues of the 2×2 Hessian matrix in (63), and in the semi-definite or indefinite case, we set $\hat{\lambda} = -(1 + \delta)\hat{\lambda}_{min}$, where $\hat{\lambda}_{min} \leq 0$ is the smaller eigenvalue, and $\delta > 0$ is some small positive constant. Th modified Newton iteration matrix is then positive-definite and well-conditioned, and the resulting modified Newton step is then a descent direction for the objective function (59).

The second modification involves selecting $0 < \hat{\alpha} \leq 1$ to ensure that the size of the step in the descent direction from the modified Newton step, actually leads to a decrease in the objective function (59), since this is only guaranteed for the small enough $\hat{\alpha}$. Therefore starting with an initial value of $\hat{\alpha} = 1$ in (64) which would be an unmodified Newton step, the objective function (59) evaluated, and if is not sufficiently decreased according to the well-known Armijo condition

$$\hat{\mathcal{L}}_k(\hat{\alpha} a_{k+1}, \hat{\alpha} b_{k+1}) \leq \hat{\mathcal{L}}_k(0,0) - c_1 \hat{\alpha}(a_{k+1}(\nabla_a \hat{\mathcal{L}})_k + b_{k+1}(\nabla_b \hat{\mathcal{L}})_k), \tag{65}$$

then $\hat{\alpha}$ is successively halved till the desired condition (65) is satisfied (the constant value c_1 in (65) must satisfy $0 < c_1 < 1$ and Nocedal and Wright [20] recommend using a small value, say $c_1 = 10^{-4}$). This choice of $\hat{\alpha}$ ensures that the largest possible modified Newton step is taken, consistent with a sufficient decrease in the objective function in this stage.

4.6 Optimum Number of Stages in Expansion

The optimum number of stages in the expansion (56), denoted by K, is obtained from the cross-validation estimate of the loss

$$K = \arg\min_k \sum_{I=1}^{N_{CV}} \mathcal{L}_{\{I\}}(\mu_k^{\{\backslash I\}}, \phi_k^{\{\backslash I\}}), \tag{66}$$

where $\mu_k^{\{\backslash I\}}$,$\phi_k^{\{\backslash I\}}$ denote the mean and dispersion estimates respectively at the k'th stage from the training data for the I'th fold (denoted by $\backslash I$), and $\mathcal{L}_{\{I\}}(\mu_k^{\{\backslash I\}}, \phi_k^{\{\backslash I\}})$ denotes the loss function evaluated on the test data in the I'th fold at the k'th stage. The number of cross-validation folds N_{CV} is typically 5 or 10. Other criteria such as the 1-SE rule (Breiman et al. [3]), in which the K is the smallest number of stages for which the cross-validation loss is within 1 standard error of the minimum cross-validation loss, can also be used as an alternative to (66).

4.7 Gradient Boosting Implementation details

For mean regression modeling, Friedman [8] has observed that Gradient Boosting leads to highly competitive and robust procedures, and many of his conclusions extend readily to the joint modeling situation discussed here as well. In particular, the computation of the stage basis functions, leads to a systematic and modular algorithmic approach that is independent of the specific form of the overall loss function. Furthermore, the use of the least-squares fitting criterion for for constructing the regression-tree basis functions in each stage is especially advantageous, since this particular criterion can be evaluated very rapidly using fast update algorithms, while examining the sequence of possible tree splits for the optimal split in the regression tree algorithm.

Another valuable benefit of using the least-squares fitting criterion in the stage regression tree computations, arises in the treatment of categorical covariate splits, where the convexity of the least-squares splitting criterion ensures that if Ω is the cardinality of a categorical covariate, then the best split over all the possible splits in this covariate can be obtained in just $O(\Omega)$ steps, without having to exhaustively search through $O(2^{\Omega})$ possible combinations (Breiman et al [3]). The ability to evaluate categorical splits in a linear rather than exponential number of steps allows categorical features of high cardinality to be used in the regression modeling without any preprocessing or grouping of the corresponding levels for computational tractability in the fitting procedure. In contrast, regression trees that directly use the overall loss function as the splitting criterion may not have this useful property for categorical covariates, since in (9) we have noted that the relevant loss functions may not always be convex, even in the mean regression case.

Finally, the use of regression trees as basis functions in the stagewise expansion is extremely useful for high-dimensional data sets with numerous input covariates. In this situation, the intrinsic dimensionality of the regression response surface is likely to be much smaller than the dimension of the input covariate space, and the greedy feature selection approach used for regression trees ensures that any collinear and irrelevant covariates are excluded from the regression models, so that there is no need for any separate pre-processing feature selection step in the computational procedure.

The ability of the regression functions to model low-order interaction effects in the covariates is achieved by using regression trees with depth greater than 1. In practice a depth of 2 or 3 usually suffices, since typically only the low-order interaction effects are important in the regression modeling. From the model interpretation point of view, the use of regression trees of depth of 1 has the advantage that the overall regression function is additive, making it possible to separate out the individual effects of the covariates in the final model.

Friedman [8] has described a number of regression diagnostics for Gradient Boosting models, which include quantitative measures for the relative importance of covariates appearing in the regression function, and partial dependence plots that show the trend in the conditional response as a function of each individual covariate, averaged over the remaining covariate effects. These diagnostics can be readily extended to the joint modeling case, and we refer to his paper for the relevant details.

5 Numerical Studies

It is difficult to find non-trivial, public data sets for analysing distributional and heteroscedasticity effects in regression modeling. For real-world data sets, reasonable fits and consistent interpretations may be obtained by many competing regression models. Therefore, we have also considered a few synthetic data sets with known response characteristics, in order to explicate some of the algorithmic issues in the computational results below.

5.1 Synthetic Data Sets

The response variable in the synthetic data sets considered below are always generated from a specific distribution with a known mean and dispersion dependency on the covariates. As a result, the joint regression methodology can be evaluated in terms of its ability to identify the correct response model, and recover the known signal from the data.

The generation of random variates with a given mean and dispersion profile from the Normal, Gamma and Inverse Gaussian distributions makes use of the well-known property of Tweedie distributions (Jorgensen [13]), that given a random variable $Y \sim Tw_p(\mu, \phi)$, then for $a > 0$, the random variable $aY \sim Tw_p(a\mu, a^{2-p}\phi)$ (or equivalently, aY has the same Tweedie distribution as Y, but with a scaled mean $a\mu$ and scaled variance $a^2\mu^p$) respectively.

It is well known that given a unit normal variate $Y \sim \mathcal{N}(0, 1)$, we have $\mu + \sigma Y \sim \mathcal{N}(\mu, \sigma)$, and the former can be sampled using the Cheng-Feast ratio-of-uniforms algorithm with linear pretest. Similarly, given a unit Gamma variate $Y \sim \mathcal{G}a(\alpha, 1)$, we have $\beta^{-1}Y \sim \mathcal{G}a(\alpha, \beta)$, and the former can be sampled using the Cheng-Feast ratio-of-uniforms algorithm for $\alpha > 1$, and the Ahrens-Dieter algorithm for $\alpha < 1$ (Glasserman [22], p. 126-127). Finally, given a unit Inverse

Gaussian variate $Y \sim IG(\mu\sigma^2, 1)$, we have $\sigma^{-2}Y \sim IG(\mu, \sigma)$, and the former can be sampled using the Michael-Schucany-Haas algorithm. These sampling algorithms for the unit Normal, Gamma and Inverse Gaussian distributions are discussed, for example, in Glasserman [22].

The synthetic data used here consists of 2000 samples, which are equally divided into training and validation data sets. The covariate set is 6-dimensional, $\boldsymbol{x} = \{x_1, x_2, \ldots, x_6\}$, with continuous-valued x_1, x_2, x_3 being uniformly sampled in the interval $(0, 1)$, and categorical-valued x_3, x_4, x_5 being uniformly sampled at 4 levels denoted by $\{1, 2, 3, 4\}$ respectively. The response is given by $y = Tw_p(\mu(\boldsymbol{x}), \phi(\boldsymbol{x}))$ where $p = 0, 2, 3$ respectively, and the mean and dispersion are given functions of the covariates. In these synthetic data sets, the mean is always an additive, nonlinear function of the following form (which we have taken from [15]),

$$\mu(\boldsymbol{x}) = 7 + 10(x_1 - 0.5)^2 - x_2 + \sin 2\pi x_2 - 3x_3 + 1.5 I_{[3,4]}(x_4), \tag{67}$$

with $I_S(x)$ denoting the unit indicator function (e.g., $I_S(x) = 1 \mathrm{if} x \in S, 0 \mathrm{otherwise}$).

The first collection of synthetic data sets, termed *SYNTH1*, are generated using a constant-dispersion response model $\phi(\boldsymbol{x}) = c$, with c being chosen to achieve a signal-to-noise ratio in the range $(1, 2)$ (the simulated response for the normal distribution is checked to ensure strictly-positive values, so that the data can also be explored using the non-normal loss functions for comparison purposes.

For example, in Table (1) the response is generated from the Normal, Gamma and Inverse Gaussian distributions respectively, and these data sets are fitted using all three corresponding loss models. In all these cases, the identity link was used for the mean sub-model and the log-link for the dispersion sub-mmodel. The results show that the "correct" loss model typically has the lowest validation-set loss, and in each case, the examination of the model fit also shows a consistent signal recovery. For small c, the identification of the correct model fit can be confounded (as can be noticed in the case of Gamma and Inverse Gaussian distributions) in Table (1), however, for large c the correct loss model always clearly provides the best model fit.

These results Table (1) also demonstrate the utility of the joint modeling formulation even in the constant-dispersion case, since the likelihood-based, variable-dispersion formulation allows alternative distributional models for the conditional response to be compared, which is not possible using the deviance-based loss function formulations typically used for mean regression modeling.

Table (2) considers the synthetic data set generated for Normal-distributed response case, with the Normal loss function being used for the model fit. The results show the effect of varying the tree-depth size in the mean and dispersion sub-models on the model accuracy. The best model accuracy is obtained using the tree depths $(1, 0)$, which are the simplest basis functions that consistent with the assumed covariate effects in the synthetic data set. The results show that using more-complex basis functions in this case, apparently leads to overfitting in the individual stages, which is reflected by the overall sub-optimal solution in the gradient boosting expansion.

The second collection of synthetic data sets, termed *SYNTH2*, is generated with the same mean sub-model as *SYNTH1*, and with a variable-dispersion sub-model of the form

$$\phi(\boldsymbol{x}) = c_1 I_{[1,2]}(x_4) + c_2 I_{[3,4]}(x_4). \tag{68}$$

Response Type	Model Loss Function		
	Normal	Gamma	Inverse Gaussian
$\mathcal{N}(\mu(\boldsymbol{x}), 0.6)$	$\underline{1.290}$	1.376	1.575
$Ga(\mu(\boldsymbol{x}), 0.08)$	2.100	$\underline{2.000}$	2.042
$Ga(\mu(\boldsymbol{x}), 1.0)$	3.386	$\underline{2.895}$	7.749
$IG(\mu(\boldsymbol{x}), 0.005)$	1.728	$\underline{1.604}$	1.628
$IG(\mu(\boldsymbol{x}), 0.2)$	3.661	2.916	$\underline{2.789}$

Table 1: Validation-set loss $\mathcal{L}_V$ for *SYNTH1* data set, with tree depths $l = 1$ for the mean and $k = 0$ for the dispersion. The estimated standard errors are negligible and have been omitted for brevity.

	Tree depths (l, k)				
	$(1, 0)$	$(1, 1)$	$(2, 0)$	$(2, 1)$	$(2, 2)$
$\mathcal{L}_V$	$\underline{1.317}$	1.393	1.386	2.588	69.77

Table 2: Validation-set loss $\mathcal{L}_V$ for *SYNTH1* data set for the response given by $\mathcal{N}(\mu(\boldsymbol{x}), 0.6)$ for varying tree depth sizes (l, k) .

Table (3) shows the effect of varying the tree depth in the respective fitted models, for the different response distributions in this variable dispersion case, using the "correct" loss function for the model fit. For all three response distributions, the choice $k = 1$ yields the best model fit, and we note that this choice also leads to the simplest basis function that is consistent with the assumed piecewise-constant variation in the generating model for the synthetic data.

5.2 Sniffer Data Set

The Sniffer data set (used in Smyth [23]) models the amount y of hydrocarbon vapors escaping while filling a gasoline tank. This data set, which can be obtained from `http://www.statsci.`

Tree depths	Response Type		
	Normal	Gamma	Inverse Gaussian
	$(c_1 = 0.2, c_2 = 2.0)$	$(c_1 = 0.08, c_2 = 1.0)$	$(c_1 = 0.08, c_2 = 0.8)$
$l = 1, k = 0$	1.598	2.822	2.810
$l = 1, k = 1$	$\underline{1.421}$	$\underline{2.566}$	2.655
$l = 2, k = 0$	1.661	2.876	2.817
$l = 2, k = 1$	1.490	2.597	$\underline{2.634}$
$l = 2, k = 2$	1.587	2.596	2.660

Table 3: Validation-set loss $\mathcal{L}_V$ for *SYNTH2* data set fitted with the "correct" loss model for the simulated response, for the case of varying tree depths (l, k). The estimated standard errors are negligible and excluded for brevity.

loss function	Identity link		
	$(l,k)=(1,0)$	$(l,k)=(1,1)$	$(l,k)=(2,1)$
Normal	2.838 (0.184)	3.231 (0.18)	3.227 (0.093)
Gamma	2.660(0.119)	3.056 (0.254)	2.815 (0.132)
Inverse Gaussian	2.558 (0.104)	2.968 (0.084)	2.735 (0.123)
	Canonical link		
	$(l,k)=(1,0)$	$(l,k)=(1,1)$	$(l,k)=(2,1)$
Normal	2.838 (0.184)	3.231 (0.18)	3.231 (0.18)
Gamma	2.770(0.137)	3.209 (0.099)	3.169 (0.050)
Inverse Gaussian	2.642 (0.110)	2.949 (0.221)	2.798 (0.096)

Table 4: Cross-validation loss estimates (with standard errors in brackets) for the Sniffer Data Set using various response distributions with different tree depths for the mean (l) and dispersion (k).

`org/data/general/gasvapor.html`, comprises of 125 measurement records, with four continuous explanatory variables, viz., the respective temperatures (x_1 and x_2) and vapor pressures (x_3 and x_4) of the original and dispensed gasoline in the tank.

In Smyth [23], several heteroscedastic GLM's were fitted to this data, and in his results, when assuming constant dispersion, the Gamma and Inverse Gaussian models yielded a better fit than the Normal model, indicating that the variance increased with the mean for the conditional response distribution. For the variable dispersion case, his results showed that the Normal model with the mean sub-model comprising of linear terms and an interaction term between the x_1 and x_4 variates, and the dispersion sub-model comprising of linear terms in x_2 and x_4, was considered to be the most suitable for the data (however, the conclusions in that paper appear to have be based solely on the training set loss estimates, so that possibility of overfitting in this result cannot be ruled out).

Table 4 shows the cross-validation estimates of the negative log-likelihood for heteroscedastic modeling with the present Gradient Boosting-based approach. The use of different tree depths l and k for the regression functions in the mean and dispersion sub-models controls the degree of the interaction effects in the respective sub-models. Our results are in agreement with Smyth's conclusions [23] for the constant dispersion case ($k = 0$), confirming the suggestion that the response variance increases with the mean. However, in marked contrast to Smyth's conclusions, our results show that the Normal model with a variable-dispersion sub-model (i.e., $k = 1$), or with a mean sub-model with first-order interaction terms (i.e., $k = 2$), are not so suitable for this data set. In fact, the best model fit is obtained with the Inverse Gaussian model with a constant dispersion (irrespective of whether the identity link or its canonical squared-reciprocal link was used in the mean regression model). The negative log-likelihood for the training data with this model for $l = 1, k = 0$ was 2.32 (which compares favorably with the value of 2.368 obtained the equivalent GLM fit in Smyth [23]). It is unclear if the identification of the conditional response as an Inverse Gaussian distribution with constant dispersion leads to any insights that are consistent with the physical origins of the data.

	$c_1 = 0.5, c_2 = 1.0$		$c_1 = 0.5, c_2 = 2.0$	
N_{Tr}	$(l,k) = (1,0)$	$(l,k) = (1,1)$	$(l,k) = (1,0)$	$(l,k) = (1,1)$
500	-0.279 (0.027)	-0.293 (0.024)	-0.222 (0.038)	-0.234 (0.032)
1000	-0.308 (0.030)	-0.325 (0.027)	-0.222 (0.042)	-0.293 (0.031)
2000	-0.305 (0.029)	-0.328 (0.026)	-0.229 (0.040)	-0.290 (0.031)

Table 5: Validation set loss (with standard errors in brackets) for modeling claims costs using heteroscedastic Gamma regression for a synthetic insurance data set for varying levels of dispersion heterogeneity, with the same mean response heterogeneity (N_{Tr} is the number of records in the training data set).

5.3 Synthetic Insurance Data Set

The synthetic insurance data sets are based on the well-known Canadian Auto Insurance data set of Bailey and Simon [5] (see `http://www.statsci.org/data/general/carinsca.html`), which contains aggregated claim frequency and claims cost data as a function of 2 covariate rating factors, viz., *Merit* (with 4 levels) and *Class* (with 5 levels), for a total of 20 cells. This web site for this data also describes GLM fits for the claims frequency (using Poisson regression) and the claims cost (using Gamma regression) for models incorporating the main effects.

The synthetic data sets for heteroscedastic modeling, therefore, use the mean profile from those GLM fits for generating the random variates for the claim frequency and cost, but with the claim cost also having an assumed variable-dispersion profile given by

$$\phi(\boldsymbol{x}) = c_1 I_{[1,2,3,5]}(Class) + c_2 I_{[4]}(Class), \tag{69}$$

so that the dispersion is a constant c_1, except in the cells with the rating factor $Class = 4$, when it another constant c_2 (the cells with $Class = 4$ comprise roughly 6% of the claim records in the data, so that this dispersion constrast is over a relatively small fraction of the overall data).

The following procedure is then used to generating the individual claim records in the synthetic data set. For each claim record to be generated, a combination of rating-factor levels is generated by random sampling using observed cell frequency of insured claims in the Bailey-Simon data set. The assumed GLM profiles for the claim frequency and claim cost corresponding to the sampled rating-factor level combination, are then used to generate a Poisson random variate for the claim frequency, and to generate Gamma random variates for the claim costs for each of these individual claim records upto the sampled claim frequency. The set of non-zero claim-frequency records then comprises the required synthetic insurance data set for heteroscedastic Gamma regression modeling, in which the cost per claim is the response variable, and the corresponding number of claims is the case weight for the record.

The results shown in Table (5) are all obtained using the log link for both the mean and dispersion sub-models. Two different sets of values are considered for (c_1, c_2), and in each case, training data sets of 3 different sizes, respectively containing approximately 50, 100 and 200 records respectively were used, in each case, with the same validation data set containng roughly 1000 records.

The results in Table (5) show that it is possible to discern variations in the dispersion parameter by comparing the constant dispersion sub-model ($k = 0$) case with the variable-dispersion sub-model ($k = 1$) case. We note that for the case with smaller contrasts in the dispersion heterogeneity (e.g., $c_1 = 0.5, c_2 = 1.0$), somewhat larger training data sets are required to elicit the signal for the variable-dispersion effect in the regression response model.

6 Summary

The extension of Gradient Boosting to the joint modeling case, must take into account the fact that the estimation of the mean and dispersion sub-models cannot be carried out independently, as these two sub-models are tightly coupled through the likelihood formulation.

The conditional response distributions considered in this paper, viz., the Normal, Gamma and Inverse Gaussian, are the only members of the exponential dispersion family for which the negative log-likelihood has the special form (12). Nevertheless, the overall methodology described in this paper can also be used for other conditional response distributions in the small dispersion limit $\phi \longrightarrow 0$, for which the leading-order term of the saddlepoint approximation to the negative log-likelihood also have the form (11) (see Jorgenson [13]), and we plan to discuss these extensions in detail elsewhere.

In real data sets (as opposed to the many synthetic data sets considered in this paper), it is often the case that a given choice of basis functions for the mean regression model is unable to effectively capture the intrinsic variation for the mean. The resulting systematic departures for the mean model in this case, can lead to a spurious inference of dispersion variability in the regression model. A careful cross-validation analysis, along with a detailed examination of a variety of regression model fits using different regression-tree depths for the stage basis functions, is required in order to rule out any such spurious inferences of variable dispersion.

The Gradient Boosting approach has been widely studied for mean regression modeling, where it is known to have certain advantages over GLM's and other comparable regression methods (see Friedman [8]), which carry over to the extension of this methodology to the joint regression modeling case as well. For example, the adaptive, non-parametric basis functions that are used in Gradient Boosting are suitable not only for capturing nonlinear additive effects, but also for capturing low-order covariate interaction effects in the regression sub-models. Furthermore, the Gradient Boosting procedure is relatively unaffected by the presence of collinear covariates and noise covariates, which are commonly encountered in high-dimensional data sets found in applications, and for which other comparable methods require careful modeling with specialized feature selection and regularization algorithms. Finally, in contrast to comparable methods, the Gradient Boosting approach can easily incorporate high-cardinality categorical covariates in the regression modeling, and there is no need for any preprocessing or grouping of these feature levels in order to reduce the feature cardinality for tractable computational modeling.

References

[1] M. Abramovitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York (1964).

[2] M. Aitkin, *Modelling Variance Heterogeneity in Normal Regression Using GLM*, Appl. Statist., Vol. 36(3), pp. 332-339 (1987).

[3] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *CART: Classification and Regression Trees*, Wadsworth, New York (1983).

[4] D. Chan, R. Kohn, D. J. Nott, and C. Kirby, *Adaptive nonparametric estimation of mean and variance functions*, J. Computational and Graphical Statistics, Vol. 15(4), pp. 915-936, (2006).

[5] R. A. Bailey and L. J. Simon, Two studies in automobile insurance ratemaking. ASTIN Bulletin, Vol. 1, pp. 192-217 (1960).

[6] B. Efron, *Double Exponential Families and Their Use in Generalized Linear Regression*, Journal of the American Statistical Association, Vol 81(395), pp. 709-721 (1986).

[7] W. Feller, *An Introduction to Probability Theory and its Applications*, Wiley, New York (1971).

[8] J. H. Friedman, *Greedy Function Approximation: A Gradient Boosting Machine*, Annals of Statistics, Vol. 29(5), pp. 1189-1232 (2001).

[9] I. Gijbels, I. Prosdocimi and G. Claeskens, *Nonparametric Estimation of Mean and Dispersion Functions in Extended Generalized Linear Models*, Report KBI 0815, Dept. of Decision Sciences and Information Management, Katholieke Universiteit, Leuven (2008).

[10] C. Goutis and G. Casella, *Explaining the Saddlepoint Approximation*, The American Statistician, Vol. 53(3), pp. 216-224, (1999).

[11] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer, New York (2001).

[12] G. Z. Heller, D. M. Stasinopoulos, R. A. Rigby and P. de Jong, *Mean and Dispersion Modeling for Policy Claims Costs*, Scandinavian Actuarial Journal, Vol. 107(4), pp. 281-292 (2007).

[13] B. Jorgensen, *The theory of exponential dispersion models*, Monographs on Statistics and Applied Probability, Vol. 76, Chapman and Hall, London (1997).

[14] D. Leslie, R. Kohn, and D. J. Nott, *A general approach to heteroscedastic linear regression*, Statistics and Computing, Vol. 17, pp. 131-146, (2007).

[15] B. Li and P. K. Goel, *Additive Regression Trees and Smoothing Splines: Predictive Modeling and Interpretation in Data Mining*, Department of Statistics Technical Report, Ohio State University (2006).

[16] C. Mano, E. Rasa, *Use of Classification Analysis for Grouping Multi-level Rating Factors*, 28th International Congress of Actuaries, Paris (2006).

[17] P. McCullagh, and J. A. Nelder, *Generalized Linear Models*, Chapman and Hall, London (1983).

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

[19] J. A. Nelder and D. Pregibon, *An Extended Quasi Likelihood Funtion*, Biometrika, Vol 74(2), pp. 221-232 1987).

[20] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, New York (1999).

[21] D. J. Nott, *Semiparametric estimation of mean and variance functions for non-Gaussian data*, Computational Statistics, Vol 21(3-4), pp. 603-620 (2006).

[22] P. Glasserman, *Monte Carlo Methods in Finanacial Engineering*, Springer-Verlag, New York (2004).

[23] G. K. Smyth, *Generalized Linear Models With Varying Dispersion*, J. Royal Statist. Soc. B., Vol. 51(1), pp. 47-60 (1989).

[24] G. K. Smyth, A. F. Huele and A. P. Verbyla, *Exact and Approximate REML for Heteroscedastic Regression*, Statistical Modeling, Vol. 1, pp. 161-175 (2001).

[25] G. G. Venter, *Generalized Linear Models Beyond the Exponential Family with Loss Reserve Applications*, Casualty Actuarial Society Forum, Arlington VA (2007).