Constrained Generalized Additive Model with Zero-Inflated Data

Hai Liu

The University of Iowa, Iowa City, U.S.A.

Kung-Sik Chan †

The University of Iowa, Iowa City, U.S.A.

Summary. Zero inflation problem is very common in ecological studies as well as other areas. We propose the COnstrained Zero-Inflated Generalized Additive Model (COZIGAM) for analyzing zero-inflated data. Our approach assumes that the response follows some distribution from the zero-inflated 1-parameter exponential family, with the further assumption that the probability of zero inflation is some monotone function of the mean response function. When the latter assumption obtains, the new approach provides a unified framework for modeling zero-inflated data. This bypasses the problems of two popular methods for analyzing zero-inflated data that either focus only on the non-zero data or model the presence-absence data and the non-zero data separately. We develop an iterative algorithm for penalized likelihood estimation with a COZIGAM, and derive formulas for constructing confidence intervals. The new approach is illustrated with both simulated data and two real applications.

Keywords: EM algorithm; Observed information; Penalized-iteratively re-weighted least squares; Penalized quasi-likelihood; Linear constraints

1. Introduction

Generalized additive models (GAMs) (Hastie and Tibshirani, 1990; Wood, 2006) are widely used in applied statistics, e.g., in ecological analysis; see, e.g., Ciannelli et al. (2008) and the references therein. Penalized likelihood methods provide powerful tools for estimating GAMs, see Wahba (1983), Green and Silverman (1994), Gu (2002) ,Wood (2000) and Wood (2006). In the GAM framework, the unknown smooth component functions can be estimated by maximizing the penalized likelihood which, in a simple case, equals

$$L(f) - (\lambda/2)J(f) \tag{1}$$

where f is the unknown regression function on the link scale, $L(f)$ is the log likelihood function, $J(f)$ is some roughness penalty and λ is the smoothing parameter that controls the trade-off between the goodness-of-fit and the smoothness of the function. A commonly used roughness measure is $J(f) = \int \|D^2 f\|^2$ where D^2 is the second derivative operator and $\|\cdot\|$ denotes the square norm. This roughness measure will be adopted in the real applications. Based on reproducing kernel Hilbert space theory and under mild regularity conditions, it can be shown that the maximizer of (1) is a linear combination of finitely

†*Address for Correspondence*: Kung-Sik Chan, Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA 52245, U.S.A.
E-mail: kung-sik-chan@uiowa.edu

many basis functions (the number of which generally increases with sample size), see Gu (2002). In particular, for $J(f) = \int \|D^2 f\|^2$, the maximizer is a smoothing spline, being natural cubic spline in the 1-dimensional case and thin-plate spline in higher dimensional cases, see Wood (2003) and Gu (2002). These results extend to the case of GAM when the mean function is the sum of more than 1 component functions on the link scale, and form the basis of some approaches for empirical GAM analysis; see Gu (2002) and Wood (2006).

A common problem encountered in ecological data is the presence of high number of zeroes, a problem known as zero inflation. For example, fisheries trawl survey data often contain a large number of zero catches, due to the fact that fish swim in schools influenced by food availability and irregular current pattern, see Ciannelli et al. (2008). Zero-inflation also occur in other fields, for example, in marketing where data on consumer choice may contain many non-purchase observations. Indeed, zero-inflated data abound in science and quantitative studies. Zero-inflated data are often analyzed via a mixture model that specifies the response variable as a probabilistic mixture of zero and a random variable belonging to some 1-parameter exponential family, the latter of which will be referred to as the regular component of the response, or simply regular response. See Rigby and Stasinopoulos (2005). The mixture model is sometimes analyzed with a two-stage approach that firstly analyzes the data with the responses dichotomized into zero or non-zero, the so-called absence-presence analysis, and then a second analysis with all non-zero data. For example, if the response distribution consists of a probabilistic mixture of zero and a log-normal distribution, then the two-stage approach models the log-transformed positive data by some additive model whereas the absence-presence pattern is performed by another GAM via, say, the logistic link. A draw-back of the two-stage approach is that the two separate model fits may result in conflicting conclusions. The same potential problem persists even with a likelihood analysis using all data including zeroes and non-zero data, as the regression function linked to the zero-inflation probability and that linked to the mean regular response are unconstrained. In other words, it is not surprising that different conclusions might be drawn from the zero data and the non-zero data under an unconstrained model. On the other hand, the presence-absence analysis are generally much less informative than the analysis with the non-zero data so that, even if the true regression functions are alike on the link scales, their estimates may well show conflicting conclusions owing to sampling variability. For recent surveys on zero-inflated data, see, e.g., Welsh et al. (1996), Agarwal et al. (2002) and Cunningham and Lindenmayer (2005).

In some cases, it is reasonable to expect that the mixing probability of the zero atom is a monotone function of the mean response. For example, if zero inflation results from under-reporting, then its probability may increase with lower mean response. Incorporating such a constraint on the GAM with zero-inflated data effectively removes the potential problem of having conflicting conclusions from a two-stage analysis. Here, we implement this new approach with the simplifying assumption that, on the link scales, the mixing probability of the zero atom is a linear function of the mean regular response which itself is modeled by a GAM with 1-parameter exponential-family response; below this new model is referred to as the COnstrained Zero-Inflated Generalized Additive Model (COZIGAM). We propose to estimate the COZIGAM by penalized likelihood. We introduce several useful parametrizations of the COZIGAM in Section 2, and propose an iterative algorithm for maximizing the constrained penalized likelihood. In the case that zero is a possible outcome for the regular response, e.g. if the regular response is conditionally Poisson, the penalized likelihood becomes more complex and the iterative estimation procedure has to

be augmented by steps based on the expectation-maximization (EM) algorithm (Dempster et al., 1977). The iterative estimation method and the formula for computing the observed Fisher information are presented in Section 3, together with some Monte-Carlo studies on the empirical coverage rates of associated confidence intervals. In Section 4, we illustrate the COZIGAM by two real examples. We briefly conclude in Section 5.

2. Model Formulation

2.1. Parametrization 1: Homogeneous Zero Inflation

Let the data be $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ and the covariates be $X = (X_1, X_2, \ldots, X_n)$ where Y_i are scalars and X_i are possibly high-dimensional vectors. In the first formulation, the probability of zero inflation is assumed to be constant. Assume that given the covariates X, the Y_i's are independently distributed. Moreover, the marginal conditional distribution of Y_i depends on the covariates only through x_i, which is a mixture distribution given by

$$Y_i|x_i \sim h_i(y_i) = \begin{cases} 0 & \text{with probability } 1-p \\ f(y_i|\theta_i) & \text{with probability } p, \end{cases} \tag{2}$$

where the zero atom models the zero inflation explicitly, and $f(y_i|\theta_i)$ is the probability density (mass) function pdf (pmf) that belongs to some 1-parameter exponential family distribution with θ_i as the canonical parameter (Nelder and Wedderburn, 1972) to be linked to the covariate x_i (see below). The exponential-family density can be expressed as

$$f(y_i|\theta_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi)\right\},$$

where it is assumed that $a_i(\phi) = \frac{\phi}{\omega_i}$, with ω_i being some known constants, often equal to 1, and ϕ is a dispersion parameter. Then

$$f(y_i|\theta_i) = \exp\left\{\frac{\omega_i(y_i\theta_i - b(\theta_i))}{\phi} + c_i(y_i, \phi)\right\}. \tag{3}$$

In the GAM setting, on the link scale,

$$g_\mu(\mu_i) = s(x_i)$$

where $\mu_i = E(Y_i) = b'(\theta_i)$ is the expectation of Y_i evaluated under f; $g_\mu(\cdot)$ is the link function and $s(\cdot)$ some smooth function to be estimated by the penalized likelihood approach. As discussed in the introduction, the penalized likelihood estimator of s generally equals some linear combination of certain basis functions. Moreover, the smooth function evaluated at x_i could be expressed as $\mathbf{X}_i\beta$, where $\mathbf{X}_i$ is the ith row of the design matrix $\mathbf{X}$ of the basis functions, and β is the parameter vector to be estimated. Consequently, without loss of generality, we have

$$g_\mu(\mu_i) = s(x_i) = \mathbf{X}_i\beta. \tag{4}$$

Hence, the unknown parameters of the model consist of $\Theta = (\beta^T, p)^T$.
(The extension to the case of replacing s by a sum of smooth functions with lower-dimensional arguments is straightforward.) Note that if $p \equiv 1$, then the model is a GAM whereas, in the general case, the model is a Zero-Inflated Generalized Additive Model (ZIGAM); below we shall refer the distribution with f as its pdf the regular distribution.

If the regular distribution assigns positive probability to zero, which is the case for many distributions including Poisson and binomial, the likelihood function becomes rather complex. The complexity owes to the fact that a zero observation may result from the zero atom or the regular distribution. If, however, the status of the zero observations are known, the likelihood becomes more tractable. This suggests the use of the EM algorithm for maximizing the penalized likelihood. Augment the data by the indicator variables $\mathbf{Z} = (Z_1, \ldots, Z_n)$ defined as follows

$$Z_i = \begin{cases} 1 & \text{if } Y_i \sim f(y_i|\theta_i) \\ 0 & \text{if } Y_i \sim 0. \end{cases} \tag{5}$$

The sequence $\{Z_i\}$ is independent and identically distributed as Bernoulli(p). The joint density of the complete data equals

$$f(\mathbf{y}, \mathbf{z}|p, \beta) = \prod_{i=1}^{n} \{pf(y_i|\theta_i)\}^{z_i} \{(1-p)I(y_i = 0)\}^{1-z_i}$$

and the complete-data log-likelihood equals

$$l(p, \beta) = \sum_{i=1}^{n} z_i \log\{pf(y_i|\theta_i)\} + (1 - z_i) \log(1 - p) + (1 - z_i) \log(I(y_i = 0)),$$

where $I(\cdot)$ is the indicator variable of the event enclosed in parentheses; $0 \log(0)$ is defined to be 0. Note that $z_i = 1$ if $y_i \neq 0$, in which case the above convention that $0 \log(0) = 0$ ensures that the corresponding term $(1 - z_i) \log(I(y_i = 0))$ has no contribution to the complete-data log-likelihood, as it should be. The roughness penalty term $(\lambda/2)J(f)$ can often be expressed as a quadratic form $\frac{1}{2}\beta^T \mathbf{S}\beta$ where $\mathbf{S}$ is a penalty matrix that is known up to the multiplicative smoothing parameter λ, see Gu (2002) and Wood (2006). Consequently, the penalized complete-data log-likelihood becomes

$$l_p(p, \beta) = \sum_{i=1}^{n} [z_i \log\{pf(y_i|\theta_i)\} + (1 - z_i) \log(1 - p) + (1 - z_i) \log(I(y_i = 0))] - \frac{1}{2}\beta^T \mathbf{S}\beta. \tag{6}$$

Estimation can be done by maximizing the above penalized log-likelihood, via an iterative algorithm detailed in Section 3.

2.2. Parametrization 2: Linear Constraint on the Zero Inflation Rate

A more general model is obtained by letting the zero-inflation probability to link to a smooth function of the covariate. However, as argued in section 1, it is of interest to impose the constraint that the smooth function linked to the zero-inflation probability is linearly related to the smooth function linked to the mean regular response. Specifically, we put a linear constraint on p on the link scale. Equation (2) is now modified to

$$Y_i|x_i \sim h(y_i) = \begin{cases} 0 & \text{with probability } 1 - p_i \\ f(y_i|\theta_i) & \text{with probability } p_i, \end{cases} \tag{7}$$

with the constraint that for some constants α and δ,

$$g_p(p_i) \quad = \quad \alpha + \delta s(x_i),$$

where $g_p(\cdot)$ is the link function of p, e.g., the logit function: $g_p(p) = \text{logit}(p) = \log \frac{p}{1-p}$. Recall $g_\mu(\mu_i) = s(x_i)$. Below, we sometimes write η_i for $s(x_i)$ so that $g_p(p_i) = \alpha + \delta\eta_i$. Denote the parameters by $\Theta = (\alpha, \delta, \beta^T)^T$. This constrained model will be called the COnstrained Zero-Inflated Generalized Additive Model (COZIGAM). Notice that now the zero atom contains information about β. Indeed,

$$\frac{\partial p_i}{\partial \beta_j} = \frac{\delta X_{ij}}{\dot{g}_p(p_i)}, \tag{8}$$

where for any function h, $\dot{h}$ denotes its first derivative and $\ddot{h}$ its second derivative. The penalized complete-data log-likelihood equals

$$l_p(\alpha, \delta, \beta) = \sum_{i=1}^{n} [z_i \log\{p_i f(y_i|\theta_i)\} + (1 - z_i) \log(1 - p_i) + (1 - z_i) \log(I(y_i = 0))] - \frac{1}{2}\beta^T \mathbf{S}\beta, \tag{9}$$

with the smoothing parameter λ included in the penalty matrix $\mathbf{S}$.

2.3. Parametrization 3: Linear Constraint on the Expectation

The preceding parametrization specifies $g_\mu(\mu)$ as a smooth function of the covariate and that $g_p(p)$ is a linear function of $g_\mu(\mu)$. Such a parametrization facilitates a framework for checking whether or not the zero inflation rate is homogeneous by testing whether or not $\delta = 0$. But it is invalid for the case that the expectation μ of the regular distribution is constant whereas the zero inflation rate is non-homogeneous. To deal with this case, we propose the third parametrization which specifies that $g_p(p)$ is a smooth function of the covariate and puts the linear constraint on $g_\mu(\mu)$. The model is then defined by (7), but with a different linear constraint:

$$g_\mu(\mu_i) \quad = \quad \alpha + \delta \; s(x_i),$$

where $g_p(p_i) = s(x_i)$. Again, we shall write η_i for $s(x_i)$ so that the linear constraint becomes $g_\mu(\mu_i) = \alpha + \delta \; \eta_i$. Note that the second and third parametrizations are equivalent if the slope parameter δ in one of the parametrizations is non-zero. The third parametrization, however, enables us to check whether or not the expectation μ of the regular response is homogeneous by testing whether or not $\delta = 0$.

The above two COZIGAM parametrizations use different bases for setting up the linear constraints. The two linear constraints can be subsumed as special cases of the constraint that $g_p(p)$ and $g_\mu(\mu)$ are linearly related, i.e. there exist constants κ and ξ, not both zero, such that $\kappa g_p(p) + \xi g_\mu(\mu)$ is a constant. If both κ and ξ are non-zero, then the second and third parametrizations are equivalent. However, if $\kappa = 0$, only the third parametrization is valid whereas if $\xi = 0$, only the second parametrization is valid. The advantage of the second and third parametrizations of the COZIGAM is that they facilitate testing for homogeneous zero inflation or homogeneous regular mean response. Furthermore, these two parametrizations have clear interpretation, and admits computationally simpler estimation algorithm (see below). However, before fitting the model, we may not know which parametrization is valid. One way to bypass this problem is to use a representation that is always valid under the general condition that $g_p(p)$ and $g_\mu(\mu)$ are linearly related. Below

is such a representation:

$$\begin{cases} g_\mu(\mu_i) = \alpha_1 + \delta s(x_i) \\ g_p(p_i) = \alpha_2 + (1-\delta)s(x_i), \end{cases}$$

where $\alpha_i, i = 1, 2$ and δ are constants. This is a symmetric representation which essentially uses the sum $g_\mu(\mu) + g_p(p)$ as the basis function. In this parametrization, $\delta = 1$ represents the interesting hypothesis that $g_p(p)$ is a constant function (homogeneous zero-inflation) whereas $\delta = 0$ is equivalent to the constancy of $g_\mu(\mu)$ (constant mean regular response). Note that for the model to be identifiable, the smooth function $s(\cdot)$ must be centered, i.e., of zero mean. The estimation algorithm for this more general parametrization is similar but more complex than the other ones, especially when computing the observed information. Recall that in the generic case when both κ and ξ are non-zero, the three COZIGAM parametrizations are equivalent. For conciseness, all subsequent theoretical development and real applications are carried out using the second parametrization, but the methods can be readily lifted to the other parametrizations.

3. Model Estimation

The proposed algorithm for estimating a COZIGAM is motivated by the Penalized Iteratively Re-weighted Least Squares (PIRLS) method (Wood, 2006, page 169) and the *Penalized Quasi-Likelihood* (PQL) method. The PQL method was exploited by Green (1987) for semiparametric regression. See, also, Breslow and Clayton (1993) for its use in estimating generalized linear mixed models (GLMM). As we mentioned earlier, if the regular distribution assigns positive probability to zero, the nature of the zero observations is unknown. The values of the indicator variable stating whether a zero observation is a realization of the zero atom or the regular response are then missing. Were these missing data available, the likelihood is more tractable. Thus, the EM algorithm will be made use in the proposed algorithm. We shall also derive the formulas for computing the observed Fisher information for the penalized estimators, which are useful for computing standard errors and confidence intervals. Throughout this section, the analysis will be done conditional on the covariate x. For simplicity, the dependency on x is generally suppressed from the notations, and we set $\omega_i \equiv 1$. Furthermore, we assume that the smoothing parameter is known in the derivation below. In practice, the smoothing parameter is generally unknown and need to be estimated by various criteria, e.g., GCV or UBRE; see Wood (2006). We shall return to the issue of estimating the smoothing parameter later.

3.1. Optimization with Homogeneous Zero Inflation

The optimization of the penalized likelihood of the homogeneous zero-inflated GAM can be implemented via the EM algorithm with Z defined by (5) as missing data. We first derive the conditional distribution of Z given the data. Write $f(y_i|\theta_i) = f(y_i)$. The joint density of $(\mathbf{Y}, \mathbf{Z})$ equals

$$f(\mathbf{y}, \mathbf{z}|p, \beta) = \prod_{i=1}^{n} \{pf(y_i)\}^{z_i} \{(1-p)I(y_i = 0)\}^{1-z_i},$$

hence the conditional distribution of $\mathbf{Z}$ given $\mathbf{Y}$ are independent with marginal conditional pdf given by

$$f(z_i|y_i;p,\beta) = \frac{f(y_i,z_i|p,\beta)}{f(y_i|p,\beta)} = \frac{\{pf(y_i)\}^{z_i}\{(1-p)I(y_i=0)\}^{1-z_i}}{pf(y_i)+(1-p)I(y_i=0)}$$

Therefore

$$Z_i|y;p,\beta \ \sim \text{Bernoulli}\left(\frac{pf(y_i)}{pf(y_i)+(1-p)I(y_i=0)}\right).$$

Denote $\psi_i = E(Z_i|y;p,\beta) = \frac{pf(y_i)}{pf(y_i)+(1-p)I(y_i=0)}$. Armed with these results, we can now state the EM algorithm for maximizing the penalized likelihood. Given the rth parameter iterate, the E-step and M-step are implemented as follows.

E-step
Let

$$\psi_i^{(r)} = E(Z_i|y_i,p^{(r)},\beta^{(r)}) = \frac{p^{(r)}f(y_i|\theta_i^{(r)})}{p^{(r)}f(y_i|\theta_i^{(r)})+(1-p^{(r)})I(y_i=0)}.$$

Then, up to an additive constant, the expected complete-data log-likelihood equals

$$E(l(p,\beta)|Y,p^{(r)},\beta^{(r)}) = \sum_{i=1}^{n}\psi_i^{(r)}\log pf(y_i|\theta_i) + (1-\psi_i^{(r)})\log(1-p)$$

and thence the expected penalized complete-data log-likelihood is given by

$$E(l_p(p,\beta)|Y,p^{(r)},\beta^{(r)}) = E(l(p,\beta)|Y,p^{(r)},\beta^{(r)}) - \frac{1}{2}\beta^T\mathbf{S}\beta.$$

M-step

- The next iterate for p equals

$$p^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n}\psi_i^{(r)} = \frac{1}{n}\sum_{i=1}^{n}\frac{p^{(r)}f(y_i|\theta_i^{(r)})}{p^{(r)}f(y_i|\theta_i^{(r)})+(1-p^{(r)})I(y_i=0)}.$$

  For simplicity, we henceforth write El_p for $E(l_p(p,\beta|\mathbf{Z})|Y,\Theta^{(r)})$.

- For estimating β, consider the score

$$\frac{\partial El_p}{\partial\beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\frac{\psi_i^{(r)}(y_i-\mu_i)}{V(\mu_i)}\frac{\partial\mu_i}{\partial\beta_j} - [\mathbf{S}\beta]_j \quad j=1,\ldots,K.$$

  The next iterate $\beta^{(r+1)}$ can be obtained by maximizing El_p as a function of β, which can be done via the PIRLS algorithm by simply treating $\psi_i^{(r)}$ as 'weight' at each iteration.

3.2. Optimization with the Linear Constraint

The above procedure can be modified for estimating the COZIGAM. We illustrate the proposed method in the setting of the second parametrization of the COZIGAM; the method works similarly for other parametrizations. The E-step requires a slight modification as follows: in the rth iteration,

$$\psi_i^{(r)} = E(Z_i|y_i, \Theta^{(r)}) = \frac{p_i^{(r)} f(y_i|\theta_i^{(r)})}{p_i^{(r)} f(y_i|\theta_i^{(r)}) + (1-p_i^{(r)}) I(y_i = 0)},$$

and

$$E(l(\Theta)|Y, \Theta^{(r)}) = \sum_{i=1}^{n} \psi_i^{(r)} \log\{p_i f(y_i|\theta_i)\} + (1-\psi_i^{(r)}) \log(1-p_i).$$

The objective function El_p for the M-step can then be readily computed.

The M-step is to find the maximizer of El_p with respect to the parameter $\Theta = (\alpha, \delta, \beta^T)^T$. Taking the first derivatives of the objective function, we get

$$\frac{\partial El_p}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{\psi_i^{(r)}(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} + \sum_{i=1}^{n} \frac{\psi_i^{(r)} - p_i}{p_i(1-p_i)} \frac{\partial p_i}{\partial \beta_j} - [\mathbf{S}\beta]_j, \quad j = 1, \ldots, K, \tag{10}$$

$$\frac{\partial El_p}{\partial \alpha} = \sum_{i=1}^{n} \frac{\psi_i^{(r)} - p_i}{p_i(1-p_i)} \frac{1}{\dot{g}_p(p_i)}, \tag{11}$$

$$\frac{\partial El_p}{\partial \delta} = \sum_{i=1}^{n} \frac{\psi_i^{(r)} - p_i}{p_i(1-p_i)} \frac{g_\mu(\mu_i)}{\dot{g}_p(p_i)}, \tag{12}$$

where β is assumed to be K-dimensional. Let ρ and τ be $n \times 1$ vectors whose components equal

$$\begin{aligned} \rho_i &= \frac{Z_i - p_i}{\dot{g}_p(p_i) p_i (1-p_i)}, \\ \tau_i &= \frac{Z_i (y_i - \mu_i)}{\dot{g}_\mu(\mu_i) \phi V(\mu_i)}. \end{aligned} \tag{13}$$

Define $\tilde{\rho}^{(r)} = E\left(\rho|\Theta^{(r)}\right)$, and $\tilde{\tau}^{(r)} = E\left(\tau|\Theta^{(r)}\right)$. Then equation (10) becomes

$$\frac{\partial El_p}{\partial \beta} = \delta \mathbf{X}^T \tilde{\rho}^{(r)} + \mathbf{X}^T \tilde{\tau}^{(r)} - \mathbf{S}\beta = 0.$$

The first set of equations (10) can be solved iteratively by modifying the PIRLS algorithm. The major obstacle for applying the PIRLS algorithm is that (10) involves two GAMs, one defined in terms of μ and another through p. The solution to this problem may be better understood by considering a more general equation:

$$\frac{1}{\phi_1} \sum_{i=1}^{n} \frac{w_{1i}(y_{1i} - \mu_{1i})}{V_1(\mu_{1i})} \frac{\partial \mu_{1i}}{\partial \beta_j} + \frac{1}{\phi_2} \sum_{i=1}^{n} \frac{w_{2i}(y_{2i} - \mu_{2i})}{V_2(\mu_{2i})} \frac{\partial \mu_{2i}}{\partial \beta_j} - [\mathbf{S}\beta]_j = 0 \quad \forall j,$$

where the two sums correspond to contributions from two GAMs with mean μ_{ki} linked to $\mathbf{X}_k\beta$ by the link function g_k, and variance function $V_k(\mu_{ki})$, $k = 1, 2$. However, these

equations are exactly the optimality conditions for finding β that minimizes the following non-linear weighted least squares:

$$\mathcal{S}_p = \mathcal{S}_1 + \mathcal{S}_2 + \beta \mathbf{S} \beta^T,$$

where

$$\begin{aligned} \mathcal{S}_1 &= \sum_{i=1}^{n} \frac{w_{1i}(y_{1i} - \mu_{1i})^2}{\phi_1 V_1(\mu_{1i})}, \\ \mathcal{S}_2 &= \sum_{i=1}^{n} \frac{w_{2i}(y_{2i} - \mu_{2i})^2}{\phi_2 V_2(\mu_{2i})}, \end{aligned}$$

assuming the weights $V_1(\mu_1)$ and $V_2(\mu_2)$ were known and independent of β.
The nonlinear least square problem can be solved iteratively. Let $\hat{\beta}^{[k]}$ be the kth iterate of β. Denote $\mu_t^{[k]}$ as the value of μ_t evaluated at $\hat{\beta}^{[k]}$. Defining diagonal matrices $\mathbf{V}_{t[k]}$ with the diagonal elements $V_{t[k]ii} = V_t(\mu_{ti}^{[k]})$, and the diagonal matrices $\mathbf{W}_t^*$ with $W_{tii}^* = w_{ti}/\phi_t$, $t = 1, 2$, we have

$$\mathcal{S}_t = \left\| \sqrt{\mathbf{V}_{t[k]}^{-1} \mathbf{W}_t^*} \left(\mathbf{y}_t - \mu_t(\beta) \right) \right\|^2, \quad t = 1, 2$$

Next approximate μ_t by its first order Taylor expansion around the kth estimate $\hat{\beta}^{[k]}$. Hence,

$$\mathcal{S}_t \approx \left\| \sqrt{\mathbf{V}_{t[k]}^{-1} \mathbf{W}_t^*} \mathbf{G}_t^{-1} \left(\mathbf{G}_t(\mathbf{y}_t - \mu_t^{[k]}) + \eta_t^{[k]} - \mathbf{X}_t \beta \right) \right\|^2, \quad t = 1, 2,$$

where $\mathbf{G}_t$ is a diagonal matrix with elements $G_{tii} = \dot{g}_t(\mu_{ti}^{[k]})$. Furthermore, by defining the 'pseudodata'

$$z_{ti}^{[k]} = \dot{g}_t(\mu_{ti}^{[k]})(y_{ti} - \mu_{ti}^{[k]}) + \eta_{ti}^{[k]}$$

and the diagonal weight matrices $\mathbf{W}_t^{[k]}$ with elements

$$W_{tii}^{[k]} = \frac{w_{ti}}{\phi_t V_t(\mu_{ti}^{[k]}) \dot{g}_t^2(\mu_{ti}^{[k]})}$$

we have

$$\mathcal{S}_t \approx \left\| \sqrt{\mathbf{W}_t^{[k]}} \left(\mathbf{z}_t^{[k]} - \mathbf{X}_t \beta \right) \right\|^2, \quad t = 1, 2.$$

Hence, at the kth iteration,

$$\mathcal{S}_p \approx \left\| \sqrt{\mathbf{W}_1^{[k]}} \left(\mathbf{z}_1^{[k]} - \mathbf{X}_1 \beta \right) \right\|^2 + \left\| \sqrt{\mathbf{W}_2^{[k]}} \left(\mathbf{z}_2^{[k]} - \mathbf{X}_2 \beta \right) \right\|^2 + \beta \mathbf{S} \beta^T.$$

Write

$$\mathbf{W}^{[k]} = \begin{pmatrix} \mathbf{W}_1^{[k]} & 0 \\ 0 & \mathbf{W}_2^{[k]} \end{pmatrix}$$

and

$$\mathbf{z}^{[k]} = \begin{pmatrix} \mathbf{z}_1^{[k]} \\ \mathbf{z}_2^{[k]} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

The two weighted sum of squares can be combined into a single penalized sum of squares:

$$\mathcal{S}_p \approx \left\| \sqrt{\mathbf{W}^{[k]}} \left(\mathbf{z}^{[k]} - \mathbf{X}\beta \right) \right\|^2 + \beta \mathbf{S} \beta^T, \tag{14}$$

the minimization of which yields the next iterate of β. In the case of unknown smoothing parameter, it can be estimated, e.g., by minimizing the GCV of the model corresponding to the preceding approximate weighted least squares; see Wood (2006) for a review of other criteria and efficient algorithms.

We apply this modified PIRLS algorithm for solving equation (10). After updating β using the modified PIRLS algorithm, the parameters (α, δ) can be updated by fitting the generalized linear model with ψ_i as the response, using the *quasi-binomial* family that links p_i to $\alpha + \delta\eta_i$ where, given the current estimate of β, $\eta_i = s(x_i)$ is known. The iteration can be repeated until all parameters converge according to some stopping criterion.

3.3. Computing the Observed Information

To compute the standard errors of the penalized likelihood estimators, we follow Louis' method (Louis, 1982) for computing the observed information matrix. Normal approximation of the sampling distribution of the estimators then yields a simple approach for constructing point-wise confidence intervals. Some simulation results will be given to examine the empirical performance of this approach.

3.3.1. Observed Information

We illustrate the application of Louis' method only for the second parametrization of the COZIGAM, as its application to other parametrizations is similar. We follow the notations in Louis (1982). Write $l_p^* = El_p$ as the imputed penalized complete-data log-likelihood where the expectation is conditional on the observed data and evaluated under the penalized parameter estimate. Recall $\Theta = (\alpha, \delta, \beta^T)^T$. Let $S(\mathbf{y}, \mathbf{Z}, \Theta)$ and $S^*(\mathbf{y}, \Theta)$ be the gradient vectors of l_p and l_p^* respectively, and $B(\mathbf{y}, \mathbf{Z}, \Theta)$, $B^*(\mathbf{y}, \Theta)$ be the corresponding negative second derivative matrices. Let I_Θ be the negative second derivative of the penalized likelihood of the observed data. From Louis (1982),

$$I_\Theta = E\left\{B(\mathbf{y}, \mathbf{Z}, \Theta)\right\} - E\left\{S(\mathbf{y}, \mathbf{Z}, \Theta) S^T(\mathbf{y}, \mathbf{Z}, \Theta)\right\} + S^*(\mathbf{y}, \Theta) S^{*T}(\mathbf{y}, \Theta), \tag{15}$$

where all expectations within this section are conditional on the observed data and computed under Θ.

The formula given in Louis (1982) is for the case of unpenalized likelihood estimation, but its extension to the penalized likelihood is straightforward. An interesting question arises as to how it relates to the observed information of the unpenalized log-likelihood function. By routine algebra, it can be shown that

$$I_{\text{penalized}}(\hat{\Theta}) = I_{\text{unpenalized}}(\hat{\Theta}) + \mathbf{S} \tag{16}$$

Thus, the penalized observed information is the unpenalized observed information plus the *prior* information specified by $\mathbf{S}$. In other words, we can view it as the observed *posterior* information assuming the estimated smoothing parameters are fixed.

We now implement the above approach for the COZIGAM. Partition I_Θ into a 3×3 block matrix according to the partition $\Theta = (\alpha, \delta, \beta^T)^T$, so that its $(1,1)$ block is denoted by I_α, the $(1,2)$ denoted by $I_{\alpha\delta}$, etc. Recall the n-dimensional vectors ρ and τ with components defined by (13). It is readily checked that

$$\dot{\rho}_i = \frac{\partial \rho_i}{\partial p_i} = \frac{-\dot{g}_p(p_i)p_i(1-p_i) - (Z_i - p_i)\left\{\ddot{g}_p(p_i)p_i(1-p_i) + \dot{g}_p(p_i)(1-2p_i)\right\}}{\dot{g}_p^2(p_i)p_i^2(1-p_i)^2} \tag{17}$$

and

$$\dot{\tau}_i = \frac{\partial \tau_i}{\partial \mu_i} = \frac{Z_i\left\{-\dot{g}_\mu(\mu_i)V(\mu_i) - (y_i - \mu_i)\left[\ddot{g}_\mu(\mu_i)V(\mu_i) + \dot{g}_\mu(\mu_i)\dot{V}(\mu_i)\right]\right\}}{\dot{g}_\mu^2(\mu_i)\phi V^2(\mu_i)}.$$

Then the first derivatives of the penalized log-likelihood equal

$$\begin{aligned}
\frac{\partial l_p}{\partial \alpha} &= \sum \rho_i, \\
\frac{\partial l_p}{\partial \delta} &= \sum \eta_i \rho_i, \\
\frac{\partial l_p}{\partial \beta} &= \delta \mathbf{X}^T \rho + \mathbf{X}^T \tau - \mathbf{S}\beta.
\end{aligned}$$

Also it can be readily checked that

$$EZ_iZ_j - \psi_i\psi_j = \begin{cases} 0 & i \neq j \\ \psi_i(1-\psi_i) & i = j. \end{cases} \tag{18}$$

For the scalar parameter α, we have

$$\frac{\partial^2 l_p}{\partial \alpha^2} = \sum \frac{\partial \rho_i}{\partial p_i}\frac{\partial p_i}{\partial \alpha} = \sum \frac{\dot{\rho}_i}{\dot{g}_p(p_i)}.$$

Combined with equation (18),

$$I_\alpha = \sum_{i=1}^n \left\{ \frac{-E\dot{\rho}_i}{\dot{g}_p(p_i)} - \frac{\psi_i(1-\psi_i)}{\dot{g}_p^2(p_i)p_i^2(1-p_i)^2} \right\} \tag{19}$$

Notice that in the above equation $E\dot{\rho}_i$ has exactly the same form as $\dot{\rho}_i$ defined by (17) except that Z_i is replaced by ψ_i. Similarly, since

$$\begin{aligned}
\frac{\partial^2 l_p}{\partial \delta^2} &= \sum \eta_i \frac{\partial \rho_i}{\partial p_i}\frac{\partial p_i}{\partial \delta} = \sum \frac{\eta_i^2 \dot{\rho}_i}{\dot{g}_p(p_i)} \\
\frac{\partial^2 l_p}{\partial \alpha \partial \delta} &= \sum \frac{\partial \rho_i}{\partial p_i}\frac{\partial p_i}{\partial \delta} = \sum \frac{\eta_i \dot{\rho}_i}{\dot{g}_p(p_i)},
\end{aligned}$$

it can be shown that

$$I_\delta = \sum_{i=1}^n \eta_i^2 \left\{ \frac{-E\dot{\rho}_i}{\dot{g}_p(p_i)} - \frac{\psi_i(1-\psi_i)}{\dot{g}_p^2(p_i)p_i^2(1-p_i)^2} \right\}, \tag{20}$$

and

$$I_{\alpha\delta} = \sum_{i=1}^n \eta_i \left\{ \frac{-E\dot{\rho}_i}{\dot{g}_p(p_i)} - \frac{\psi_i(1-\psi_i)}{\dot{g}_p^2(p_i)p_i^2(1-p_i)^2} \right\}. \tag{21}$$

Now we turn to β. To get the second derivatives, first note that $\frac{\partial \tau_i}{\partial \beta_j} = \frac{\partial \tau_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j} = \frac{\dot{\tau}_i X_{ij}}{\dot{g}_\mu(\mu_i)}$, and $\frac{\partial \rho_i}{\partial \beta_j} = \frac{\partial \rho_i}{\partial p_i}\frac{\partial p_i}{\partial \beta_j} = \frac{\delta \dot{\rho}_i X_{ij}}{\dot{g}_p(\mu_i)}$. Let $\mathbf{G}_\tau$ and $\mathbf{G}_\rho$ be two diagonal matrices with elements $G_{\tau ii} = \frac{-\dot{\tau}_i}{\dot{g}_\mu(\mu_i)}$ and $G_{\rho ii} = \frac{-\dot{\rho}_i}{\dot{g}_p(\mu_i)}$ respectively, then

$$\frac{\partial^2 l_p}{\partial \beta \partial \beta^T} = -\mathbf{X}^T \mathbf{G}_\tau \mathbf{X} - \delta^2 \mathbf{X}^T \mathbf{G}_\rho \mathbf{X} - \mathbf{S}.$$

It follows from (18) that

$$\begin{aligned}
E\tau\tau^T - E\tau E\tau^T &= \operatorname{diag}\left\{\frac{\psi_i(1-\psi_i)(y_i-\mu_i)^2}{\dot{g}_\mu^2(\mu_i)\phi^2 V^2(\mu_i)}\right\} := \mathbf{W}_\tau \\
E\rho\rho^T - E\rho E\rho^T &= \operatorname{diag}\left\{\frac{\psi_i(1-\psi_i)}{\dot{g}_p^2(p_i)p_i^2(1-p_i)^2}\right\} := \mathbf{W}_\rho \\
E\left(\tau\rho^T + \rho\tau^T\right) - \left(E\tau E\rho^T + E\rho E\tau^T\right) &= \operatorname{diag}\left\{\frac{2\psi_i(1-\psi_i)(y_i-\mu_i)}{\dot{g}_\mu(\mu_i)\phi V(\mu_i)\dot{g}_p(p_i)p_i(1-p_i)}\right\} := 2\mathbf{W}_{\tau\rho},
\end{aligned}$$

and hence

$$E\left(\frac{\partial l_p}{\partial \beta}\frac{\partial l_p}{\partial \beta^T}\right) - E\left(\frac{\partial l_p}{\partial \beta}\right)E\left(\frac{\partial l_p}{\partial \beta^T}\right) = \mathbf{X}^T\left\{\mathbf{W}_\tau + 2\delta\mathbf{W}_{\tau\rho} + \delta^2\mathbf{W}_\rho\right\}\mathbf{X}.$$

So the observed information matrix of β equals

$$I_\beta = \mathbf{X}^T\left(\tilde{\mathbf{G}}_\tau - \mathbf{W}_\tau\right)\mathbf{X} + \delta^2\mathbf{X}^T\left(\tilde{\mathbf{G}}_\rho - \mathbf{W}_\rho\right)\mathbf{X} - 2\delta\mathbf{X}^T\mathbf{W}_{\tau\rho}\mathbf{X} + \mathbf{S}, \tag{22}$$

where $\tilde{\mathbf{G}}_\tau = E\mathbf{G}_\tau$, $\tilde{\mathbf{G}}_\rho = E\mathbf{G}_\rho$.
Similarly, it can be checked that

$$I_{\alpha\beta} = -\delta\mathbf{X}^T\upsilon^{(\alpha)} - \mathbf{X}^T\nu^{(\alpha)} \tag{23}$$

where $\upsilon_i^{(\alpha)} = \frac{E\dot{\rho}_i}{\dot{g}_p(p_i)} + \frac{\psi_i(1-\psi_i)}{\dot{g}_p^2(p_i)p_i^2(1-p_i)^2}$, and $\nu_i^{(\alpha)} = \frac{\psi_i(1-\psi_i)(y_i-\mu_i)}{\dot{g}_\mu(\mu_i)\phi V(\mu_i)\dot{g}_p(p_i)p_i(1-p_i)}$;

$$I_{\delta\beta} = -\delta\mathbf{X}^T\upsilon^{(\delta)} - \mathbf{X}^T\nu^{(\delta)}, \tag{24}$$

where $\upsilon_i^{(\delta)} = \eta_i\left\{\frac{E\dot{\rho}_i}{\dot{g}_p(p_i)} + \frac{\psi_i(1-\psi_i)}{\dot{g}_p^2(p_i)p_i^2(1-p_i)^2}\right\}$, and $\nu_i^{(\delta)} = \frac{\psi_i - p_i}{\dot{g}_p(p_i)p_i(1-p_i)} + \frac{\eta_i\psi_i(1-\psi_i)(y_i-\mu_i)}{\dot{g}_\mu(\mu_i)\phi V(\mu_i)\dot{g}_p(p_i)p_i(1-p_i)}$.

Combining all the pieces together, we have the observed information matrix of Θ given by

$$I_\Theta = \begin{pmatrix} I_\alpha & I_{\alpha\delta} & I_{\alpha\beta}^T \\ I_{\alpha\delta} & I_\delta & I_{\delta\beta}^T \\ I_{\alpha\beta} & I_{\delta\beta} & I_\beta \end{pmatrix}. \tag{25}$$

Then $V_\Theta = I_\Theta^{-1}$ is the observed covariance matrix of the estimator $\hat{\Theta}$, and the square root of its diagonal elements yield the standard errors of the estimates.

3.3.2. Simulation Results

Assuming asymptotic normality for the estimators, point-wise confidence intervals can be readily constructed for each parameter and the smooth functions. The confident intervals are constructed based on the assumption that the smoothing parameters are fixed, while in fact they are estimated from the data by some criterion, e.g., GCV. So in this section we examine the empirical coverage probabilities of the confidence intervals via simulation.

The simulations are based on two test functions, denoted by s_1 and s_2, which are taken from Wood (2006). The test function s_1 has a 1-dimensional argument, while s_2 has a 2-dimensional argument(see Fig. 1).

$$\begin{aligned} s_1(x) &= 0.2x^{11}(10(1-x))^6 + 10(10x)^3(1-x)^{10} \\ s_2(x,z) &= 0.3 \times 0.4\pi \left\{ 1.2e^{-(x-0.2)^2/0.3^2-(z-0.3)^2} + 0.8e^{-(x-0.7)^2/0.3^2-(z-0.8)^2/0.4^2} \right\}. \end{aligned}$$

Poisson responses are generated so that, on the link scale, the mean equals the test func-

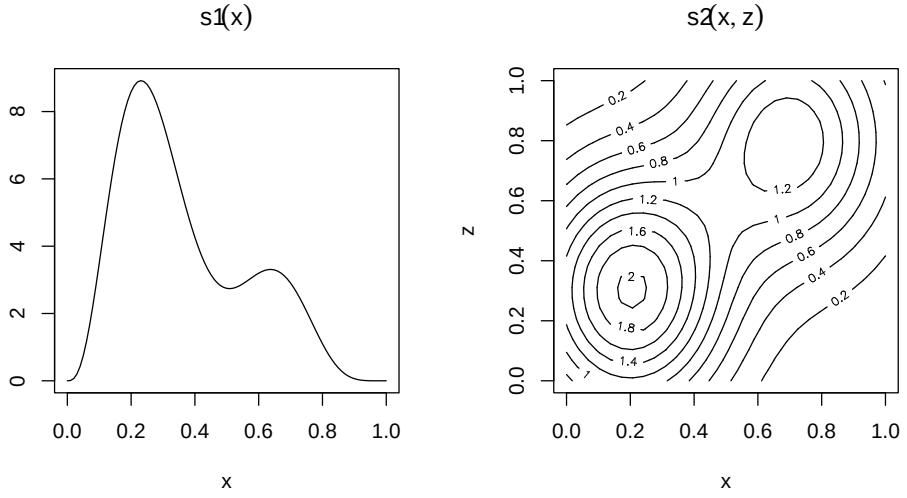

Fig. 1. Test Functions

tion, after some rescaling. Zero inflation occurs at a rate that is proportional to the Poisson mean on the logit scale (see the legend of Fig. 2). The smoothing parameter is chosen by GCV, as explained in Section 3. The fitting results for one simulated data are shown in Fig. 2. Notice that the plots are on the link scale and are centered.

We examined the performance of confidence intervals by checking the Across-the-Function Coverage (AFC), as was discussed by Wahba (1983) and Gu (2002). The average coverage proportion is defined as the coverage rate over the sampling points as follows.

$$\mathrm{ACP}(\alpha) = \frac{1}{n} \sharp \left\{ i : |\hat{s}(x_i) - s(x_i)| < z_{\alpha/2} \hat{\sigma}_{s(x_i)} \right\}$$

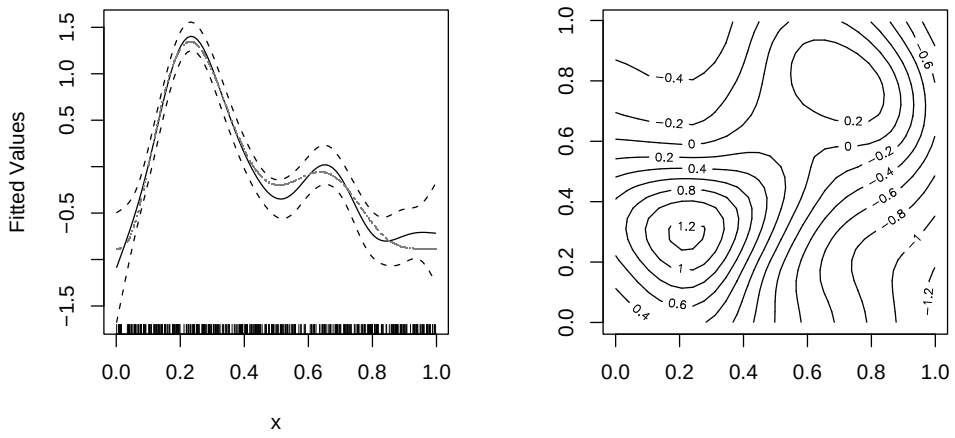

Fig. 2. Model Fits of Simulated Data: the left panel depicts an estimate of the test function s_1 with true $\alpha = -0.5$, $\delta = 1.5$, whose estimated values are $\hat{\alpha} = -0.550(0.377)$, $\hat{\delta} = 1.529(0.315)$. The gray dash-dot line is the true centered function, the black line is the estimated function, and the dashed lines are the 95% point-wise confidence band; The right panel displays an estimate of the test function s_2 with true $\alpha = -1.0$, $\delta = 1.5$, whose estimated values are $\hat{\alpha} = -1.042(0.420)$, $\hat{\delta} = 1.448(0.332)$. Sample size $n = 400$.

where $\hat{s}(x_i)$ is the predictor at point x_i obtained by assuming that the estimated smoothing parameter λ as known and fixed; denote $\hat{\sigma}_{s(x_i)}$ as the standard error of the predictor and $z_{\alpha/2}$ as the upper $\alpha/2$ quantile of standard normal distribution. The main results are listed the Table 1 based on 1000 replications, with nominal coverage probability 0.95.

The simulation results showed that the empirical coverage probabilities were very close to the nominal value for both the 1-dimensional and 2-dimensional test cases. As the test data are highly zero-inflated and nearly 50% of the responses are zeroes, there are, on the average, about 200 and 300 non-zero responses in each simulated dataset. For these simulation studies, the observed information matrix provided adequate approximation for assessing the variability in the estimator. Furthermore, the simulation results lend support

Table 1. Simulation Results

	Avg. Coverage Prop.	Coverage Prop. of α	Coverage Prop. of δ
$n = 400$			
s_1	0.945	0.951	0.958
s_2	0.942	0.946	0.948
$n = 600$			
s_1	0.944	0.953	0.957
s_2	0.952	0.949	0.939

to the result that

$$E[\text{ACP}(\alpha)] \approx 1 - \alpha,$$

see Wahba (1983).

4. Two Real Applications

4.1. Pollock Egg Density

The data analyzed in this example is part of an extensive survey data on walleye pollock egg density (numbers $10m^{-2}$) collected during the ichthyoplankton surveys of the Alaska Fisheries Science Center (AFSC, Seattle) in the Gulf of Alaska (GOA) from 1972 to 2000. Ciannelli et al. (2007) showed that the spatial-temporal distribution of the pollock egg in the GOA underwent a change around 1989-90. However, their analysis was confined to positive catch data and information from the zero catches were ignored. Here, we illustrate the use of the COZIGAM for extracting information from all data including zero catches. For simplicity, we only analyze the data from 1986 which contain 337 observations sampled from the 88th to the 137th Julian day over sites with bottom depth ranging 24-4171m. Among the 337 observations, 74 are zeroes, which make up over 20% of the data. The histogram of the pollock egg data, combining the zero catches and the log-transformed positive catches, in Fig. 3 shows the occurrence of high proportion of zero catches. The main goal is to explore the spatial and timing patterns of pollock spawning aggregations in the GOA. Pollock egg density is the response variable, and covariates include location (longitude and latitude), bottom depth and Julian day of the year. Bottom depth is log-transformed. We assume that the conditional response is a mixture distribution that equals zero with probability $1 - p$ but otherwise is log-normal with mean μ given by

$$\mu = c + s(lon, lat) + s(day) + s(log(depth)) \tag{26}$$

and

$$\text{logit}(p) = \alpha + \delta\mu, \tag{27}$$

where c, α, δ are parameters, s are assumed to be distinct smooth functions if they have distinct arguments; for model identifiability, the smooth functions are constrained to be of zero mean and hence the corresponding function estimates are centered over the data, i.e. of zero mean. In this and the next application, the roughness measure of the smooth functions are defined to be $J(s) = \int \|D^2 s\|^2$ where D^2 is the second derivative operator and $\|\cdot\|$ denotes the square norm. In particular, the penalized likelihood estimators are splines, being natural cubic spline in the 1-dimensional case and thin-plate spline in higher dimensional cases.

Under the above model assumptions, the regression function specified by (26) may be estimated by fitting an additive model with the log-transformed positive catch data, in which case the second equation concerning $\text{logit}(p)$ is dropped from the model. In particular, no systematic bias on the function estimates is expected from fitting the model with positive data only, which is borne out by the almost identical function estimates (unreported) whether we use only the positive data or all data in the model fit. Fig. 4 shows the estimated location and time effects based on the COZIGAM fitted with all data. But a narrower 95% point-wise confidence band for the bottom depth effect (Fig. 5) emerges from the model fit using all data including zero catches, as it makes use of all information in the data. These plots show that (i) egg density decreased monotonically starting from

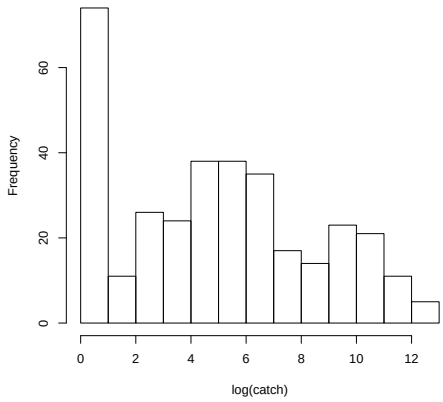

Fig. 3. Histogram of the combined log-transformed positive pollock egg catches and untransformed zero catches in 1986 at GOA

Table 2. Pollock Egg Density Estimation

Parameter	Point Estimate	Standard Error
α	-1.323	0.430
δ	0.536	0.093

the 88th Julian day and (ii) the density seemed to be more concentrated over deeper areas than shallower areas, see Fig. 5.

Table 2 reports the parameter estimates for Equation (27). The proportionality parameter δ is significantly positive. Thus, there is strong evidence indicating that zero inflation is more likely to occur at locations with less egg density.

As mentioned earlier, a two-stage approach that models the positive data and the presence-absence data separately may lead to contradictory result. This point can be illustrated using the pollock egg data. For example, using only the presence-absence data, we fit the model with Bernoulli response specified by the equation:

$$\text{logit}(p) = c + s(lon, lat) + s(day) + s(log(depth)).$$

Fig. 6 shows the estimated Julian day effect and bottom depth effect from the model fit with the presence-absence data. Note that these plots are very different from the fits from the COZIGAM fit using all data, or the GAM using only positive data.

It is tricky to carry out model diagnostics for models with zero-inflated data. Nonetheless, the validity of the log-normal regression assumption for the positive data may be explored with the model fit using only the non-zero data. Fig. 7 shows three model diagnostics plots: the normal QQ plot of residuals, residuals vs. fitted values plot, and observed (log-transformed) egg density vs. fitted value plot. These plots suggest that the model assumptions for the positive data are generally valid except for the presence of an outlier.

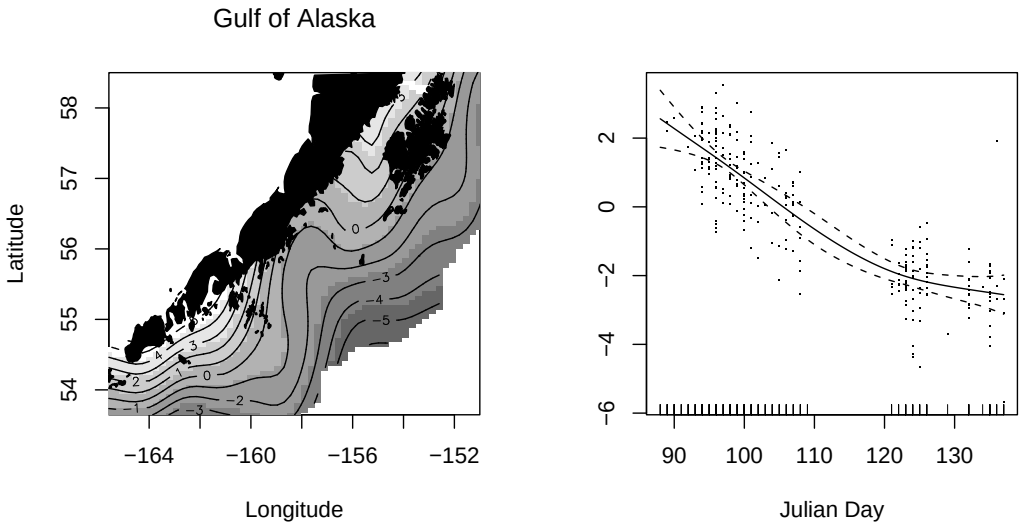

Fig. 4. Effects of Location and Time: the left diagram shows the contour plot of $s(lon, lat)$ on the right side of Equation (26) and the right diagram displays $s(day)$ with the partial residuals represented by dots.

After removing the outlier and refitting the model, the results, however, do not change very much. Therefore the log-normal regression assumption is reasonable according to the model diagnostics.

4.2. Larynx Cancer

The second application concerns an analysis of the Iowa larynx cancer data, obtained from SEER (http://seer.cancer.gov/data/access.html). The larynx, also known as the voice box, is an organ in the neck, which plays a role in breathing, swallowing, and talking. Larynx cancer is a 'rare disease' with an incidence rate of about 5 in 100,000 in the U.S.A. Although the exact causes of larynx cancer are unknown, known risk factors include smoking, alcohol consumption and exposure to sulfuric acid mist or nickel. The data analyzed below consist of the number of larynx cancer cases in each county of the state of Iowa, U.S.A. from 1980 to 1981. There are 99 counties in Iowa, among which 27 reported zero cases over the two-year period. A total of 286 larynx cancer cases were reported, see the histogram of the county number of larynx cancer cases reported in Fig. 8. (Because of confidentiality requirement, raw data cannot be reproduced herein.) So there are nearly 30% of zeroes in the dataset, which suggests possible zero-inflation. Zero-inflation may occur due to, for example, under-reporting of larynx cancer incidence which may be more likely to occur when the disease is relatively rare.

Two interesting questions concern estimating the overall larynx cancer rate in Iowa and its spatial variation over counties, after adjusting for zero inflation. We do this using the COZIGAM framework, with comparison to results based on regular GAM, assuming conditionally Poisson response. The covariates include geographic coordinates (longitude

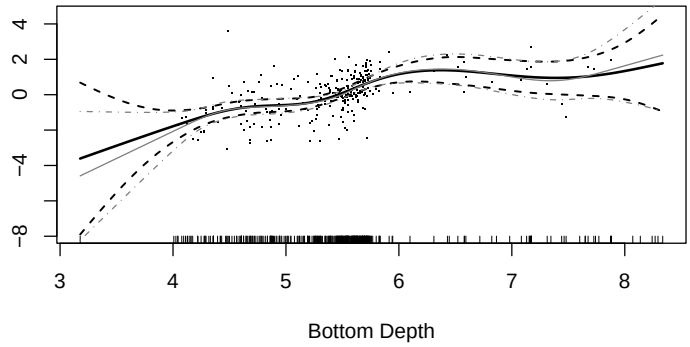

Fig. 5. Comparison of the COZIGAM (black) with the model using only the positive data (gray) on bottom depth effect; the dots represent the partial residuals from the COZIGAM.

Table 3. Iowa Larynx Cancer Estimation

Parameter	Point Estimate	Standard Error
β_0	-9.098	0.754
β_1	0.925	0.065
α	1.165	0.975
δ	2.366	1.451

and latitude) and log-transformed county population size (sum of yearly population over the study period). Specifically, the county numbers of larynx cancer cases constitute the response which is modeled to have a mixture distribution which for the ith county is zero with probability $1-p_i$ but otherwise a Poisson random variable with mean μ_i. The Poisson mean μ_i is a function of population, longitude and latitude:

$$\log(\mu_i) = \beta_0 + s(\log(pop_i)) + s(lon_i, lat_i)$$

where β_0 is the intercept which is related to the overall incidence rate; $s(\log(pop_i))$ and $s(lon_i, lat_i)$ are two smooth functions that are centered, i.e., of zero mean, over the observations. The probability p_i is specified as

$$\text{logit}(p_i) = \alpha + \delta \log(\mu_i),$$

where α and δ are parameters. Preliminary analysis (unreported) suggests that the covariate $\log(pop)$ affects the response linearly on the log scale, hence the model is simplified as follows:

$$\log(\mu_i) = \beta_0 + \beta_1 \log(pop_i) + s(lon_i, lat_i)$$

The corresponding model estimation results are listed in Table 3.

Note that the slope parameter δ is marginally significant, suggesting that zero inflation occurs more frequently with lower incidence of the disease. Because the location effect is

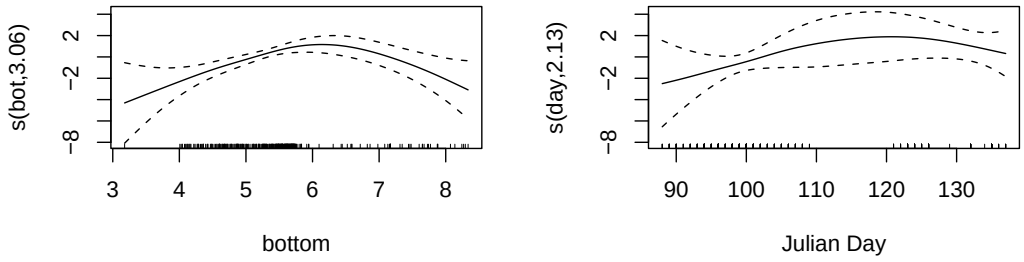

Fig. 6. Estimated Bottom Depth and Time Effect from the Presence-Absence Analysis of the Pollock Egg Data.

already centered at 0, the overall larynx cancer rate per 100,000 Iowans is estimated to be $100000^{\beta_1}e^{\beta_0}$, which is 4.727 according to the COZIGAM fit. The estimated larynx cancer incidence rate of 4.727 in 100,000 Iowans is close to the reported U.S. incidence rate of 5 in 100,000. In contrast, fitting the same dataset with a regular GAM with conditionally Poisson response distribution yields the estimated incidence rate of 4.501 in 100,000, an approximately 5% drop from the COZIGAM estimate. That the regular GAM yields a lower estimate owes to the fact that it ignores zero inflation in the data.

Below, we report further inference and model diagnostics with the constrained GAM fit. The county incidence rates per 100,000 Iowans can be estimated by $100000^{\beta_1}e^{\beta_0+s(lon_i,lat_i)}$, and they are shown in Fig. 9. Generally speaking, larynx cancer was relatively more prevalent in the southeast part of Iowa than its northwest counterpart. There were two areas located in the east boundary and the south boundary of Iowa that had relatively higher larynx cancer rates. The COZIGAM model pools information across counties to produce county larynx cancer rates that vary more smoothly across space than the crude rates calculated directly from raw data, the latter of which are subject to greater variability due to the small number of cases.

Model diagnostics with GAM (Wood, 2006) may be proceeded by examining the Pearson residuals, which are obtained by rescaling the raw residuals $\hat{\epsilon}_i = y_i - \hat{\mu}_i$ by their estimated standard deviation:

$$\hat{\epsilon}_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

For the COZIGAM, we can standardize the raw residuals in a similar way. Because of zero inflation, the mean and variance of Y_i have to be computed differently as follows.

$$\begin{aligned} E(Y_i) &= p_i\mu_i \\ Var(Y_i) &= p_i\mu_i(1+\mu_i - p_i\mu_i) \end{aligned}$$

Hence, the standardized residuals are given by

$$\hat{\epsilon}_i^* = \frac{y_i - \hat{p}_i\hat{\mu}_i}{\sqrt{\hat{p}_i\hat{\mu}_i(1+\hat{\mu}_i - \hat{p}_i\hat{\mu}_i)}}$$

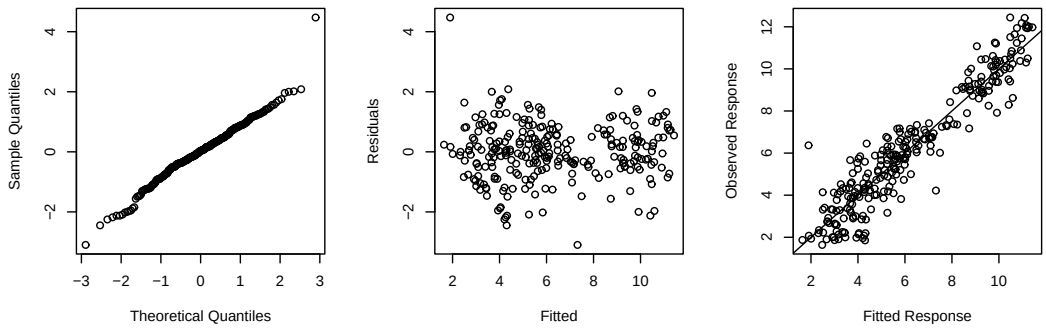

Fig. 7. Model Diagnostics Based on the Non-Zero Pollock Catch Data

Fig. 10 displays the Pearson residuals against the fitted values. Two outliers appear in the diagram, namely, Wayne county and Pocahontas county. Both counties had more larynx cancer cases in the years 1980 to 1981 relative to nearby counties after adjusting for population size. An interesting epidemiological question is to identify factors causing these two outliers.

5. Conclusion

In summary, we have presented a new approach for analyzing zero-inflated data, and a modified penalized-iteratively re-weighted least squares algorithm for model estimation. Simulation studies suggest that Louis' method for computing the observed information works well with the COZIGAM. The real data analysis reported in Section 4 illustrates the usefulness of the new approach. So far, the COZIGAM studied in this paper imposes linear constraints on the link scale. More general form of constraints remains to be studied. Also it is of interest to develop tests for the constraints imposed by the COZIGAM, i.e., tests for misspecification. Although simulation results in Section 3 provide some justification for our estimation approach, the asymptotic properties of the estimator constitutes an interesting open problem. These are some interesting directions for future work.

6. Acknowledgments

We thank Lorenzo Ciannelli for helpful comments and Michele West for assistance in obtaining the Iowa larynx cancer data via SEER*Stat. We gratefully acknowlege partial support from the US National Science Foundation (CMG-0620789) and North Pacific Research Board.

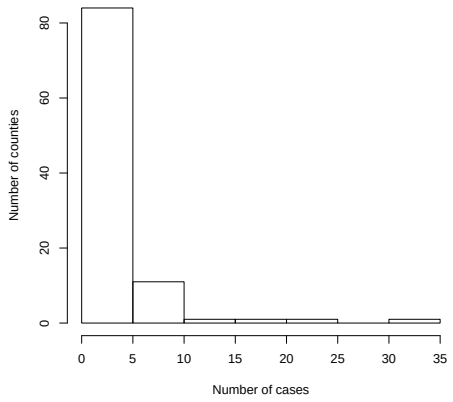

Fig. 8. Histogram of County Number of larynx Cancer Cases in Iowa, U.S.A, over 1980 and 1981.

References

Agarwal, D., A. Gelfand, and S. Citron-Pousty (2002). Zero-inflated models with application to spatial count data. *Environmental And Ecological Statistics 9*(4), 341–355.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*(421), 9–25.

Ciannelli, L., K. Bailey, K. S. Chan, and N. C. Stenseth (2007). Phenological and geographical patterns of walleye pollock spawning in the gulf of alaska. *Canadian Journal of Aquatic and Fisheries Sciences 64*, 713–722.

Ciannelli, L., P. Fauchald, K. S. Chan, V. Agostini, and G. Dingsr (2008). Spatial fisheries ecology: recent progress and future prospects. *Journal of Marine Systems*.

Cunningham, R. and D. Lindenmayer (2005). Modeling count data of rare species: Some statistical issues. *Ecology 85*(5), 1135–1142.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *J. R. Statist. Soc. B 39*, 1–38.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review 55*, 245–259.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer-Verlag.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. London: Chapman and Hall.

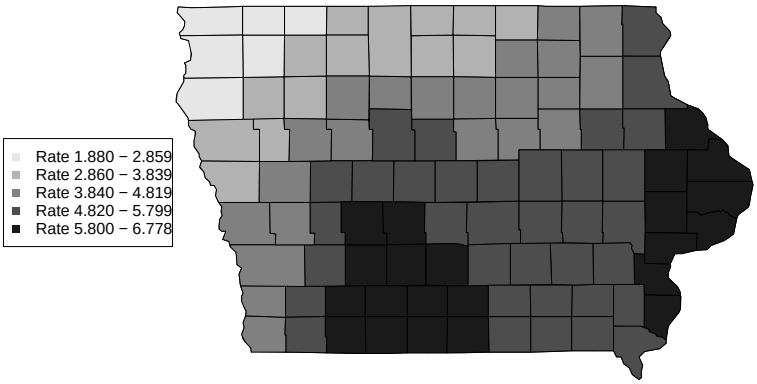

Fig. 9. Estimated Incidence Rate of Larynx Cancer in Iowa, 1980-81.

Louis, T. A. (1982). Finding the observed information matrix when using em algorithm. *J. R. Statist. Soc. B 44*, 226–233.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *J. R. Statist. Soc. A 135*, 370–384.

Rigby, R. and D. Stasinopoulos (2005). Generalized additive models for location, scale and shape (with discussion). *Appl. Statist. 54*, 507–554.

Wahba, G. (1983). Bayesian 'confidence intervals' for the cross-validated smoothing spline. *J. R. Statist. Soc. B 45*, 133–150.

Welsh, A., R. Cunningham, C. Donnelly, and D. Lindenmayer (1996). Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling 88*(1–3), 297–308.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Statist. Soc. B 62*, 413–428.

Wood, S. N. (2003). Thin plate regression splines. *J. R. Statist. Soc. B 65*, 95–114.

Wood, S. N. (2006). *Generalized Additive Models, An Introduction with R*. London: Chapman and Hall.

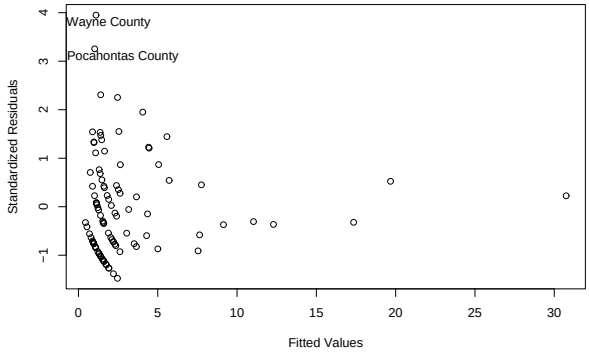

Fig. 10. Iowa Larynx Cancer Data: Model Diagnostics