Constrained Generalized Additive Models for Zero-Inflated Data

HAI LIU
Department of Statistics and Actuarial Science,
University of Iowa, Iowa City, Iowa 52242, U.S.A.
hai-liu@uiowa.edu

KUNG-SIK CHAN
Department of Statistics and Actuarial Science,
University of Iowa, Iowa City, Iowa 52242, U.S.A.
kung-sik-chan@uiowa.edu

Technical Report # 388 of Department of Statistics & Actuarial Science,
University of Iowa; First version Feb 2008; Second version March 2009

Summary

Zero-inflated data abound in ecological studies as well as in other scientific and quantitative fields. Nonparametric regression with zero-inflated response may be studied via the Zero-Inflated Generalized Additive Model (ZIGAM). ZIGAM assumes that the response variable follows a probabilistic mixture distribution of a zero atom and a regular component whose distribution belongs to some 1-parameter exponential family, where the zero atom explicitly accounts for zero-inflation. We propose the COnstrained Zero-Inflated Generalized Additive Model (COZIGAM) for analyzing zero-inflated data, with the further assumption that the probability of non-zero-inflation is some monotone function of the mean of the regular component. When the latter assumption obtains, the new approach provides a unified framework for modeling zero-inflated data, which is more parsimonious and efficient than the unconstrained ZIGAM. We develop an iterative algorithm for model estimation based on the penalized likelihood approach, and derive formulas for constructing confidence intervals of the maximum penalized likelihood estimator. Some asymptotic properties including the consistency of the regression function estimator and the limiting distribution of the parametric estimator are derived. We also propose a Bayesian model selection criterion for choosing between the unconstrained and constrained ZIGAMs. The new methods are illustrated with both simulated data and a real application.

Key Words: Asymptotic normality; Convergence rate; EM algorithm; Model selection; Penalized quasi-likelihood.

1 Introduction

Generalized additive models (GAMs) (Hastie and Tibshirani, 1990; Wood, 2006) are widely used in applied statistics, see, for instance, Ciannelli et al. (2007) and the references therein in ecological analysis. Penalized likelihood method provides a powerful tool for estimating GAMs, see Green and Silverman (1994), Wood (2000) and Gu (2002). In the GAM framework, the unknown smooth function components can be estimated by maximizing the penalized likelihood which generally equals

$$L(\eta) - \tilde{\lambda}^2 J^2(\eta) \tag{1}$$

where η is the unknown regression function on the link scale, $L(\eta)$ is the log-likelihood function, $J^2(\eta)$ is some roughness penalty, and $\tilde{\lambda}$ 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 $J^2(\eta) = \int \|D^k \eta\|^2$ where D^k is the k-th derivative operator with $k \geq 1$ as a fixed integer, and $\| \cdot \|^2$ denotes the square norm. This roughness measure will be adopted in the following discussions. 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 many basis functions (the number of which generally increases with sample size), see Wahba (1990) and Gu (2002). In particular, for $k = 2$, the maximizer is a smoothing spline, being natural cubic spline in the 1-dimensional case and thin-plate spline (Wood, 2003) in higher dimensional cases. These results extend to the case of GAM when the mean function is the sum of more than one component functions on the link scale, and form the basis of some approaches for empirical GAM analysis.

A common problem encountered in scientific 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. Zero-inflation also occurs in other fields, for example, in marketing where data on consumer choice may contain many non-purchase observations. Indeed, zero-inflated data abound in scientific and quantitative studies. These data are often analyzed via a two-component mixture model specifying that the distribution of the response variable belongs to a zero-inflated 1-parameter exponential family, that is, a probabilistic mixture of zero and a regular component whose distribution (to be referred below as the regular distribution) belongs to the 1-parameter exponential family; see Mullahy (1986), Lambert (1992), Heilbron (1994) and Lam et al. (2006). GAM has been generalized to include the zero-inflated exponential family (Barry and Welsh, 2002; Chiogna and Gaetan, 2007), which requires (i) linking a smooth function of the covariates, say $s_p(T)$ where T is the vector covariate, to the probability that the response is not zero-inflated (not from the zero atom), and (ii) linking another smooth function, say $s_\mu(T)$, to the mean of the (non-zero-inflated) 1-parameter exponential family distribution; the generalization will be referred to as the zero-inflated generalized additive model (ZIGAM).

The functional forms of the two smooth predictors $s_p(T)$ and $s_\mu(T)$ are generally unconstrained, because the zero-inflation process may be uncoupled from the process generating the (non-zero-inflated) data. However, for many ecological data, the zero-inflation process is coupled with the underlying population process. For example, in trawl survey studies, zero-inflation often arises from the spatio-temporal aggregation of fish due to their schooling behavior. For such data, the probability of positive catch is positively correlated to the volume occupied by the schools of fish which increases with the mean (local) abundance of the fish. On the other hand, some grasshopper species may suddenly change from solitary behavior to swarm behavior as their abundance greatly increases upon suitable environmental conditions. In the latter case, the probability of positive catch is a decreasing function of the mean (local) abundance of the locusts, over the transition stage from solitude to swarming. In sum, for survey data involving spatio-temporally aggregated subjects, the probability of positive catch is likely a function of the mean (local) abundance of the study population. Incorporating such a constraint in a ZIGAM reflects the mechanistic nature of the zero-inflation process, and promotes estimation efficiency by effectively reducing the degrees of freedom of the parameter space. Here, we implement this approach with the simplifying assumption that, on the link scales, the non-zero-inflation probability is a linear function of the conditional mean of the 1-parameter exponential family, that is, we assume that $s_p(T) = \alpha + \delta s_\mu(T)$ for some constants α and δ. This new model is referred to as the constrained zero-inflated generalized additive model (COZIGAM) below. A harbinger of our new approach is the ZIP(τ) model, proposed by Lambert (1992), which is a parametric zero-inflated Poisson regression model with the zero-inflation probability constrained to be proportional to the Poisson mean.

In practice, the validity of the constraint imposed by the COZIGAM needs to be assessed, which can be checked via model selection between an unconstrained ZIGAM and a COZIGAM. We derive a Bayesian model selection criterion, via Laplace approximation, for choosing between a ZIGAM and a COZIGAM. It is interesting to note that the proposed model selection criterion depends on the roughness penalty but otherwise does not depend on the explicit form of the prior.

Another approach to modeling zero-inflated data proceeds in two stages: (i) model the presence/absence pattern by a GAM and (ii) model the response given it is non-zero by another GAM (Barry and Welsh, 2002). For the case of a continuous regular distribution, the two approaches are equivalent, otherwise the two approaches are generally different. In stage (ii), the two-stage approach generally specifies the conditional response distribution given it is non-zero to belong to a zero-truncated 1-parameter exponential family, which requires more complex link functions for non-continuous regular distributions. We will not further pursue the two-stage approach.

The structure of this paper is as follows. We introduce the model formulation of the unconstrained ZIGAM and the COZIGAM, and propose an estimation algorithm based on the penalized likelihood approach in Section 2. Some large sample properties including the convergence rate of the maximum penalized likelihood estimator and the limiting distribution of the parametric part of the estimator are derived in Section 3, followed by some discussion on the computation of the observed Fisher information in order to assess the variability of the estimator. A Bayesian model selection criterion for choosing between the unconstrained and constrained ZIGAMs is derived in Section 4. Some Monte Carlo studies on the model estimation as well as the performance of the proposed model selection criterion will be discussed in Section 5. In Section 6, we illustrate the COZIGAM by a real example. We briefly conclude in Section 7.

2 Model Formulation and Penalized Likelihood Estimation

2.1 Zero-Inflated Generalized Additive Model

Let the data be $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ and the covariates be $\boldsymbol{T} = (T_1, T_2, \ldots, T_n)$ where Y_i are scalars and T_i are possibly high-dimensional vectors. Assume that given the covariate T_i, the Y_i's are independently distributed. Moreover, the marginal conditional distribution of Y_i depends on the covariates only through t_i, which is a mixture distribution given by

$$Y_i | T_i = t_i \sim h_i(y_i) = \begin{cases} 0 & \text{with probability } 1 - p_i \\ f(y_i|\vartheta_i) & \text{with probability } p_i, \end{cases} \tag{2}$$

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

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

where ω_i are some known constants, often equal to 1, and ϕ is the dispersion parameter. Then the expectation of Y_i under f is $\mu_i = E_f(Y_i) = \dot{b}(\vartheta_i)$ (for any function h, $\dot{h}$ denotes its first derivative and $\ddot{h}$ its second derivative) and variance $Var_f(Y_i) = \phi V(\mu_i) = \phi \ddot{b}(\vartheta_i)$. Below we will refer to f as the regular (exponential family) distribution, and μ_i as the regular mean.

The regular mean is assumed to be linked to some smooth function of the covariate:

$$g_\mu(\mu_i) = \eta(t_i),$$

where $g_\mu(\cdot)$ is the link function, and $\eta(\cdot)$ is some smooth function to be estimated by the penalized likelihood approach. (The extension to the case of replacing η by a sum of smooth functions with lower-dimensional arguments is straightforward.) The non-zero-inflation probability p_i is linked to the covariate as follows:

$$g_p(p_i) = \xi(t_i), \tag{3}$$

where $g_p(\cdot)$ is another link function, for instance, the logit function and $\xi(\cdot)$ is an unknown smooth function. If η and ξ are functionally orthogonal (infinite-dimensional) parameters, the model is an unconstrained zero-inflated GAM (ZIGAM) in which case zero-inflation could be caused by a mechanism different from that underlying the non-zero-inflated responses. On the other hand, if the zero-inflation process is coupled with the process generating the non-zero-inflated data, for example, data from surveys of spatio-temporally aggregated subjects, we may expect some relationship between η and ξ. In particular, we consider the case that ξ is constrained to be a linear function of η:

$$\xi = \alpha + \delta \cdot \eta, \tag{4}$$

where α and δ are some unknown parameters. Note that η must be a non-constant function over the support of the covariate, otherwise the model is non-identifiable because α and δ will then be non-unique. The constrained model assuming (4) adds only two more degrees of freedom to a GAM. Hence, when the constraint (4) obtains, the constrained ZIGAM provides a more parsimonious model that promotes estimation efficiency, as compared to the unconstrained ZIGAM whose estimation generally requires much larger sample size than the constrained ZIGAM; in particular, the estimation error of ξ of an unconstrained ZIGAM could be substantial for small to moderate samples. On the other hand, if the constraint does not hold, fitting a constrained ZIGAM introduces bias. Thus, it is important to assess the validity of (4), a task we shall return in Section 4. We will refer to the zero-inflated model with constraint (4) as the COnstrained Zero-Inflated Generalized Additive Model (COZIGAM). In Section 6, we will illustrate the use of COZIGAM by a real application. Note that if $\alpha = \infty$ and $\delta < \infty$, then the model is a GAM whereas, in the general case, the model is a zero-inflated GAM; if $|\alpha| < \infty$ and $\delta \equiv 0$, then the zero-inflation probability is the same across all sampling points, in which case the COZIGAM degenerates into a homogeneous ZIGAM.

To write the penalized log-likelihood function of the COZIGAM, first define the binary variables $E_i, i = 1, \dots, n$, with

$$E_i = \begin{cases} 1 & \text{if } Y_i \neq 0 \\ 0 & \text{if } Y_i = 0. \end{cases}$$

If the underlying regular exponential family distribution is continuous, for instance, Gaussian, the penalized log-likelihood then equals

$$l_p(\alpha, \delta, \eta) = \sum_{i=1}^{n} \Big[e_i \log\{p_i f(y_i|\vartheta_i)\} + (1 - e_i) \log (1 - p_i) \Big] - \tilde{\lambda}_n^2 J^2(\eta), \tag{5}$$

where η is an infinite-dimensional parameter, $\tilde{\lambda}_n$ is the smoothing parameter and $J^2(\eta)$ is the roughness penalty of η.

If the regular distribution assigns positive probability to zero, which is the case for many discrete distributions including Poisson and binomial, the penalized log-likelihood function becomes somewhat complex

$$l_p(\alpha, \delta, \eta) = \sum_{i=1}^{n} \Big[e_i \log p_i f(y_i|\vartheta_i) + (1 - e_i) \log (1 - p_i + p_i f(0|\vartheta_i)) \Big] - \tilde{\lambda}_n^2 J^2(\eta). \tag{6}$$

The complexity owes to the fact that a zero observation may result from the zero atom or the regular distribution. In some literature, the zeroes from the zero atom are called structural zeroes and those from the regular distribution are called sampling zeroes. If, however, the nature of the zero observations is known, the likelihood becomes simpler. This suggests the use of the EM algorithm (Dempster et al., 1977) for maximizing the penalized likelihood in (6), whose M-step admits an iterative algorithm with closed-form solutions. Augment the data by an indicator variable $\boldsymbol{Z} = (Z_1, \dots, Z_n)^T$ defined as follows

$$Z_i = \begin{cases} 1 & \text{if } Y_i \sim f(y_i|\vartheta_i) \\ 0 & \text{if } Y_i \sim 0. \end{cases} \tag{7}$$

The sequence $\{Z_i\}_{i=1,\cdots,n}$ is independently distributed, with the marginal distribution of Z_i being Bernoulli(p_i). The joint density of the complete data equals

$$f(\boldsymbol{y}, \boldsymbol{z}|\alpha, \delta, \eta) = \prod_{i=1}^{n} \{p_i f(y_i|\vartheta_i)\}^{z_i} \{(1-p_i)I(y_i = 0)\}^{1-z_i}, \tag{8}$$

and after dropping some constant term which does not depend on the unknown parameters, the complete-data penalized log-likelihood equals

$$l_p^c(\alpha, \delta, \eta) = \sum_{i=1}^{n} \Big[z_i \log\{p_i f(y_i|\vartheta_i)\} + (1 - z_i) \log(1 - p_i) \Big] - \tilde{\lambda}_n^2 J^2(\eta).$$

Estimation can be done by maximizing the above complete-data penalized log-likelihood, via an iterative algorithm detailed in Section 2.2. Below, a COZIGAM will be referred to as a continuous (discrete) COZIGAM if its penalized likelihood function is given by Equation (5) (Equation (6)).

2.2 Model Estimation

According to the reproducing kernel Hilbert space theory, under some mild conditions, the maximum penalized likelihood estimator of the smooth function η is a linear combination of some basis functions. More specifically, the functional value of η evaluated at t_i can be written as

$$\eta(t_i) = \boldsymbol{X}_i \boldsymbol{\beta},$$

where $\boldsymbol{X}_i$ is the i-th row of the design matrix $\boldsymbol{X}$ of the basis functions, and $\boldsymbol{\beta}$ is the parameter vector to be estimated. So without loss of generality, for a given set of covariates, we can reparametrize the infinite-dimensional parameter η as a finite-dimensional vector parameter $\boldsymbol{\beta}$ for the model estimation purpose. (The dimensionality may be further reduced by knot-based or principle component approximation.) Moreover, the penalty term $\tilde{\lambda}_n^2 J^2(\eta)$ can often be expressed as a quadratic form $\tilde{\lambda}_n^2 \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}/2$ where $\boldsymbol{S}$ is a penalty matrix, see Gu (2002) and Wood (2006). (In the case that η equals a sum of smooth functions, $\boldsymbol{S}$ is block diagonal, with each block submatrix corresponding to a smooth component and the smoothing-parameter multiplier being component-specific. For ease of exposition, we shall confine to the case that η is a single smooth function.) Hence the penalized likelihood functions of the continuous and discrete COZIGAMs become

$$l_p(\alpha, \delta, \boldsymbol{\beta}) = \sum_{i=1}^{n} \Big[e_i \log\{p_i f(y_i|\vartheta_i)\} + (1 - e_i) \log(1 - p_i) \Big] - \frac{1}{2} \tilde{\lambda}_n^2 \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}, \tag{5'}$$

and

$$l_p(\alpha, \delta, \boldsymbol{\beta}) = \sum_{i=1}^{n} \Big[e_i \log p_i f(y_i|\vartheta_i) + (1 - e_i) \log(1 - p_i + p_i f(0|\vartheta_i)) \Big] - \frac{1}{2} \tilde{\lambda}_n^2 \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}. \tag{6'}$$

We propose an iterative algorithm to find the maximizers of (5') and (6') with respect to the parameter $\boldsymbol{\theta} = (\alpha, \delta, \boldsymbol{\beta}^T)^T$.

The proposed algorithm for estimating the COZIGAM is motivated by the Penalized Iteratively Reweighted Least Squares (PIRLS) method (Wood, 2006, pg. 169-170) 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.

For a discrete COZIGAM, direct optimization of the penalized likelihood (6') is challenging because it complicates the use of generalized cross validation (GCV) or unbiased risk estimation (UBRE) for choosing the smoothing parameters. Instead, we develop an estimation scheme based on the EM algorithm. The EM algorithm relies on the fact that were the latent binary indicator variable $\boldsymbol{Z}$ defined by (7) available, the complete-data likelihood l_p^c is simpler, the optimization of which admits an iterative algorithm each step of which has a closed-form solution with known smoothing parameter, and, moreover, GCV or UBRE can be applied iteratively to determine the smoothing parameters; see Wood (2006, Chapter 4) for further discussions about GCV and UBRE. In particular, the optimization of the penalized likelihood can be implemented via the EM algorithm with $\boldsymbol{Z}$ as missing data. The penalized likelihood (5') of a continuous COZIGAM, however, can be maximized directly without the E-step, because $\boldsymbol{Z}$ is observable and coincides with the indicator $\boldsymbol{E}$. Throughout this section, the analysis will be done conditional on the observed values of the covariate t. For simplicity, the dependency on t is generally suppressed from the notations. For ease of exposition, the smoothing parameter is initially assumed known in the derivation below.

We first derive the conditional distribution of $\boldsymbol{Z}$ given the data. Write $f(y_i|\vartheta_i) = f(y_i)$. From the joint density of $(\boldsymbol{Y}, \boldsymbol{Z})$ given in (8), the conditional distribution of the components of $\boldsymbol{Z}$ given $\boldsymbol{Y}$ are independent with marginal conditional pdf

$$f(z_i|y_i;\boldsymbol{\theta}) = \frac{f(y_i, z_i|\boldsymbol{\theta})}{f(y_i|\boldsymbol{\theta})} = \frac{\{p_i f(y_i)\}^{z_i}\{(1-p_i)I(y_i=0)\}^{1-z_i}}{p_i f(y_i) + (1-p_i)I(y_i=0)}.$$

Therefore

$$Z_i|y_i;\boldsymbol{\theta} \sim \text{Bernoulli}\left(\frac{p_i f(y_i)}{p_i f(y_i) + (1-p_i)I(y_i=0)}\right).$$

Denote $\psi_i = E(Z_i|y_i;\boldsymbol{\theta}) = p_i f(y_i)/\left\{p_i f(y_i) + (1-p_i)I(y_i=0)\right\}$. Armed with these results, we can now state the EM algorithm for maximizing the penalized likelihood. Given the r-th parameter iterate $\boldsymbol{\theta}^{[r]}$, in the *E-step*, compute

$$\psi_i^{[r]} = E(Z_i|y_i, \boldsymbol{\theta}^{[r]}) = \frac{p_i^{[r]} f(y_i|\vartheta_i^{[r]})}{p_i^{[r]} f(y_i|\vartheta_i^{[r]}) + (1-p_i^{[r]})I(y_i=0)}.$$

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

$$E\{l_p^c(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}^{[r]}\} = \sum_{i=1}^{n}\left[\psi_i^{[r]}\log\{p_i f(y_i|\vartheta_i)\} + (1-\psi_i^{[r]})\log(1-p_i)\right] - \frac{1}{2}\tilde{\lambda}_n^2\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}.$$

Denote the above objective function as $\tilde{l}_p^c$. In the *M-step* we want to find the maximizer of $\tilde{l}_p^c$ with respect to the parameter $\boldsymbol{\theta}$. Taking the first derivatives of the objective function, we get

$$\frac{\partial \tilde{l}_p^c}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^n \frac{\psi_i^{[r]}(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} + \sum_{i=1}^n \frac{\psi_i^{[r]} - p_i}{p_i(1-p_i)} \frac{\partial p_i}{\partial \beta_j} - \tilde{\lambda}_n^2 [\boldsymbol{S}\boldsymbol{\beta}]_j, \quad j = 1, \ldots, K, \tag{9a}$$

$$\frac{\partial \tilde{l}_p^c}{\partial \alpha} = \sum_{i=1}^n \frac{\psi_i^{[r]} - p_i}{p_i(1-p_i)} \frac{1}{\dot{g}_p(p_i)}, \tag{9b}$$

$$\frac{\partial \tilde{l}_p^c}{\partial \delta} = \sum_{i=1}^n \frac{\psi_i^{[r]} - p_i}{p_i(1-p_i)} \frac{g_\mu(\mu_i)}{\dot{g}_p(p_i)}, \tag{9c}$$

where $\boldsymbol{\beta}$ is assumed to be K-dimensional. Equation (9) can be solved iteratively by modifying the PIRLS algorithm. The major obstacle for applying the PIRLS algorithm is that (9a) 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:

$$\frac{1}{\phi_1} \sum_{i=1}^n \frac{w_{1i}(y_{1i} - \mu_{1i})}{V_1(\mu_{1i})} \frac{\partial \mu_{1i}}{\partial \beta_j} + \frac{1}{\phi_2} \sum_{i=1}^n \frac{w_{2i}(y_{2i} - \mu_{2i})}{V_2(\mu_{2i})} \frac{\partial \mu_{2i}}{\partial \beta_j} - \tilde{\lambda}_n^2 [\boldsymbol{S}\boldsymbol{\beta}]_j = 0, \quad \text{for all } j,$$

where the two sums correspond to contributions from two GAMs with mean μ_{ki} linked to $\boldsymbol{X}_k\boldsymbol{\beta}$ by the link function g_k, and variance function V_k, $k = 1, 2$. However, these equations are exactly the optimality conditions for finding $\boldsymbol{\beta}$ that minimizes the following non-linear weighted least squares:

$$\mathcal{S}_p = \mathcal{S}_1 + \mathcal{S}_2 + \tilde{\lambda}_n^2 \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta},$$

where, for $k = 1, 2$,

$$\mathcal{S}_k = \sum_{i=1}^n \frac{w_{ki}(y_{ki} - \mu_{ki})^2}{\phi_k V_k(\mu_{ki})},$$

assuming the weights $V_1(\mu_1)$ and $V_2(\mu_2)$ were known and independent of $\boldsymbol{\beta}$.
The nonlinear least square problem can be solved iteratively. Let $\boldsymbol{\beta}^{[r]}$ be the r-th iterate of $\boldsymbol{\beta}$. Denote $\boldsymbol{\mu}_k^{[r]}$ as the value of $\boldsymbol{\mu}_k$ evaluated at $\boldsymbol{\beta}^{[r]}$. Defining diagonal matrices $\boldsymbol{V}_{k[r]}$ with the diagonal elements $V_{k[r]ii} = V_k(\mu_{ki}^{[r]})$, and the diagonal matrices $\boldsymbol{W}_k^*$ with $W_{kii}^* = w_{ki}/\phi_k$, $k = 1, 2$, we have

$$\mathcal{S}_k = \left\| \sqrt{\boldsymbol{V}_{k[r]}^{-1} \boldsymbol{W}_k^*} \, (\boldsymbol{y}_k - \boldsymbol{\mu}_k(\boldsymbol{\beta})) \right\|^2, \quad k = 1, 2$$

Next approximate $\boldsymbol{\mu}_k$ by its first order Taylor expansion around the r-th estimate $\boldsymbol{\beta}^{[r]}$. Hence,

$$\mathcal{S}_k \approx \left\| \sqrt{\boldsymbol{V}_{k[r]}^{-1} \boldsymbol{W}_k^*} \boldsymbol{G}_{k[r]}^{-1} \left(\boldsymbol{G}_{k[r]}(\boldsymbol{y}_k - \boldsymbol{\mu}_k^{[r]}) + \boldsymbol{\eta}_k^{[r]} - \boldsymbol{X}_k \boldsymbol{\beta} \right) \right\|^2, \quad k = 1, 2,$$

where $\boldsymbol{G}_{k[r]}$ is a diagonal matrix with elements $G_{k[r]ii} = \dot{g}_k(\mu_{ki}^{[r]})$. Furthermore, by defining the 'pseudodata'

$$z_{ki}^{[r]} = \dot{g}_k(\mu_{ki}^{[r]})(y_{ki} - \mu_{ki}^{[r]}) + \eta_{ki}^{[r]}$$

and the diagonal weight matrices $\boldsymbol{W}_k^{[r]}$ with elements

$$W_{kii}^{[r]} = \frac{w_{ki}}{\phi_k V_k(\mu_{ki}^{[r]}) \dot{g}_k^2(\mu_{ki}^{[r]})}$$

we have

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

Hence, at the r-th iteration,

$$\mathcal{S}_p \approx \left\| \sqrt{\boldsymbol{W}_1^{[r]}} \left(\boldsymbol{z}_1^{[r]} - \boldsymbol{X}_1 \boldsymbol{\beta} \right) \right\|^2 + \left\| \sqrt{\boldsymbol{W}_2^{[r]}} \left(\boldsymbol{z}_2^{[r]} - \boldsymbol{X}_2 \boldsymbol{\beta} \right) \right\|^2 + \tilde{\lambda}_n^2 \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta},$$

which can be readily combined into a single penalized sum of squares by appropriately defining $\boldsymbol{z}^{[r]}, \boldsymbol{X}$ and $\boldsymbol{W}^{[r]}$:

$$\mathcal{S}_p \approx \left\| \sqrt{\boldsymbol{W}^{[r]}} \left(\boldsymbol{z}^{[r]} - \boldsymbol{X} \boldsymbol{\beta} \right) \right\|^2 + \tilde{\lambda}_n^2 \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta},$$

the minimization of which yields the next iterate of $\boldsymbol{\beta}$. In the case of unknown smoothing parameter, it can be estimated, e.g., by minimizing the GCV or UBRE of the model corresponding to the preceding approximate weighted least squares; see Wood (2006, Chapter 4).

We apply this modified PIRLS algorithm for solving Equation (9a). After updating $\boldsymbol{\beta}$ 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 $\alpha + \delta\eta_i$ via the link function g_p where, given the current estimate of $\boldsymbol{\beta}$, $\eta_i = \boldsymbol{X}_i \boldsymbol{\beta}$ is known. The iteration can be repeated until all parameters converge according to some stopping criterion.

3 Asymptotic Properties and the Observed Information

3.1 Asymptotic Properties

In this section, we derive some asymptotic properties of the proposed penalized likelihood estimator of the COZIGAM. Our approach builds on earlier work on the asymptotics of semiparametric inference, see Mammen and van de Geer (1997) and van der Vaart (1998). For the theoretical analysis, recall the original parameter space $\Theta = \{\boldsymbol{\theta} = (\alpha, \delta, \eta)^T : \alpha, \delta \in \mathcal{R}, J(\eta) < \infty\}$, and assume that the roughness penalty takes the form $J^2(\eta) = \int \|D^k \eta\|^2$. Below, we derive the consistency and convergence rate of the estimator $\widehat{\boldsymbol{\theta}}_n$, as well as the asymptotic normality of $(\widehat{\alpha}_n, \widehat{\delta}_n)^T$ under suitable regularity conditions. For ease of exposition, we state the main results under the assumptions that (R1) the covariate T takes value in the unit interval $[0, 1]$ over which the probability density function is bounded away from 0, (R2) the function η is non-constant over $[0, 1]$, (R3) the data $\{(Y_i, T_i), i = 1, 2, \ldots, n\}$ are independent and identically distributed, (R4) the function $Q_1(y; \mu)$ (defined by Equation (15)) is a concave function of μ for every y, and (R5) any zero observation must come from the zero atom. Condition (R5) entails that we deal with the case of continuous COZIGAM. For discrete COZIGAM, the regularity conditions are more cumbersome, which will be discussed at the end of this section.

3.1.1 Notations and Further Assumptions

Reparametrize the smoothing parameter by $\lambda_n^2 = \tilde{\lambda}_n^2/n$, and denote $F_1 = g_\mu^{-1}$, $F_2 = g_p^{-1}$. Then,

$$\mu(t) = E(Y|T = t, Z = 1) = F_1(\eta(t)),$$
$$p(t) = E(Z|T = t) = F_2(\xi(t)).$$

The true parameter is denoted as $\boldsymbol{\theta}_0 = (\alpha_0, \delta_0, \eta_0)$ (so that the functions $\xi_0 = \alpha_0 + \delta_0 \cdot \eta_0$, $\mu_0 = F_1 \circ \eta_0$ and $p_0 = F_2 \circ \xi_0$, where $\circ$ denotes function composition.) Write $W = Y - \mu_0(T)$ and $R = Z - p_0(T)$, which are of zero mean.

Suppose that $f_1(x) = dF_1(x)/dx$ and $f_2(x) = dF_2(x)/dx$ exist for all $x \in \mathcal{R}$, and denote $l_1(x) = f_1(x)/V(F_1(x))$, $l_2(x) = f_2(x)/F_2(x)(1 - F_2(x))$, where recall V is the variance function of the regular distribution. Write $l_{10} = l_1 \circ \eta_0$, $f_{10} = f_1 \circ \eta_0$, $l_{20} = l_2 \circ \xi_0$, and $f_{20} = f_2 \circ \xi_0$. For any measurable function $a : \mathcal{R} \times [0, 1] \to \mathcal{R}$, $E(a(Y, T))$ denotes the expectation of $a(Y, T)$ under the true distribution and the following norms will be useful: $\|a\|^2 = Ea^2(Y, T)$, $\|a\|_n^2 = \frac{1}{n} \sum_{i=1}^n a^2(Y_i, T_i)$ and $|a|_\infty = \sup_{t \in [0,1]} |a(t)|$.

The asymptotic properties of the estimators depend on the smoothing parameter λ_n as $n \to \infty$. Recall the roughness functional $J(\eta)$ equals the integral of the squared k-th derivative of η. We assume that (R6) $\lambda_n = o_{\mathbf{P}}(n^{-1/4})$ and $1/\lambda_n = O_{\mathbf{P}}(n^{k/(2k+1)})$, and (R7) given T, W and R are uniformly *sub-Gaussian* (that is, for some constant $0 < C_0 < \infty$,

$$E_0 \left(\exp(W^2/C_0) | T \right) \leq C_0 \quad \text{almost surely},$$

and a similar inequality holds for R), as well as some technical conditions (R8)-(R14) listed in Appendix A. Note that assumptions (R6) and (R7) are standard conditions used in the literature, see Mammen and van de Geer (1997).

3.1.2 Main Results

Theorem 1. *Under conditions (R1) to (R13), it holds that*

$$J^2(\widehat{\eta}_n) = O_{\mathbf{P}}(1),$$
$$\|\widehat{\eta}_n - \eta_0\|_n = O_{\mathbf{P}}(\lambda_n), \tag{10}$$
$$|\widehat{\alpha}_n - \alpha_0| = O_{\mathbf{P}}(\lambda_n), \tag{11}$$
$$|\widehat{\delta}_n - \delta_0| = O_{\mathbf{P}}(\lambda_n).$$

The above convergence rate can be proved by adapting the arguments in Mammen and van de Geer (1997), and hence we omit the proof for saving space.

Before stating the asymptotic normality result for $(\widehat{\alpha}_n, \widehat{\delta}_n)^T$, we define the following two functions:

$$h_1(t) = -\frac{\delta_0 f_{20}(t) l_{20}(t)}{p_0(t) f_{10}(t) l_{10}(t) + \delta_0^2 f_{20}(t) l_{20}(t)},$$
$$h_2(t) = -\frac{\delta_0 \eta_0(t) f_{20}(t) l_{20}(t)}{p_0(t) f_{10}(t) l_{10}(t) + \delta_0^2 f_{20}(t) l_{20}(t)}.$$

Theorem 2. *Suppose that all conditions of Theorem 1 hold, and that assumption (R14) listed in Appendix A is valid. Moreover, assume that*

$$J(h_i) < \infty, \quad i = 1, 2, \tag{12}$$

$$p_0(t) f_{10}(t) l_{10}(t) + \delta_0^2 f_{20}(t) l_{20}(t) \neq 0, \quad \forall\ t \in [0, 1].$$

Then $\left(\sqrt{n}(\widehat{\alpha}_n - \alpha_0), \sqrt{n}(\widehat{\delta}_n - \delta_0)\right)^T$ *is asymptotically bivariate normal with zero mean and covariance matrix equal to* $\boldsymbol{A}^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}$ *with the elements of* $\boldsymbol{A}$ *given by*

$$\begin{aligned} a_{11} &= \left\| (m_0 p_0 f_{10} f_{20})^{1/2} \right\|^2, \\ a_{12} = a_{21} &= \left\| (m_0 p_0 \eta_0 f_{10} f_{20})^{1/2} \right\|^2, \\ a_{22} &= \left\| (m_0 p_0 f_{10} f_{20})^{1/2} \eta_0 \right\|^2, \end{aligned}$$

and $\boldsymbol{V}$ *equals the covariance matrix of* $m_0(T)\{R p_0(T) f_{10}(T) - \delta_0 W Z f_{20}(T)\}(1, \eta_0)^T$, *where*

$$m_0(t) = \frac{l_{10}(t) l_{20}(t)}{p_0(t) f_{10}(t) l_{10}(t) + \delta_0^2 f_{20}(t) l_{20}(t)}.$$

An outline of the proof is given in Appendix B; detailed proofs of Theorems 1 and 2 are given in Appendix E.

Finally, we remark that condition (R5) holds only for continuous COZIGAMs, but it can be relaxed by the following device. In a discrete COZIGAM, the conditional response distribution can be alternatively represented as a mixture of zero and a positive regular distribution. This can be done by merging the original zero atom with the zero realized under the original regular distribution, and redefine the regular distribution as the conditional regular distribution given that it is non-zero. This reparametrization, however, leads to much more complex regularity conditions. Theorems 1 and 2 can be generalized for discrete COZIGAMs under further assumptions including the boundedness of the parameter space, in which case the concavity condition (R4) can be dispensed with. Note that among discrete COZIGAMs, the zero-inflated Bernoulli model does not satisfy one of the regularity conditions, and it is not identifiable because the zero-truncated Bernoulli degenerates into a singleton. See Appendix E for the details and proofs.

3.2 Computing the Observed Information

Although we have shown the consistency of the maximum penalized likelihood estimator of the smooth function η, it is not useful for assessing the accuracy of the estimator. Even in the simplest case of a single spline regression function plus noise model, there is currently no theoretical results on the asymptotic distribution of the estimator of the smooth function. Because the maximum penalized likelihood estimator can be regarded as the Bayesian posterior mode, with the roughness penalty inducing an improper Gaussian prior (Gu, 2002), Wahba (1983) proposed to use the Bayesian confidence intervals for the smoothing splines, whose frequency properties was investigated by Nychka (1988), among other authors. We follow this approach of making use of the observed information matrix of the penalized likelihood to compute pointwise confidence intervals of the smooth functions, as well as computing the standard errors of $\widehat{\alpha}$ and $\widehat{\delta}$; this approach is preferred to computing their standard errors based on their complex forms of the asymptotic formula given in Theorem 2. (Note that the calculation is conditional on the finite basis functions chosen for the data on hand.) It can be shown that the two methods of computing the standard errors of $\widehat{\alpha}$ and $\widehat{\delta}$ are asymptotically equivalent for some simple cases; see Appendix E. The empirical performance of the confidence intervals so constructed will be studied in Section 5 by the Monte Carlo method.

Since the likelihood of a COZIGAM has an explicit form given by (5') or (6'), the computation of the information matrix can be readily done by computing the Hessian matrix, even though it is tedious. The covariance matrix of the estimator can be approximately computed by inverting the observed information matrix. Normal approximation of the sampling distribution of the estimators then yields a simple approach for constructing pointwise confidence intervals. The formulas for computing the observed information are listed in Appendix C.

4 Model Selection

In statistical analysis, one important issue is model selection or model comparison among multiple competing models. In this section, we introduce a Bayesian model selection criterion for selecting between the unconstrained ZIGAM and the COZIGAM. One of the widely used model selection criteria is the BIC, which selects the model with maximum posterior model probability. In the Bayesian framework and assuming constant prior model probabilities, the posterior probability of a model, say M_i, is proportional to the the marginal likelihood $P(D|M_i) = \int P(D|\boldsymbol{\theta}, M_i)P(\boldsymbol{\theta}|M_i)d\boldsymbol{\theta}$, where D denotes the data, $\boldsymbol{\theta}$ is the parameter under the model M_i. Therefore we will use the maximum marginal likelihood as the model selection criterion, as in the case of the BIC. The model with larger marginal likelihood will be preferred. For the unconstrained ZIGAM and the COZIGAM, there is generally no closed-form solution for the integral in the marginal likelihood. Laplace method (see, for example, Tierney and Kadane, 1986) will be used to approximately compute the marginal likelihood. Below, all marginal likelihood calculations are conditional on the finite basis functions chosen for the data on hand.

The penalized likelihood of the constrained model defined by (5') or (6') implicitly assumes a Gaussian prior with pdf $\propto \exp\left\{-\frac{1}{2}\tilde{\lambda}_n^2\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}\right\}$, which is generally improper. In Appendix D, we first consider the use of proper priors obtained by multiplying the preceding Gaussian prior by $\tilde{g}(\theta)$ and show that, under some mild regularity conditions on $\tilde{g}$, the logarithmic marginal likelihood is equal to

$$\log E \approx l_p(\widehat{\alpha}, \widehat{\delta}, \widehat{\boldsymbol{\beta}}) - \frac{K+2}{2}\log n - \frac{1}{2}\log\left|\boldsymbol{H}\right| + \frac{K+2-m}{2}\log 2\pi + \frac{1}{2}\log\left|\tilde{\lambda}_n^2\boldsymbol{S}_+\right|,$$

up to an error term of $O_{\mathbf{P}}(1)$, where $\widehat{\boldsymbol{\theta}} = (\widehat{\alpha}, \widehat{\delta}, \widehat{\boldsymbol{\beta}}^T)^T$ is the maximum penalized likelihood estimator, $K = \dim(\boldsymbol{\beta})$, $\boldsymbol{H}$ is the negative Hessian matrix of l_p/n evaluated at $\widehat{\boldsymbol{\theta}}$, and $\boldsymbol{S}_+$ is the submatrix in the penalty matrix $\boldsymbol{S}$ associated with the basis functions having positive roughness penalties, so that $\boldsymbol{S}_+$ is of full rank. Moreover, by passing to the limit, it is shown in Appendix D that the approximate formula holds also for the case of the improper Gaussian prior discussed above. Similarly, we can derive the Laplace approximation of the logarithmic marginal likelihood of an unconstrained ZIGAM, see Appendix D.

5 Simulation Results

In this section we first examine the empirical coverage probabilities of the confidence intervals constructed via the observed information. Then we report some simulation results on the success rates of the proposed Bayesian model selection criterion for choosing between the unconstrained ZIGAM and the COZIGAM. All simulation results reported below are based on 1000 replications.

Assuming asymptotic normality for the estimators, pointwise confidence intervals can be readily constructed for each parameter and the smooth functions. The confidence 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, for example, GCV or UBRE which was adopted in our simulation study. The simulation results reported below suggest that the omission of the variability in the smooth parameter seems to have asymptotically negligible effects on the coverage rate of the confidence intervals.

The simulations are based on two test functions, denoted by s_1 and s_2, which are taken from Wood (2006, pg. 197). The test function s_1 has a 1-dimensional argument, while s_2 has a 2-dimensional argument (see Figure 1).

$$s_1(t) = 0.2t^{11}(10(1-t))^6 + 10(10t)^3(1-t)^{10}, \quad 0 \leq t \leq 1$$

$$s_2(t_1, t_2) = 0.3 \times 0.4\pi \left\{1.2e^{-(t_1-0.2)^2/0.3^2-(t_2-0.3)^2} + 0.8e^{-(t_1-0.7)^2/0.3^2-(t_2-0.8)^2/0.4^2}\right\}, \ 0 \leq t_1, t_2 \leq 1.$$

Both Gaussian and Poisson regular distributions are considered. Responses from the regular distributions are generated so that, on the link scale, the regular means equal the test functions after some rescaling. Zero-inflation occurs at a rate that is proportional to the regular mean on the link scale. The smoothing parameter is chosen by the GCV for Gaussian regular distribution and UBRE for Poisson regular distribution, as explained in Section 2.2. The fitting results for two sets of simulated zero-inflated Poisson count data are shown in Figure 1. Notice that the plots are on the link scale.

We examined the performance of confidence intervals by checking the Average Coverage Probability (ACP), as was discussed by Wahba (1983) and Gu (2002). The ACP is defined as the coverage rate over the sampling points as follows.

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

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

Table 1: Model Estimation and Coverage Probabilities

	Mean $\widehat{\alpha}$	SD $\widehat{\alpha}$	Cov. $\widehat{\alpha}$	Mean $\widehat{\delta}$	SD $\widehat{\delta}$	Cov. $\widehat{\delta}$	ACP$_c$ $\widehat{\eta}$	ACP$_u$ $\widehat{\eta}$	ACP$_u$ $\widehat{\xi}$
					Gaussian				
True	-0.500		0.950	0.600		0.950	0.950	0.950	0.950
s_1									
$n = 100$	-0.532	0.545	0.948	0.626	0.344	0.962	0.926	0.955	0.892
$n = 200$	-0.541	0.375	0.951	0.614	0.236	0.956	0.914	0.948	0.874
s_2									
$n = 200$	-0.535	0.500	0.955	0.612	0.246	0.961	0.939	0.980	0.921
$n = 400$	-0.515	0.343	0.946	0.601	0.169	0.953	0.943	0.974	0.928
					Poisson				
True	-0.500		0.950	1.000		0.950	0.950	0.950	0.950
s_1									
$n = 200$	-0.535	0.470	0.959	1.053	0.372	0.958	0.941	0.898	0.792
$n = 400$	-0.513	0.320	0.958	1.015	0.252	0.947	0.940	0.906	0.851
s_2									
$n = 400$	-0.681	0.398	0.945	1.137	0.307	0.943	0.954	0.944	0.812
$n = 600$	-0.651	0.318	0.937	1.111	0.244	0.942	0.952	0.952	0.871

The simulation results show that the empirical coverage probabilities are very close to the nominal levels in each case with different regular distributions and test functions. The test data are highly zero-inflated and about 30% to 50% of the responses are zeroes, with the true constraint coefficients $\alpha_0 = -0.5$ and $\delta_0 = 0.6$ in the Gaussian case and $\alpha_0 = -0.5$ and $\delta_0 = 1.0$ in the Poisson case. For both test functions, with increasing sample size, the bias of the estimators decreases. There is a tendency of overestimating the parameter δ which is the slope of the proportional constraint. A similar problem arises in a simulation study with Lambert's ZIP(τ) model reported by Lambert (1992); the ZIP(τ) model is a special case of the COZIGAM in the parametric zero-inflated Poisson regression setting with the intercept term α fixed to be 0. We also fit (unconstrained) ZIGAMs to the data and compare the ACPs of the smooth functions from both models (ACP$_c$ $\widehat{\eta}$ and ACP$_u$ $\widehat{\eta}$ denote the ACP of η for the constrained and unconstrained models respectively, and similarly defined is ACP$_u$ $\widehat{\xi}$). We find that the discrepancy between the ACPs and the nominal 95% level is generally greater for a ZIGAM than a COZIGAM. The discrepancy is greatest for the function estimate of ξ based on the unconstrained ZIGAM, as its estimation is based on the presence/absence data. This confirms that fitting a COZIGAM gains efficiency when the constraint obtains. Note that the estimator of α has larger variability than that of δ, which is expected because both test functions are non-negative, rendering lesser information for estimating the intercept term.

These simulation studies suggest that the observed information matrix provided adequate approximation for assessing the variability in the estimator. Furthermore, the

Table 2: Proportions of Choosing the True Model

s_1	Regular Distribution	$n = 100$	$n = 200$	$n = 300$
Unconstrained Model	Gaussian	0.538	0.755	0.857
	Poisson	0.753	0.899	0.954
Constrained Model	Gaussian	0.808	0.863	0.963
	Poisson	0.776	0.866	0.936
s_2		$n = 300$	$n = 400$	$n = 500$
Unconstrained Model	Gaussian	0.772	0.924	0.983
	Poisson	0.932	0.962	0.983
Constrained Model	Gaussian	0.960	0.931	0.926
	Poisson	0.881	0.961	0.977

simulation results lend support to the result that

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

see Wahba (1983). We remark that this frequency property may not hold for estimating the functional value at a specific point but it seems to hold on average across the sampling points.

Table 2 shows that the proposed Bayesian model selection criterion correctly chooses the true models with very high proportions in several different situations. In general the proportions increase with the sample size within each case. For the COZIGAMs, the data generating processes are same as above. For the (unconstrained) ZIGAMs, the non-zero-atom probability is linked to $\xi(t) = 2\sin(\pi t) - 1$ for the model using test function s_1 and linked to $\xi(t_1, t_2) = 2t_1 - t_2^2$ in the model using s_2 as the test function, both via the logistic link function.

6 A Real Application: 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 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 in the year 1987 which contain 274 observations sampled from the 93th to the 116th Julian day over sites with bottom depth ranging 28-5200m. Among the 274 observations, 84 are zeroes, which make up over 30% of the data. The main goal is to explore the spatial patterns of pollock spawning aggregations in the GOA. Pollock egg density is the response variable, and the covariates include longitude, latitude, and (log-transformed) bottom depth. (Preliminary analysis suggests that the covariate Julian day does not enter the model due to the relatively short period of the sampling dates in this year, and hence not included in the analysis.) 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

$$\mu = c + s(lon, lat) + s(\log(depth)), \tag{13}$$

and

$$\text{logit}(p) = \alpha + \delta \cdot \mu, \tag{14}$$

where c, α, δ are parameters, each function s in (13) is assumed to be a distinct smooth function; for model identifiability, the smooth functions are constrained to be of zero mean and hence the corresponding function estimates are centered over the data.

Under the above model assumptions, the regression function specified by (13) with constraint (14) may be estimated by fitting a COZIGAM. We have also fitted an unconstrained ZIGAM to the data. Using the model selection criterion developed in Section 4, the logarithmic marginal likelihood of the unconstrained model equals -464.24, whereas that of the COZIGAM equals -455.88. Thus it provides some justification for choosing the COZIGAM over the unconstrained model.

Figure 2 shows the estimated smooth functions of the location and bottom depth effects based on the COZIGAM fitted with all data. The estimated smooth functions give the spatial density distribution of pollock egg in the GOA in the year 1987 and the density seemed to be more concentrated over deeper areas than shallower areas. The estimated parametric coefficients in Equation (14) are $\widehat{\alpha} = -1.816$ (0.347) and $\widehat{\delta} = 0.489$ (0.064) which is significantly positive. Thus, there is strong evidence indicating that zero-inflation is more likely to occur at locations with less egg density.

The validity of the fitted COZIGAM may be explored based on the residuals for the cases with non-zero catch. The model diagnostic plots including the Q-Q normal score plot of these residuals and the plot of residuals vs. fitted values (unreported) suggest that the model assumptions for the positive data are generally valid. Therefore the log-normal regression assumption is reasonable according to the model diagnostics.

7 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. Some large sample properties of the maximum penalized likelihood estimator including consistency and asymptotic normality have been proved. We propose a Bayesian model selection criterion for choosing between the unconstrained ZIGAM and the COZIGAM. The new methods are illustrated with both simulation studies and a real example with application to the pollock egg density data analysis.

So far we have considered imposing proportional constraints (on the link scales) on the non-zero-inflation probability in the COZIGAM. An interesting problem is to relax the proportional constraint to allow possibly different proportionality coefficients for different covariates for the zero-inflation process. Another future problem is to relax the linear constraint to other more general constraints. Currently not much is known about the limiting distributions of the smooth components even in simple cases such as a spline function plus noise model. These are some interesting directions for future work.

Acknowledgements

The authors thank Lorenzo Ciannelli and Luke Tierney for helpful suggestions. We gratefully acknowledge partial support from the US National Science Foundation (CMG-0620789) and the North Pacific Research Board.

References

Barry, S. C. and Welsh, A. H. (2002). Generalized additive modelling and zero inflated count data. *Ecological Modelling* **157,** 179–188.

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

Chiogna, M. and Gaetan, C. (2007). Semiparametric zero-inflated poisson models with application to animal abundance studies. *Environmetrics* **18,** 303–314.

Ciannelli, L., Bailey, K., Chan, K. S., and Stenseth, N. C. (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., Fauchald, P., Chan, K. S., Agostini, V. N., and Dingsør, G. E. (2007). Spatial fisheries ecology: recent progress and future prospects. *Journal of Marine Systems* **71,** 223–236.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (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 Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

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

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

Heilbron, D. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal* **36,** 531–547.

Lam, K. F., Xue, H. Q., and Cheung, Y. B. (2006). Semiparametric analysis of zero-inflated count data. *Biometrics* **62,** 996–1003.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* **34,** 1–14.

Mammen, E. and van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *The Annals of Statistics* **25,** 1014–1035.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33,** 341–365.

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

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association* **83,** 1134–1143.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6,** 461–464.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **18,** 82–86.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

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

Wahba, G. (1990). *Spline Models for Observational Data.* Volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia, SIAM.

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.* Chapman and Hall, London.

Appendix A: Regularity Conditions Needed For Theorems 1 And 2

The proof of the convergence rate and asymptotic normality makes heavy use of the empirical process theory (see Pollard (1984) and the references therein). We first list the conditions used in the proof of the consistency and convergence rate: for some constants $0 < C_1, C_2, C_3, C_4 < \infty$,

$$V(s) \geq 1/C_1, \quad \forall\ s \in F_1(\mathcal{R}), \tag{R8}$$

$$1/C_2 \leq |\bar{l}_1(x)| \leq C_2, \quad \forall\ x \in \mathcal{R}, \tag{R9}$$

$$|f_2(x)| \leq C_3, \quad \forall\ x \in \mathcal{R}, \tag{R10}$$

and

$$1/C_4 \leq F_2(\xi_0(t)) \leq 1 - 1/C_4, \quad \forall\ t \in [0, 1], \tag{R11}$$

where $\bar{l}_1(x) = f_1(x)/V(\bar{F}_1(x))$ and $\bar{F}_1(x) = \{F_1(x) + F_1(\eta_0)\}/2$. Also, for some constants $0 < C_5, h, b < \infty$ and for all $t \in [0, 1]$, we have

$$|f_2(x)| \geq 1/C_5, \quad \text{for all } |x - \xi_0| \leq h, \tag{R12}$$

together with an identifiability condition

$$\inf_{\|\xi - \xi_0\| > b} \|F_2(\xi) - F_2(\xi_0)\| > 0, \quad \text{for all } b > 0. \tag{R13}$$

Furthermore, we have the assumption

$$f_i \text{ and } l_i, i = 1, 2, \text{are bounded and Lipschitz continuous functions.} \tag{R14}$$

Appendix B: Proof of Theorem 2

We now list some key steps for proving the asymptotic normality result stated in Theorem 2. It follows from Theorem 1 and because T has a probability density that is bounded away from zero over its support, it can be shown by similar arguments as in Mammen and van de Geer (1997) that we can, without loss of generality, assume that the unknown functions η and ξ satisfy the condition that for some constants $0 < d_1, d_2 < \infty$,

$$|\eta - \eta_0|_\infty \leq d_1 \quad \text{and} \quad |\xi - \xi_0|_\infty \leq d_2.$$

The proof exploits some properties of the penalized quasi-likelihood estimation method. Define the quasi-(log-)likelihood functions

$$Q_1(y; \mu) = \int_y^\mu \frac{(y - s)}{V(s)} ds, \tag{15}$$

$$Q_2(z; p) = \int_z^p \frac{(z - t)}{t(1 - t)} dt.$$

The quasi-likelihood equals

$$Q(\alpha,\delta,\eta)=\bar{Q}_{1n}(F_1(\eta))+\bar{Q}_{2n}(F_2(\alpha+\delta\eta)),$$

and hence the penalized quasi-likelihood becomes

$$Q_p(\alpha,\delta,\eta)=Q(\alpha,\delta,\eta)-\lambda_n^2J^2(\eta), \tag{16}$$

where

$$\begin{aligned}\bar{Q}_{1n}(\mu)&=\frac{1}{n}\sum_{i=1}^{n}Z_iQ_1(Y_i;\mu(T_i)),\\ \bar{Q}_{2n}(p)&=\frac{1}{n}\sum_{i=1}^{n}Q_2(Z_i;p(T_i)).\end{aligned}$$

Then the penalized quasi-likelihood (PQL) estimator equals

$$\widehat{\boldsymbol{\theta}}_n=\arg\max_{\boldsymbol{\theta}\in\Theta}Q_p(\alpha,\delta,\eta).$$

We then perturb $\widehat{\boldsymbol{\theta}}_n$ along two paths $s\mapsto\widehat{\boldsymbol{\theta}}_{ns}$,

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^{I}_{ns}&=\widehat{\boldsymbol{\theta}}_{ns}+s(1,0,h_1)^T\\ \widehat{\boldsymbol{\theta}}^{II}_{ns}&=\widehat{\boldsymbol{\theta}}_{ns}+s(0,1,h_2)^T,\end{aligned}$$

for some measurable functions h_1, h_2 (to be determined below) and $s\in\mathcal{R}$. Thus,

$$\begin{aligned}\frac{d}{ds}\left\{Q(\widehat{\boldsymbol{\theta}}^{I}_{ns})-\lambda_n^2J^2(\hat{\eta}^{I}_{ns})\right\}\Big|_{s=0}&=0\ ,\\ \frac{d}{ds}\left\{Q(\widehat{\boldsymbol{\theta}}^{II}_{ns})-\lambda_n^2J^2(\hat{\eta}^{II}_{ns})\right\}\Big|_{s=0}&=0\ ,\end{aligned}$$

where $\hat{\eta}^{I}_{ns}=\hat{\eta}_n+sh_1$ and $\hat{\eta}^{II}_{ns}=\hat{\eta}_n+sh_2$. By routine analysis, it can be shown that

$$\begin{aligned}\frac{d}{ds}\lambda_n^2J^2(\hat{\eta}^{I}_{ns})\Big|_{s=0}&\leq 2\lambda_n^2J(\hat{\eta}_n)J(h_1)=o_{\mathbf{P}}(n^{-1/2}), \qquad (17)\\ \frac{d}{ds}\lambda_n^2J^2(\hat{\eta}^{II}_{ns})\Big|_{s=0}&\leq 2\lambda_n^2J(\hat{\eta}_n)J(h_2)=o_{\mathbf{P}}(n^{-1/2}).\end{aligned}$$

It remains to choose h_1 and h_2 such that both $dQ(\widehat{\boldsymbol{\theta}}^{I}_{ns})/ds\big|_{s=0}$ and $dQ(\widehat{\boldsymbol{\theta}}^{II}_{ns})/ds\big|_{s=0}$ admit a linear stochastic expansion in terms of $\hat{\alpha}_n-\alpha_0$ and $\hat{\delta}_n-\delta_0$, specifically

$$\begin{aligned}\frac{d}{ds}Q(\widehat{\boldsymbol{\theta}}^{I}_{ns})\Big|_{s=0}=&\ \frac{1}{n}\sum\{Z_iW_il_{10}(T_i)h_1(T_i)+R_il_{20}(T_i)(1+\delta_0h_1(T_i))\}\\ &-(\hat{\alpha}_n-\alpha_0)\frac{1}{n}\sum f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i))\\ &-(\hat{\delta}_n-\delta_0)\frac{1}{n}\sum\eta_0(T_i)f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i))\\ &-\frac{1}{n}\sum(\hat{\eta}_n(T_i)-\eta_0(T_i))\big\{Z_if_{10}(T_i)l_{10}(T_i)h_1(T_i)\\ &+\delta_0f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i))\big\}\\ &+o_{\mathbf{P}}(n^{-1/2}),\end{aligned} \tag{18}$$

which can be shown to hold if

$$h_1(t) = -\frac{\delta_0 f_{20}(t) l_{20}(t)}{p_0(t) f_{10}(t) l_{10}(t) + \delta_0^2 f_{20}(t) l_{20}(t)},$$

using similar techniques employed in the proof of Theorem 2.4 in Mammen and van de Geer (1997). Combining (17) and (18), we obtain

$$\begin{aligned} 0 &= \frac{1}{n}\sum \{Z_i W_i l_{10}(T_i) h_1(T_i) + R_i l_{20}(T_i)(1 + \delta_0 h_1(T_i))\} \\ &\quad -(\hat{\alpha}_n - \alpha_0)\frac{1}{n}\sum f_{20}(T_i) l_{20}(T_i)(1 + \delta_0 h_1(T_i)) \\ &\quad -(\hat{\delta}_n - \delta_0)\frac{1}{n}\sum \eta_0(T_i) f_{20}(T_i) l_{20}(T_i)(1 + \delta_0 h_1(T_i)) \\ &\quad + o_{\mathbf{P}}(n^{-1/2}). \end{aligned} \tag{19}$$

Similarly, for the second path, if we let

$$h_2(t) = -\frac{\delta_0 \eta_0(t) f_{20}(t) l_{20}(t)}{p_0(t) f_{10}(t) l_{10}(t) + \delta_0^2 f_{20}(t) l_{20}(t)},$$

then

$$\begin{aligned} 0 &= \frac{1}{n}\sum \{Z_i W_i l_{10}(T_i) h_2(T_i) + R_i l_{20}(T_i)(\eta_0(T_i) + \delta_0 h_2(T_i))\} \\ &\quad -(\hat{\alpha}_n - \alpha_0)\frac{1}{n}\sum f_{20}(T_i) l_{20}(T_i)(\eta_0(T_i) + \delta_0 h_2(T_i)) \\ &\quad -(\hat{\delta}_n - \delta_0)\frac{1}{n}\sum \eta_0(T_i) f_{20}(T_i) l_{20}(T_i)(\eta_0(T_i) + \delta_0 h_2(T_i)) \\ &\quad + o_{\mathbf{P}}(n^{-1/2}). \end{aligned} \tag{20}$$

The asymptotic normality result stated in Theorem 2 follows from (19) and (20) and the fact that both $\boldsymbol{A}$ and $\boldsymbol{V}$ can be readily checked to be finite matrices and that $\boldsymbol{A}$ is non-singular.

Appendix C: Formulas for Computing the Observed Information

The observed information matrix is given by

$$\boldsymbol{I_\theta} = -\frac{\partial^2 l_p}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\left.\begin{pmatrix} \frac{\partial^2 l_p}{\partial\alpha^2} & \frac{\partial^2 l_p}{\partial\alpha\partial\delta} & \frac{\partial^2 l_p}{\partial\alpha\partial\boldsymbol{\beta}^T} \\ \frac{\partial^2 l_p}{\partial\delta\partial\alpha} & \frac{\partial^2 l_p}{\partial\delta^2} & \frac{\partial^2 l_p}{\partial\delta\partial\boldsymbol{\beta}^T} \\ \frac{\partial^2 l_p}{\partial\boldsymbol{\beta}\partial\alpha} & \frac{\partial^2 l_p}{\partial\boldsymbol{\beta}\partial\delta} & \frac{\partial^2 l_p}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \end{pmatrix}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \tag{21}$$

For a continuous COZIGAM, let $\boldsymbol{\rho}$ and $\boldsymbol{\tau}$ be $n \times 1$ vectors with components $\rho_i = \frac{e_i - p_i}{\dot{g}_p(p_i)p_i(1-p_i)}$, and $\tau_i = \frac{e_i(y_i - \mu_i)}{\dot{g}_\mu(\mu_i)\phi V(\mu_i)}$, $i = 1, \ldots, n$. To simplify the notations, in some of the following derivations we will suppress the arguments in the functions $g_\mu(\mu_i)$, $V(\mu_i)$, $g_p(p_i)$ and their derivatives. By routine analysis, we have

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

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

Then each element in Equation (21) can be evaluated via the following formulas $\frac{\partial^2 l_p}{\partial \alpha^2} = \mathbf{1}^T \boldsymbol{G_\rho} \mathbf{1}$, $\frac{\partial^2 l_p}{\partial \delta^2} = \boldsymbol{\eta}^T \boldsymbol{G_\rho} \boldsymbol{\eta}$, $\frac{\partial^2 l_p}{\partial \alpha \partial \delta} = \mathbf{1}^T \boldsymbol{G_\rho} \boldsymbol{\eta}$, $\frac{\partial^2 l_p}{\partial \boldsymbol{\beta} \partial \alpha} = \delta \boldsymbol{X}^T \boldsymbol{G_\rho} \mathbf{1}$, $\frac{\partial^2 l_p}{\partial \boldsymbol{\beta} \partial \delta} = \delta \boldsymbol{X}^T \boldsymbol{G_\rho} \boldsymbol{\eta} + \boldsymbol{X}^T \boldsymbol{\rho}$, and $\frac{\partial^2 l_p}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \boldsymbol{X}^T \boldsymbol{G_\tau} \boldsymbol{X} + \delta^2 \boldsymbol{X}^T \boldsymbol{G_\rho} \boldsymbol{X} - \tilde{\lambda}_n^2 \boldsymbol{S}$, where $\boldsymbol{G_\tau}$ and $\boldsymbol{G_\rho}$ are two diagonal matrices with elements $G_{\tau ii} = \frac{\dot{\tau}_i}{\dot{g}_\mu(\mu_i)}$ and $G_{\rho ii} = \frac{\dot{\rho}_i}{\dot{g}_p(p_i)}$ respectively, and $\boldsymbol{\eta} = (\eta(t_1), \ldots, \eta(t_n))^T$.

For a discrete COZIGAM, define $\boldsymbol{\rho}^*$ and $\boldsymbol{\tau}^*$ to be $n \times 1$ vectors whose components equal to $\rho_i^* = \frac{e_i - p_i(1-f_i)}{\dot{g}_p p_i(1-p_i+p_i f_i)}$, and $\tau_i^* = \frac{\{e_i(1-p_i)+p_i f_i\}(y_i-\mu_i)}{(1-p_i+p_i f_i)\dot{g}_\mu \phi V}$ respectively. If $E_i = 1$, $\frac{\partial \tau_i^*}{\partial \mu_i} = \frac{\partial \rho_i^*}{\partial p_i} = -1$ and $\frac{\partial \tau_i^*}{\partial p_i} = \frac{\partial \rho_i^*}{\partial \mu_i} = 0$. Otherwise, if $E_i = 0$, denote $f_i^0 = f(Y_i = 0|\mu_i)$ and $A_i = 1 - p_i + p_i f_i^0$, then we have

$$\frac{\partial \tau_i^*}{\partial \mu_i} = -\frac{p_i}{\phi \dot{g}_\mu V A_i}\left\{f_i^0 - \frac{\mu_i f_i^0\left(\ddot{g}_\mu V + \dot{g}_\mu \dot{V}\right)}{\dot{g}_\mu V} + (1-p_i)\mu_i \dot{f}_i^0 / A_i\right\},$$

$$\frac{\partial \tau_i^*}{\partial p_i} = -\frac{\mu_i f_i^0}{\phi \dot{g}_\mu V A_i^2},$$

and

$$\frac{\partial \rho_i^*}{\partial \mu_i} = \frac{\dot{f}_i^0}{\dot{g}_p A_i^2},$$

$$\frac{\partial \rho_i^*}{\partial p_i} = \frac{\left(1 - f_i^0\right)\left\{\ddot{g}_p A_i - \dot{g}_p(1 - f_i^0)\right\}}{\dot{g}_p^2 A_i^2}.$$

Let $\boldsymbol{G_{\tau\mu}}$, $\boldsymbol{G_{\tau p}}$, $\boldsymbol{G_{\rho\mu}}$, and $\boldsymbol{G_{\rho p}}$ be four diagonal matrices with elements on the leading diagonals $G_{\tau\mu ii} = \frac{1}{\dot{g}_\mu(\mu_i)}\frac{\partial \tau_i^*}{\partial \mu_i}$, $G_{\tau p ii} = \frac{1}{\dot{g}_p(p_i)}\frac{\partial \tau_i^*}{\partial p_i}$, $G_{\rho\mu ii} = \frac{1}{\dot{g}_\mu(\mu_i)}\frac{\partial \rho_i^*}{\partial \mu_i}$, and $G_{\rho p ii} = \frac{1}{\dot{g}_p(p_i)}\frac{\partial \rho_i^*}{\partial p_i}$ respectively. Then the second derivatives in (21) are given as follows: $\frac{\partial^2 l_p}{\partial \alpha^2} = \mathbf{1}^T \boldsymbol{G_{\rho p}} \mathbf{1}$, $\frac{\partial^2 l_p}{\partial \delta^2} = \boldsymbol{\eta}^T \boldsymbol{G_{\rho p}} \boldsymbol{\eta}$, $\frac{\partial^2 l_p}{\partial \alpha \partial \delta} = \mathbf{1}^T \boldsymbol{G_{\rho p}} \boldsymbol{\eta}$, $\frac{\partial^2 l_p}{\partial \boldsymbol{\beta} \partial \alpha} = \boldsymbol{X}^T \left(\boldsymbol{G_{\rho\mu}} + \delta \boldsymbol{G_{\rho p}}\right) \mathbf{1}$, $\frac{\partial^2 l_p}{\partial \boldsymbol{\beta} \partial \delta} = \boldsymbol{X}^T \left(\boldsymbol{G_{\rho\mu}} + \delta \boldsymbol{G_{\rho p}}\right) \boldsymbol{\eta} + \boldsymbol{X}^T \boldsymbol{\rho}^*$, and $\frac{\partial^2 l_p}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \boldsymbol{X}^T \left(\boldsymbol{G_{\tau\mu}} + \delta \boldsymbol{G_{\tau p}} + \delta \boldsymbol{G_{\rho\mu}} + \delta^2 \boldsymbol{G_{\rho p}}\right) \boldsymbol{X} - \tilde{\lambda}_n^2 \boldsymbol{S}$.

Appendix D: Justification of the Laplace Approximation

For a COZIGAM, partition the parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T)^T$, where, after relabelling the basis functions if needed, $\boldsymbol{\beta}_0$ corresponds to the basis functions belonging to the null space of the roughness penalty, and $\boldsymbol{\beta}_1$ corresponds to those with positive penalties.

Denote $\boldsymbol{S}_+$ as the submatrix in the penalty matrix $\boldsymbol{S}$ associated with $\boldsymbol{\beta}_1$. Therefore $\boldsymbol{S}_+$ is of full rank and $\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta} = \boldsymbol{\beta}_1^T\boldsymbol{S}_+\boldsymbol{\beta}_1$, and hence the implicit Gaussian prior with pdf $\propto \exp\left\{-\frac{1}{2}\tilde{\lambda}_n^2\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}\right\}$ is improper whenever $\boldsymbol{\beta}_0$ is present. To bypass the problem of an improper prior in the calculation of the marginal likelihood, we first consider the case with a proper prior and derive the marginal likelihood by Laplace approximation. Specifically, consider the following prior density

$$g(\alpha, \delta, \boldsymbol{\beta}) = \frac{\left|\tilde{\lambda}_n^2\boldsymbol{S}_+\right|^{1/2}}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}\tilde{\lambda}_n^2\boldsymbol{\beta}_1^T\boldsymbol{S}_+\boldsymbol{\beta}_1\right\}\tilde{g}(\boldsymbol{\theta}),$$

where $m = \dim(\boldsymbol{\beta}_1)$, $\tilde{g}(\boldsymbol{\theta})$ is some prior information on the other parameters, with the assumption that it is continuous, bounded and locally bounded away from zero, which is similar to the condition imposed on the prior by Schwarz (1978). For example, $\tilde{g}(\boldsymbol{\theta})$ may represent some proper Gaussian prior on $\boldsymbol{\beta}_0, \alpha$ and δ. It follows from (5') and (6') that the marginal likelihood of the COZIGAM equals

$$\int p(\boldsymbol{y}|\alpha, \delta, \boldsymbol{\beta})g(\alpha, \delta, \boldsymbol{\beta})d\alpha d\delta d\boldsymbol{\beta} = \frac{\left|\tilde{\lambda}_n^2\boldsymbol{S}_+\right|^{1/2}}{(2\pi)^{m/2}}\int \exp\left\{n\bar{l}_p(\alpha, \delta, \boldsymbol{\beta})\right\}\tilde{g}(\boldsymbol{\theta})d\alpha d\delta d\boldsymbol{\beta},$$

where $\bar{l}_p(\alpha, \delta, \boldsymbol{\beta}) = \frac{1}{n}l_p(\alpha, \delta, \boldsymbol{\beta})$ is the normalized penalized log-likelihood.

Note that $\boldsymbol{\theta} = (\alpha, \delta, \boldsymbol{\beta}^T)^T$ and assume that with probability going to 1 as $n \to \infty$, $\forall\, \varepsilon > 0$, $\exists\; 0 < \varrho < \exp\left\{\bar{l}_p(\widehat{\boldsymbol{\theta}})\right\}$, such that $\widehat{\Theta} = \left\{\boldsymbol{\theta} : \exp\left\{\bar{l}_p(\boldsymbol{\theta})\right\} > \varrho\right\} \subseteq D_\varepsilon = \left\{\boldsymbol{\theta} : \left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right\| \leq \varepsilon\right\}$, where $\widehat{\boldsymbol{\theta}}$ is the point at which $\bar{l}_p(\boldsymbol{\theta})$ attains its global maximum. This assumption is similar to the condition of well-separated point of maximum discussed by van der Vaart (1998), and it holds if $\bar{l}_p(\boldsymbol{\theta})$ is strictly concave, which is the case for unconstrained zero-inflated Gaussian models with canonical links. For other cases, local concavity may still hold but is not easy to prove because of the complexity of model specification, where more future work needs to be done. By employing similar arguments in the proof of Lemma 2 in Schwarz (1978), we can show that, as $n \to \infty$,

$$\log\int \exp\left\{n\bar{l}_p(\boldsymbol{\theta})\right\}\tilde{g}(\boldsymbol{\theta})d\boldsymbol{\theta} = \log\int_{\widehat{\Theta}} \exp\left\{n\bar{l}_p(\boldsymbol{\theta})\right\}\tilde{g}(\boldsymbol{\theta})d\boldsymbol{\theta} + o_{\mathbf{P}}(1). \tag{22}$$

The Taylor expansion of $\bar{l}_p(\boldsymbol{\theta})$ over D_ε gives $\bar{l}_p(\boldsymbol{\theta}) = \bar{l}_p(\widehat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T\boldsymbol{H}(\widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})$, where $\widetilde{\boldsymbol{\theta}}$ lies between $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$, so that $\left\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right\| \leq \varepsilon$. Hence the integral on the right hand side of Equation (22) becomes

$$\exp\left\{n\bar{l}_p(\widehat{\boldsymbol{\theta}})\right\}\int_{\widehat{\Theta}} \exp\left\{-\frac{n}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T\boldsymbol{H}(\widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\right\}\tilde{g}(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

Notice that we can and shall choose ϱ to make $\widetilde{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}$ sufficiently close so that $\log\left|\boldsymbol{H}(\widetilde{\boldsymbol{\theta}})\right| = \log\left|\boldsymbol{H}(\widehat{\boldsymbol{\theta}})\right| + o_{\mathbf{P}}(1)$. Moreover, the difference $\boldsymbol{H}(\widetilde{\boldsymbol{\theta}}) - \boldsymbol{H}(\widehat{\boldsymbol{\theta}})$ does not depend on the smoothing parameter. Then, from the properties of $\tilde{g}(\boldsymbol{\theta})$ and using similar arguments as in

Schwarz (1978), it is not difficult to show that

$$\begin{aligned}&\log\int_{\widehat{\Theta}}\exp\left\{-\frac{n}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^T\boldsymbol{H}(\widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\right\}\tilde{g}(\boldsymbol{\theta})d\boldsymbol{\theta}\\ =&\ \frac{K+2}{2}\log 2\pi-\frac{K+2}{2}\log n-\frac{1}{2}\log\left|\boldsymbol{H}(\widehat{\boldsymbol{\theta}})\right|+O_{\mathbf{P}}(1).\end{aligned}$$

Therefore, the logarithmic marginal likelihood is equal to

$$\log E=l_p(\widehat{\boldsymbol{\theta}})-\frac{K+2}{2}\log n-\frac{1}{2}\log\left|\boldsymbol{H}(\widehat{\boldsymbol{\theta}})\right|+\frac{K+2-m}{2}\log 2\pi+\frac{1}{2}\log\left|\tilde{\lambda}_n^2\boldsymbol{S}_+\right|+O_{\mathbf{P}}(1).$$

Notice that this approximation holds uniformly for continuous $\tilde{g}$ that are uniformly bounded, and uniformly locally bounded away from 0. Hence, the above approximate logarithmic marginal likelihood does not depend on the explicit form of $\tilde{g}$, and so it holds for the original improper Gaussian prior density by a limiting argument.

Similarly, for an unconstrained ZIGAM, the smooth function ξ in (3) is functionally orthogonal to η. Write the functional value of ξ as

$$\xi(t_i)=\boldsymbol{M}_i\boldsymbol{\gamma},$$

where $\boldsymbol{M}_i$ is the i-th row of the design matrix $\boldsymbol{M}$ of ξ and $\boldsymbol{\gamma}$ is the parameter vector. Also the roughness penalty of ξ could be written as a quadratic form in $\boldsymbol{\gamma}$. Then the penalized log-likelihood of the unconstrained continuous and discrete ZIGAM equal

$$l_p(\boldsymbol{\beta},\boldsymbol{\gamma})=\sum_{i=1}^{n}\Big[e_i\log\{p_if(y_i|\vartheta_i)\}+(1-e_i)\log{(1-p_i)}\Big]-\frac{1}{2}\tilde{\lambda}_{1n}^2\boldsymbol{\beta}^T\boldsymbol{S}_1\boldsymbol{\beta}-\frac{1}{2}\tilde{\lambda}_{2n}^2\boldsymbol{\gamma}^T\boldsymbol{S}_2\boldsymbol{\gamma},$$

and

$$l_p(\boldsymbol{\beta},\boldsymbol{\gamma})=\sum_{i=1}^{n}\Big[e_i\log p_if(y_i|\vartheta_i)+(1-e_i)\log{(1-p_i+p_if(0|\vartheta_i))}\Big]-\frac{1}{2}\tilde{\lambda}_{1n}^2\boldsymbol{\beta}^T\boldsymbol{S}_1\boldsymbol{\beta}-\frac{1}{2}\tilde{\lambda}_{2n}^2\boldsymbol{\gamma}^T\boldsymbol{S}_2\boldsymbol{\gamma},$$

respectively, where $\boldsymbol{S}_1$, $\boldsymbol{S}_2$ are two penalty matrices, and $\tilde{\lambda}_{1n}^2$, $\tilde{\lambda}_{2n}^2$ are the smoothing parameters, corresponding to η and ξ respectively. Let the joint prior density of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ equal

$$g(\boldsymbol{\beta},\boldsymbol{\gamma})=\frac{\left|\boldsymbol{S}_{1+}\right|^{1/2}}{(2\pi)^{m_1/2}}\exp\left\{-\frac{1}{2}\tilde{\lambda}_{1n}^2\boldsymbol{\beta}^T\boldsymbol{S}_1\boldsymbol{\beta}\right\}\frac{\left|\boldsymbol{S}_{2+}\right|^{1/2}}{(2\pi)^{m_2/2}}\exp\left\{-\frac{1}{2}\tilde{\lambda}_{2n}^2\boldsymbol{\gamma}^T\boldsymbol{S}_2\boldsymbol{\gamma}\right\}\tilde{g}(\boldsymbol{\theta}),$$

where $\boldsymbol{S}_{i+}, i=1,2$ is an $m_i\times m_i$ nonsingular submatrix of $\boldsymbol{S}_i$ associated with the basis functions having positive roughness penalties. The marginal likelihood of the unconstrained ZIGAM equals

$$\int p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma})g(\boldsymbol{\beta},\boldsymbol{\gamma})\tilde{g}d\boldsymbol{\beta}d\boldsymbol{\gamma}=\frac{\left|\tilde{\lambda}_{1n}^2\boldsymbol{S}_{1+}\right|^{1/2}\left|\tilde{\lambda}_{2n}^2\boldsymbol{S}_{2+}\right|^{1/2}}{(2\pi)^{(m_1+m_2)/2}}\int\exp\left\{n\bar{l}_p(\boldsymbol{\beta},\boldsymbol{\gamma})\right\}\tilde{g}(\boldsymbol{\theta})d\boldsymbol{\beta}d\boldsymbol{\gamma}.$$

By Laplace approximation,

$$\begin{aligned}\log E = l_p(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) - \frac{K_1 + K_2}{2} \log n - \frac{1}{2} \log \left| \boldsymbol{H(\widehat{\beta}, \widehat{\gamma})} \right| \\ + \frac{K_1 + K_2 - (m_1 + m_2)}{2} \log 2\pi + \frac{1}{2} \log \left| \tilde{\lambda}_{1n}^2 \boldsymbol{S}_{1+} \right| + \frac{1}{2} \log \left| \tilde{\lambda}_{2n}^2 \boldsymbol{S}_{2+} \right| + O_{\mathbf{P}}(1),\end{aligned}$$

where $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ is the maximum penalized likelihood estimator, $K_1 = \dim(\boldsymbol{\beta})$, $K_2 = \dim(\boldsymbol{\gamma})$. The negative Hessian matrix $H(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ can be computed similarly as in the case of COZIGAM, and hence omitted.

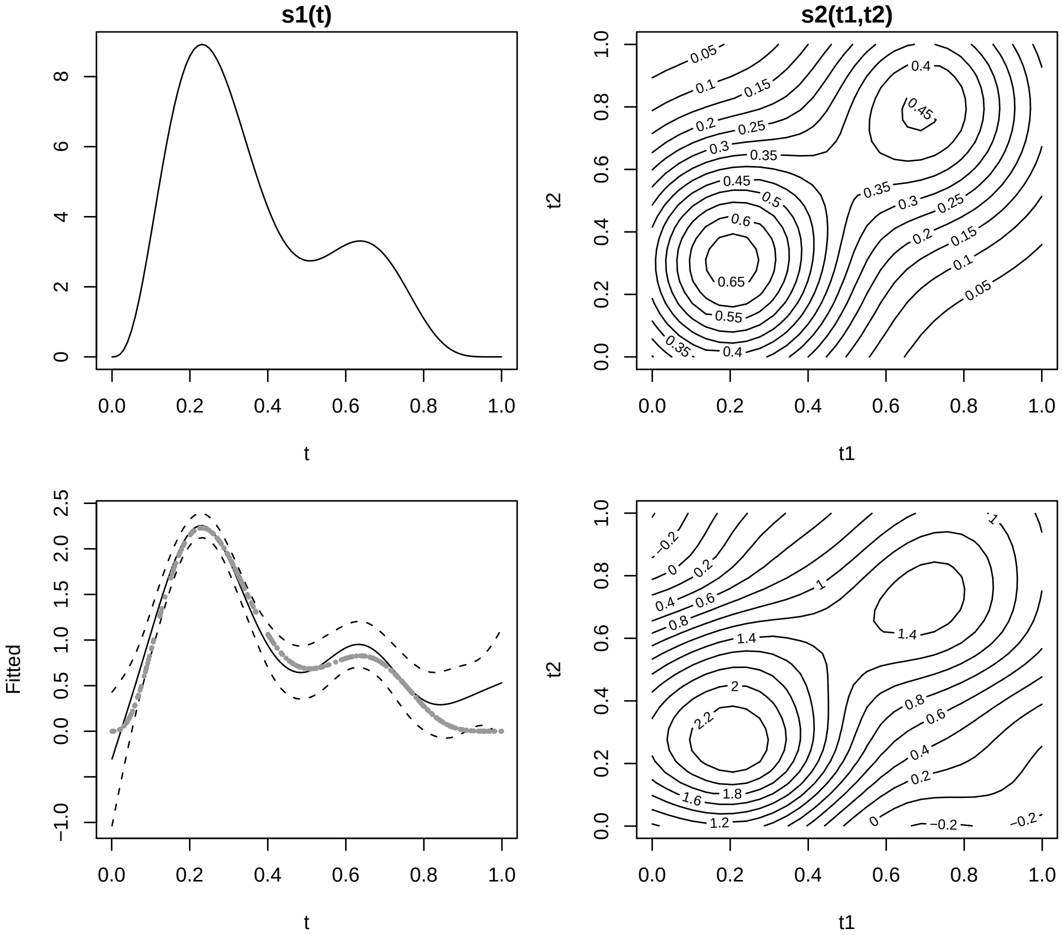

Figure 1: Simulation Study. The upper two plots display the true test functions. Model fits of two simulated series are listed in the lower two panels: the lower left panel depicts an estimate of the test function $s_1/4$ with true $\alpha_0 = -0.5$, $\delta_0 = 1.0$ and sample size $n = 300$, whose estimated values are $\widehat{\alpha} = -0.658$ (0.367), $\widehat{\delta} = 1.112$ (0.299). The gray dots are the true functional values, the black line is the estimated function, and the dashed lines are the 95% pointwise confidence bands; The lower right panel displays an estimate of the test function $3.5s_2$ with true $\alpha_0 = -0.5$, $\delta_0 = 1.0$ and sample size $n = 500$, whose estimated values are $\widehat{\alpha} = -0.404$ (0.319), $\widehat{\delta} = 1.021$ (0.246).

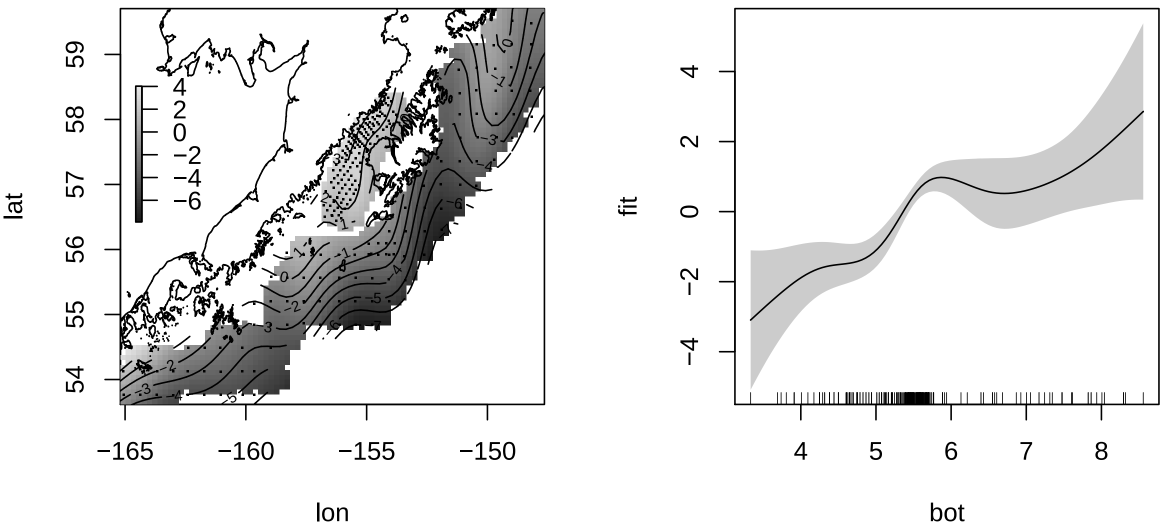

Figure 2: Effects of Location and Bottom Depth: the left diagram shows the contour plot of $s(lon, lat)$ on the right side of Equation (13); the right diagram depicts the bottom depth effect $s(\log(depth))$ with 95% pointwise confidence band.

Appendix E: Further Asymptotic Results and Detailed Technical Proofs

As we mentioned before, in the case that (R5) does not hold, that is, the regular distribution from the exponential family admits zero with positive probability (e.g. Poisson, binomial), we can still derive the consistency and limiting distribution of the estimator by adapting and augmenting the regularity conditions. Let's first state two theorems parallel to the results of Theorem 1 and 2.

Denote $f_0(\mu) = Pr(Y = 0|\mu)$ as the probability mass at zero under the regular exponential family distribution with mean μ. Let

$$E_i = \begin{cases} 1 & \text{if } Y_i \neq 0 \\ 0 & \text{if } Y_i = 0. \end{cases}$$

Then the incomplete-data log-likelihood equals

$$l = \sum_{i=1}^{n} E_i \log p_i f(y_i|\vartheta_i) + (1 - E_i) \log (1 - p_i + p_i f(0|\vartheta_i))$$

which could be rewritten as

$$l = \sum_{i=1}^{n} E_i \log \left(\frac{f(y_i|\vartheta_i)}{1 - f(0|\vartheta_i)} \right) + \sum_{i=1}^{n} E_i \log (p_i(1 - f(0|\vartheta_i))) + (1 - E_i) \log (1 - p_i + p_i f(0|\vartheta_i)),$$

where the first summand corresponds to the log-likelihood of another exponential family with mean $\mu^* = \mu/(1 - f_0(\mu))$. The second summand is exactly the log-likelihood of a Bernoulli process with probability of success equal to $p^* = p(1 - f_0(\mu))$. Let

$$\mu^* = \frac{F_1(\eta)}{1 - f_0(F_1(\eta))} := F_1^*(\eta)$$

and

$$p^* = F_2(\xi)\{1 - f_0(F_1(\eta))\}.$$

In order to prove the consistency result, we assume that the parameter space of $(\alpha, \delta)^T$ to be bounded. Specifically, let the parameter space be $\Theta^* = \left\{\boldsymbol{\theta} = (\alpha, \delta, \eta)^T : |\alpha|, |\delta|, |\eta|_\infty \leq C, J(\eta) < \infty\right\}$. The penalized likelihood estimator equals

$$\widehat{\boldsymbol{\theta}}_n = \arg \max_{\boldsymbol{\theta} \in \Theta^*} \left[\bar{Q}_{1n}\{F_1^*(\eta)\} + \bar{Q}_{2n}\{F_2(\xi)(1 - f_0(F_1(\eta)))\} - \lambda_n^2 J^2(\eta) \right],$$

with only a slight change in the form of $Q_1(y; \mu)$ by replacing V by V^*, where $V^*(\mu^*) = V(\mu)(d\mu^*/d\mu)$. And

$$V^*(s) \geq 1/C_1', \quad \forall\ s \in F_1^*(\mathcal{R}). \tag{R8'}$$

Then the convergence rate of the estimator can be verified via Lemma 3.1 in Mammen and van de Geer (1997) under conditions (R1)-(R3), (R6), (R8'), and the sub-Gaussianality of $W^* = Y^* - \mu_0^*(T)$, $R^* = E - p_0^*(T)$ where Y^* denotes a random variable whose distribution equals the conditional distribution of Y given $Y \neq 0$, $\mu_0^* = F_1^*(\eta_0)$, and $p_0^* = F_2(\xi_0)\{1 - f_0(F_1(\eta_0))\}$. Notice that condition (R8') ensures that the zero-inflated model is identifiable and excludes the situation where the regular non-zero-inflated exponential family distribution is Bernoulli ($V^* = 0$).

Theorem 3. *Under conditions (R1)-(R3), (R6), (R8'), and the sub-Gaussianality of W^* and R^*, and assuming the parameter space Θ^*, we have*

$$\begin{aligned} J(\hat{\eta}_n) &= O_{\mathbf{P}}(1), \\ \|\hat{\eta}_n - \eta_0\|_n &= O_{\mathbf{P}}(\lambda_n), \\ \|\hat{\xi}_n - \xi_0\|_n &= O_{\mathbf{P}}(\lambda_n), \end{aligned}$$

and

$$\begin{aligned} |\hat{\alpha}_n - \alpha_0| &= O_{\mathbf{P}}\left(\lambda_n\right), \\ |\hat{\delta}_n - \delta_0| &= O_{\mathbf{P}}\left(\lambda_n\right). \end{aligned}$$

For proving the asymptotic normality, let $f_1^*(x) = dF_1^*(x)/dx$, $\zeta(x) = 1 - f_0(F_1(x))$,$\zeta_0 = \zeta \circ \eta_0$ ($\hat{\zeta}_n = \zeta \circ \hat{\eta}_n$), and $\dot{\zeta}(x) = d\zeta/dx$. Also write $F_{20} = F_2 \circ \xi_0$, $\hat{p}_n^* = F_2(\hat{\xi}_n)\hat{\zeta}_n$.

Assume that

$$f_1^*, l_1^*, \zeta, F_2\dot{\zeta}/v, \text{and } f_2\zeta/v \text{ are bounded and Lipschitz continuous functions.} \qquad \text{(R1')}$$

where $l_1^*(x) = f_1^*(x)/V(F_1(x))$, $v(\boldsymbol{\theta}) = F_2(\xi)\zeta(\eta)\{1 - F_2(\xi)\zeta(\eta)\}$, and $v_0 = v(\boldsymbol{\theta}_0)$. Let $w_0 = \delta_0 f_{20}\zeta_0 + F_{20}\dot{\zeta}_0$, $l_{10}^* = l_1^* \circ \eta_0$, and $f_{10}^* = f_1^* \circ \eta_0$. Choose the directions of perturbation as

$$\begin{aligned} h_1^* &= -\frac{w_0 f_{20}\zeta_0}{p_0^* f_{10}^* l_{10}^* v_0 + w_0^2}, \\ h_2^* &= -\frac{w_0 \eta_0 f_{20}\zeta_0}{p_0^* f_{10}^* l_{10}^* v_0 + w_0^2}. \end{aligned}$$

Theorem 4. *Assume that the results in Theorem 3 hold with $\lambda_n = o_{\mathbf{P}}(n^{-1/4})$, and assumption (R1') is valid. Moreover, assume that*

$$J(h_i^*) < \infty, \quad i = 1, 2,$$

$$p_0^*(t) f_{10}^*(t) l_{10}^*(t) v_0 + w_0^2 \neq 0, \quad \forall\ t \in [0, 1].$$

Then $\left(\sqrt{n}(\hat{\alpha}_n - \alpha_0), \sqrt{n}(\hat{\delta}_n - \delta_0)\right)^T$ is asymptotically bivariate normal with zero mean and covariance matrix equal to $\boldsymbol{A}^{*-1}\boldsymbol{V}^*\boldsymbol{A}^{*-1}$ with the elements of $\boldsymbol{A}^*$ given by

$$\begin{aligned} a_{11}^* &= \left\| (m_0^* p_0^* f_{10}^* f_{20}\zeta_0)^{1/2} \right\|^2, \\ a_{12}^* &= a_{21}^* = \left\| (m_0^* p_0^* f_{10}^* f_{20}\zeta_0\eta_0)^{1/2} \right\|^2, \\ a_{22}^* &= \left\| (m_0^* p_0^* f_{10}^* f_{20}\zeta_0)^{1/2}\,\eta_0 \right\|^2. \end{aligned}$$

and $\boldsymbol{V}^*$ equals the covariance matrix of $m_0^*(T)\{R^* p_0^*(T) f_{10}^*(T) - W^* E w_0(T)\}(1, \eta_0)^T$, where

$$m_0^*(t) = \frac{l_{10}^*(t) f_{20}(t)\zeta_0(t)}{p_0^* f_{10}^*(t) l_{10}^*(t) v_0(t) + w_0^2(t)}.$$

Proof of the Convergence Rate

Proof of Theorem 1.

Let $\bar{F}_1(\hat{\eta}_n) = \{F_1(\hat{\eta}_n) + F_1(\eta_0)\}/2 = (\hat{\mu}_n + \mu_0)/2$. We have

$$\begin{aligned}
\bar{Q}_{1n}(\bar{F}_1(\hat{\eta}_n)) - \bar{Q}_{1n}(F_1(\eta_0)) &= \bar{Q}_{1n}\left(\frac{F_1(\hat{\eta}_n) + F_1(\eta_0)}{2}\right) - \bar{Q}_{1n}(F_1(\eta_0)) \\
&\geq \frac{1}{2}\left\{\bar{Q}_{1n}(F_1(\hat{\eta}_n)) + \bar{Q}_{1n}(F_1(\eta_0))\right\} - \bar{Q}_{1n}(F_1(\eta_0)) \\
&= \frac{1}{2}\left\{\bar{Q}_{1n}(F_1(\hat{\eta}_n)) - \bar{Q}_{1n}(F_1(\eta_0))\right\}
\end{aligned}$$

where the inequality holds because of the concavity of $Q_1(y;\cdot)$. For a fixed y_0, write

$$\gamma_\eta = \int_{y_0}^{\bar{F}_1(\eta)} \frac{1}{V(s)} ds,$$

and $\hat{\gamma}_n = \gamma_{\hat{\eta}_n}$, $\gamma_0 = \gamma_{\eta_0}$. Then

$$\begin{aligned}
\bar{Q}_{1n}(\bar{F}_1(\hat{\eta}_n)) - \bar{Q}_{1n}(F_1(\eta_0)) &= \frac{1}{n}\sum_{i=1}^n Z_i \int_{F_1(\eta_0)}^{\bar{F}_1(\hat{\eta}_n)} \frac{(Y_i - s)}{V(s)} ds \\
&= \frac{1}{n}\sum_{i=1}^n Z_i \int_{F_1(\eta_0)}^{\bar{F}_1(\hat{\eta}_n)} \frac{W_i}{V(s)} ds - \frac{1}{n}\sum_{i=1}^n Z_i \int_{F_1(\eta_0)}^{\bar{F}_1(\hat{\eta}_n)} \frac{s - \mu_0(T_i)}{V(s)} ds \\
&= \frac{1}{n}\sum_{i=1}^n W_i Z_i\left(\hat{\gamma}_n(T_i) - \gamma_0(T_i)\right) - \frac{1}{n}\sum_{i=1}^n Z_i \int_{\mu_0}^{\frac{1}{2}(\hat{\mu}_n + \mu_0)} \frac{s - \mu_0(T_i)}{V(s)} ds
\end{aligned}$$

For $\gamma = \int_{y_0}^{\frac{1}{2}(\mu+\mu_0)} V(s)^{-1} ds$, it can be readily checked that

$$\begin{aligned}
\frac{d}{d\gamma}\int_{\mu_0}^{\frac{1}{2}(\mu+\mu_0)} \frac{s-\mu_0}{V(s)} ds &= \frac{1}{2}(\mu - \mu_0) \\
\frac{d^2}{d\gamma^2}\int_{\mu_0}^{\frac{1}{2}(\mu+\mu_0)} \frac{s-\mu_0}{V(s)} ds &= V\left(\frac{\mu+\mu_0}{2}\right)
\end{aligned}$$

So by mean value theorem,

$$\int_{\mu_0}^{\frac{1}{2}(\hat{\mu}_n+\mu_0)} \frac{s - \mu_0(T_i)}{V(s)} ds = \frac{1}{2} V\left(\frac{\tilde{\mu}+\mu_0}{2}\right)(\hat{\gamma}_n - \gamma_0)^2$$

where $\tilde{\mu}$ is some value between $\hat{\mu}_n$ and μ_0. Therefore, by Cauchy-Schwartz inequality and

(R8),

$$
\begin{aligned}
\bar{Q}_{1n}(\bar{F}_1(\hat{\eta}_n)) - \bar{Q}_{1n}(F_1(\eta_0)) &\leq \frac{1}{n}\sum_{i=1}^{n} W_i Z_i \left(\hat{\gamma}_n(T_i) - \gamma_0(T_i)\right) \\
&\quad - \frac{1}{2C_1}\frac{m}{n}\left\{\frac{1}{m}\sum_{\{i:Z_i=1\}} (\hat{\gamma}_n(T_i) - \gamma_0(T_i))^2\right\} \\
&\leq \frac{m}{n}\left\{\frac{1}{m}\sum_{\{i:Z_i=1\}} W_i(\hat{\gamma}_n(T_i) - \gamma_0(T_i))\right\} - \frac{1}{2C_1}\frac{m}{n}\|\hat{\gamma}_n - \gamma_0\|_m^2 \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} W_i^2\right)^{1/2}\sqrt{\frac{m}{n}}\|\hat{\gamma}_n - \gamma_0\|_m - \frac{1}{2C_1}\frac{m}{n}\|\hat{\gamma}_n - \gamma_0\|_m^2,
\end{aligned}
$$

where $m = \sum_{i=1}^{n} Z_i$ as the number of observations coming from the regular response distribution. Because of (R11), $n/m = O_{\mathbf{P}}(1)$ as $n \to \infty$.

We now develop a similar inequality For Q_2. Notice that $Q_2(z;p) = z\log p + (1 - z)\log(1-p)$, which is the exact log-likelihood of a Bernoulli random variable. Define $\bar{F}_2(\xi) = \{F_2(\xi) + F_2(\xi_0)\}/2$. By the concavity of the log-function, it is easy to verify that

$$
\begin{aligned}
\bar{Q}_{2n}(\bar{F}_2(\hat{\xi}_n)) - \bar{Q}_{2n}(F_2(\xi_0)) &= \frac{1}{n}\sum_{i=1}^{n} Z_i \log\left(\frac{\bar{F}_2(\hat{\xi}_n(T_i))}{F_2(\xi_0(T_i))}\right) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(1 - Z_i)\log\left(\frac{1 - \bar{F}_2(\hat{\xi}_n(T_i))}{1 - F_2(\xi_0(T_i))}\right) \\
&\geq \frac{1}{2}\left\{\bar{Q}_{2n}(F_2(\hat{\xi}_n)) - \bar{Q}_{2n}(F_2(\xi_0))\right\}
\end{aligned}
$$

On the other hand, because $\log(x) = 2\log(\sqrt{x}) \leq 2(\sqrt{x} - 1)$,

$$
\begin{aligned}
\bar{Q}_{2n}(\bar{F}_2(\hat{\xi}_n)) - \bar{Q}_{2n}(F_2(\xi_0)) &\leq \frac{2}{n}\sum_{i=1}^{n} Z_i \left(\sqrt{\frac{\bar{F}_2(\hat{\xi}_n(T_i))}{F_2(\xi_0(T_i))}} - 1\right) \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}(1 - Z_i)\left(\sqrt{\frac{1 - \bar{F}_2(\hat{\xi}_n(T_i))}{1 - F_2(\xi_0(T_i))}} - 1\right) \\
&\leq \frac{2}{n}\sum_{i=1}^{n}\frac{R_i}{\sqrt{F_2(\xi_0(T_i))}}\left\{\sqrt{\bar{F}_2(\hat{\xi}_n(T_i))} - \sqrt{F_2(\xi_0(T_i))}\right\} \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\frac{-R_i\left\{\sqrt{1 - \bar{F}_2(\hat{\xi}_n(T_i))} - \sqrt{1 - F_2(\xi_0(T_i))}\right\}}{\sqrt{1 - F_2(\xi_0(T_i))}} \\
&\quad - \left\|\sqrt{\bar{F}_2(\hat{\xi}_n)} - \sqrt{F_2(\xi_0)}\right\|_n^2 \\
&\quad - \left\|\sqrt{1 - \bar{F}_2(\hat{\xi}_n)} - \sqrt{1 - F_2(\xi_0)}\right\|_n^2
\end{aligned}
$$

By the definition of the maximum penalized likelihood estimator and Equation (16), we have

$$\bar{Q}_{1n}(F_1(\hat{\eta}_n)) + \bar{Q}_{2n}(F_2(\hat{\xi}_n)) - \lambda_n^2 J^2(\hat{\eta}_n) \geq \bar{Q}_{1n}(F_1(\eta_0)) + \bar{Q}_{2n}(F_2(\xi_0)) - \lambda_n^2 J^2(\eta_0)$$

hence,

$$\begin{aligned} & \bar{Q}_{1n}(\bar{F}_1(\hat{\eta}_n)) - \bar{Q}_{1n}(F_1(\eta_0)) + \bar{Q}_{2n}(\bar{F}_2(\hat{\xi}_n)) - \bar{Q}_{2n}(F_2(\xi_0)) \\ \geq \; & \frac{1}{2}\left\{\bar{Q}_{1n}(F_1(\hat{\eta}_n)) - \bar{Q}_{1n}(F_1(\eta_0)) + \bar{Q}_{2n}(F_2(\hat{\xi}_n)) - \bar{Q}_{2n}(F_2(\xi_0))\right\} \\ \geq \; & \frac{1}{2}\lambda_n^2\left(J^2(\hat{\eta}_n) - J^2(\eta_0)\right) \end{aligned}$$

Let $\mathcal{A}$ be a subset of a metric space $(\mathcal{L}, \rho)$ of real-valued functions. Define the ε-entropy as $H(\varepsilon, \mathcal{A}, \rho) = \log N(\varepsilon, \mathcal{A}, \rho)$, where the ε-covering number $N(\varepsilon, \mathcal{A}, \rho)$ of $\mathcal{A}$ is the smallest value of N for which there exist functions $a_1, \cdots, a_N$ in $\mathcal{L}$, such that for each $a \in \mathcal{A}$, $\rho(a, a_j) \leq \varepsilon$ for some $j \in \{1, \cdots, N\}$.

Because the class of smooth functions η satisfies the entropy growth condition, i.e.

$$\sup_{\varepsilon>0} \varepsilon^{1/k} H\left(\varepsilon, \{\eta : |\eta|_\infty \leq C, J(\eta) \leq C\}, |\cdot|_\infty\right) < \infty,$$

it can be shown that

$$\sup_{\varepsilon>0} \varepsilon^{1/k} H\left(\varepsilon, \{\xi = \alpha + \delta\eta : |\alpha|, |\delta|, |\eta|_\infty \leq C, J(\eta) \leq C\}, |\cdot|_\infty\right) < \infty.$$

Use similar argument in Mammen and van de Geer (1997), it follows from (R9) that,

$$\sup_{\varepsilon>0} \varepsilon^{1/k} H\left(\varepsilon, \left\{\frac{\gamma_\eta - \gamma_{\eta 0}}{1 + J(\eta)} : J(\eta) < \infty, \frac{|\gamma_\eta|_\infty}{1 + J(\eta)} \leq C\right\}, |\cdot|_\infty\right) < \infty$$

Then from the sub-Gaussianality of W, applying Theorem 2.2 in Mammen and van de Geer (1997), and using the fact that $\|\hat{\gamma}_n - \gamma_0\|_n \geq (1/C_2)\|\hat{\eta}_n - \eta_0\|_n$, we have

$$\frac{(1/m)\sum_{Z_i=1} W_i(\hat{\gamma}_n(T_i) - \gamma_0(T_i))}{\|\hat{\eta}_n - \eta_0\|_m^{1-1/(2k)}(1 + J(\hat{\eta}_n))^{1/(2k)} \vee (1 + J(\hat{\eta}_n))m^{-(2k-1)/2(2k+1)}} = O_{\mathbf{P}}(m^{-1/2}).$$

Similarly, if (R10) and (R11) hold true, the entropy condition holds for the class

$$\left\{\frac{\sqrt{\bar{F}_2(\xi)} - \sqrt{F_2(\xi_0)}}{\sqrt{F_2(\xi_0)}(1 + J(\eta))} : \boldsymbol{\theta} \in \Theta\right\}.$$

Thus by the sub-Gaussianality of R,

$$\begin{aligned} & \frac{(1/n)\sum_{i=1}^n R_i\left(\sqrt{\bar{F}_2(\hat{\xi}_n(T_i))} - \sqrt{F_2(\xi_0(T_i))}\right)\Big/\sqrt{F_2(\xi_0(T_i))}}{\|\sqrt{\bar{F}_2(\hat{\xi}_n)} - \sqrt{F_2(\xi_0)}\|_n^{1-1/(2k)}(1 + J(\hat{\eta}_n))^{1/(2k)} \vee (1 + J(\hat{\eta}_n))n^{-(2k-1)/2(2k+1)}} \\ = \; & O_{\mathbf{P}}(n^{-1/2}), \end{aligned}$$

and

$$\begin{aligned}
&\frac{(1/n)\sum_{i=1}^n R_i \left(\sqrt{1-\bar{F}_2(\hat{\xi}_n(T_i))} - \sqrt{1-F_2(\xi_0(T_i))}\right) \Big/ \sqrt{1-F_2(\xi_0(T_i))}}{\|\sqrt{1-\bar{F}_2(\hat{\xi}_n)} - \sqrt{1-F_2(\xi_0)}\|_n^{1-1/(2k)}(1+J(\hat{\eta}_n))^{1/(2k)} \vee (1+J(\hat{\eta}_n))n^{-(2k-1)/2(2k+1)}} \\
= \;& O_{\mathbf{P}}(n^{-1/2}).
\end{aligned}$$

Using similar procedure as in the proof of Lemma 3.1 in Mammen and van de Geer (1997), we find that $J(\hat{\eta}_n) = O_{\mathbf{P}}(1)$, and

$$\begin{aligned}
\|\hat{\eta}_n - \eta_0\|_n &= O_{\mathbf{P}}(\lambda_n), & (23)\\
\left\|\sqrt{\bar{F}_2(\hat{\xi}_n)} - \sqrt{F_2(\xi_0)}\right\|_n &= O_{\mathbf{P}}(\lambda_n), & (24)\\
\left\|\sqrt{1-\bar{F}_2(\hat{\xi}_n)} - \sqrt{1-F_2(\xi_0)}\right\|_n &= O_{\mathbf{P}}(\lambda_n). & (25)
\end{aligned}$$

Equations (24) and (25) imply

$$\left\|F_2(\hat{\xi}_n) - F_2(\xi_0)\right\|_n = O_{\mathbf{P}}(\lambda_n).$$

Then (R12) and (R13) entail that

$$\|\hat{\xi}_n - \xi_0\|_n = O_{\mathbf{P}}(\lambda_n). \qquad (26)$$

Finally, the stated convergence rates of $\hat{\alpha}_n$ and $\hat{\delta}_n$ follow from the following lemma.

Lemma 5. *If (23), (26) and (R2) hold, then*

$$\begin{aligned}
|\hat{\alpha}_n - \alpha_0| &= O_{\mathbf{P}}(\lambda_n), \\
|\hat{\delta}_n - \delta_0| &= O_{\mathbf{P}}(\lambda_n).
\end{aligned}$$

Proof of Lemma 5.

Suppose (R2) is true so that for some $t_1 \neq t_2$, $\eta_0(t_1) \neq \eta_0(t_2)$. It follows from (23) and (26) that

$$\begin{aligned}
&|\hat{\eta}_n(t_1) - \eta_0(t_1)| = O_{\mathbf{P}}(\lambda_n) \quad \text{and} \quad |\hat{\xi}_n(t_1) - \xi_0(t_1)| = O_{\mathbf{P}}(\lambda_n), \\
&|\hat{\eta}_n(t_2) - \eta_0(t_2)| = O_{\mathbf{P}}(\lambda_n) \quad \text{and} \quad |\hat{\xi}_n(t_2) - \xi_0(t_2)| = O_{\mathbf{P}}(\lambda_n).
\end{aligned}$$

Then $\hat{\alpha}_n, \hat{\delta}_n$ satisfy

$$\begin{cases} \hat{\xi}_n(t_1) = \hat{\alpha}_n + \hat{\delta}_n\hat{\eta}_n(t_1) \\ \hat{\xi}_n(t_2) = \hat{\alpha}_n + \hat{\delta}_n\hat{\eta}_n(t_2), \end{cases}$$

and similarly,

$$\begin{cases} \xi_0(t_1) = \alpha_0 + \delta_0\eta_0(t_1) \\ \xi_0(t_2) = \alpha_0 + \delta_0\eta_0(t_2), \end{cases}$$

Because $\eta_0(t_1) \neq \eta_0(t_2)$, when n is large enough $\hat{\eta}_n(t_1) \neq \hat{\eta}_n(t_2)$, with probability → 1. Hence we can invert both groups of equations to obtain the solutions for $\hat{\alpha}_n, \hat{\delta}_n, \alpha_0$ and δ_0. Then it follows from the convergence rate of $\hat{\eta}_n$ and $\hat{\xi}_n$ that $\hat{\alpha}_n$ and $\hat{\delta}_n$ share the same convergence rate, i.e.

$$|\hat{\alpha}_n - \alpha_0| = O_{\mathbf{P}}\left(\lambda_n\right), \quad |\hat{\delta}_n - \delta_0| = O_{\mathbf{P}}\left(\lambda_n\right).$$

Proof of the Asymptotic Normality

Proof of Theorem 2.

We now elaborate the derivation of (18) as follows:

$$\frac{d}{ds}Q(\widehat{\boldsymbol{\theta}}^I_{ns})\Big|_{s=0} = \frac{d}{ds}\bar{Q}_{1n}(F_1(\hat{\eta}^I_{ns}))\Big|_{s=0} + \frac{d}{ds}\bar{Q}_{2n}(F_2(\hat{\xi}^I_{ns}))\Big|_{s=0} := A_I + B_I,$$

where $\hat{\xi}^I_{ns} = \hat{\alpha}_n + s + \hat{\delta}_n(\hat{\eta}_n + sh_1)$.

$$\begin{aligned} A_I &= \frac{1}{n}\sum Z_i\{Y_i - F_1(\hat{\eta}_n(T_i))\}l_1(\hat{\eta}_n(T_i))h_2(T_i) \\ &= \frac{1}{n}\sum Z_iW_il_1(\hat{\eta}_n(T_i))h_1(T_i) - \frac{1}{n}\sum Z_i\{F_1(\hat{\eta}_n(T_i)) - F_1(\eta_0(T_i))\}l_1(\hat{\eta}_n(T_i))h_1(T_i) \\ &:= C_I - D_I \end{aligned}$$

We shall make repeated use of the following technique. The class

$$\mathcal{A} = \{a(Y,Z,T) = Z(Y - \mu_0(T))[l_1(\eta(T)) - l_1(\eta_0(T))]h_1(T) : |\eta - \eta_0|_\infty \leq d_1, J(\eta) \leq C\}$$

satisfies condition (2.7) of Theorem 2.4 in Mammen and van de Geer (1997), which implies that $\frac{1}{\sqrt{n}}\sum_{i=1}^n\{a(Y_i,Z_i,T_i) - E(a)\} = o_{\mathbf{P}}(1)$ uniformly for $a \in \mathcal{A}$ with $||a|| \to 0$. Note that for fixed $a \in \mathcal{A}$, $E(a) = 0$. This result implies that

$$C_I = \frac{1}{n}\sum Z_iW_il_{10}(T_i)h_1(T_i) + o_{\mathbf{P}}(n^{-1/2}),$$

To see this, note that $Z(Y - \mu_0(T))$ is a centered random variable with finite variance, and $h_1(T)$ is a fixed bounded function. Also by (R14), $||\hat{\eta}_n - \eta_0||_n = o_{\mathbf{P}}(1)$ implies $||\hat{l}_{1n} - l_{10}||_n = o_{\mathbf{P}}(1)$, where we write $\hat{l}_{1n} = l_1 \circ \hat{\eta}_n$ (and $\hat{l}_{2n} = l_2 \circ \hat{\xi}_n$ for later use).

Next, we write D_I as

$$\begin{aligned} D_I &= \frac{1}{n}\sum Z_i\{F_1(\hat{\eta}_n(T_i)) - F_1(\eta_0(T_i))\}\hat{l}_{1n}(T_i)h_1(T_i) \\ &= \frac{1}{n}\sum Z_i(\hat{\eta}_n(T_i) - \eta_0(T_i))f_{10}(T_i)l_{10}(T_i)h_1(T_i) \\ &\quad + \frac{1}{n}\sum Z_i\{F_1(\hat{\eta}_n(T_i)) - F_1(\eta_0(T_i)) - (\hat{\eta}_n(T_i) - \eta_0(T_i))f_{10}(T_i)\}l_{10}(T_i)h_1(T_i) \\ &\quad + \frac{1}{n}\sum Z_i\{F_1(\hat{\eta}_n(T_i)) - F_1(\eta_0(T_i))\}(\hat{l}_{1n}(T_i) - l_{10}(T_i))h_1(T_i) \\ &:= D_{Ia} + D_{Ib} + D_{Ic}. \end{aligned}$$

By the mean value theorem and (R14), there exist a constant C that may differ in each occurrence that

$$\begin{aligned}|D_{Ib}| &= \frac{1}{n}\sum Z_i|\hat{\eta}_n(T_i)-\eta_0(T_i)||f_1(\tilde{\eta}(T_i))-f_{10}(T_i)||l_{10}(T_i)h_1(T_i)| \\ &\leq C\frac{1}{n}\sum Z_i(\hat{\eta}_n(T_i)-\eta_0(T_i))^2|l_{10}(T_i)h_1(T_i)| \\ &\leq C|h_1|_\infty\cdot||\hat{\eta}_n-\eta_0||_n^2 = o_{\mathbf{P}}(n^{-1/2}),\end{aligned}$$

where $\tilde{\eta}(T_i)$ lies between $\eta_0(T_i)$ and $\hat{\eta}_n(T_i)$; note that $|h_1|_\infty < \infty$ follows from (12) and (R14). Similarly, it can be shown that

$$|D_{Ic}| = o_{\mathbf{P}}(n^{-1/2}).$$

Thus, we have

$$D_I = \frac{1}{n}\sum Z_i(\hat{\eta}_n(T_i)-\eta_0(T_i))f_{10}(T_i)l_{10}(T_i)h_1(T_i) + o_{\mathbf{P}}(n^{-1/2}).$$

Hence,

$$\begin{aligned}A_I &= \frac{1}{n}\sum Z_iW_il_{10}(T_i)h_1(T_i) \\ &\quad -\frac{1}{n}\sum Z_i(\hat{\eta}_n(T_i)-\eta_0(T_i))f_{10}(T_i)l_{10}(T_i)h_1(T_i) + o_{\mathbf{P}}(n^{-1/2}).\end{aligned}$$

Next, let us write B_I as

$$\begin{aligned}B_I &= \frac{1}{n}\sum\{Z_i - F_2(\hat{\xi}_n(T_i))\}l_2(\hat{\xi}_n(T_i))(1+\hat{\delta}_nh_1(T_i)) \\ &= \frac{1}{n}\sum\{Z_i - F_2(\xi_0(T_i))\}\hat{l}_{2n}(T_i)(1+\hat{\delta}_nh_1(T_i)) \\ &\quad -\frac{1}{n}\sum\{F_2(\hat{\xi}_n(T_i)) - F_2(\xi_0(T_i))\}\hat{l}_{2n}(T_i)(1+\hat{\delta}_nh_1(T_i)) \\ &:= E_I - F_I\end{aligned}$$

By (R14), and applying the argument in the case of C_I, $||\hat{\xi}_n - \xi_0||_n = o_{\mathbf{P}}(1)$ implies that $||\hat{l}_{2n} - l_{20}||_n = o_{\mathbf{P}}(1)$. Together with $|\hat{\delta}_n - \delta_0| = o_{\mathbf{P}}(1)$, we have

$$E_I = \frac{1}{n}\sum R_il_{20}(T_i)(1+\delta_0h_1(T_i)) + o_{\mathbf{P}}(n^{-1/2}).$$

F_I could be written as

$$\begin{aligned}F_I &= \frac{1}{n}\sum\{F_2(\hat{\xi}_n(T_i)) - F_2(\xi_0(T_i))\}\hat{l}_{2n}(T_i)(1+\delta_0h_1(T_i)) \\ &\quad +\frac{1}{n}\sum\{F_2(\hat{\xi}_n(T_i)) - F_2(\xi_0(T_i))\}\hat{l}_{2n}(T_i)(\hat{\delta}_n-\delta_0)h_1(T_i).\end{aligned}$$

It follows from (10) and assumption (R14) that the second term in the above equation is of the order $o_{\mathbf{P}}(n^{-1/2})$. That is,

$$F_I = \frac{1}{n}\sum\{F_2(\hat{\xi}_n(T_i)) - F_2(\xi_0(T_i))\}\hat{l}_{2n}(T_i)(1+\delta_0h_1(T_i)) + o_{\mathbf{P}}(n^{-1/2}).$$

It can be further written as

$$\begin{aligned} F_I &= \frac{1}{n}\sum(\hat{\xi}_n(T_i)-\xi_0(T_i))f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i)) \\ &\quad +\frac{1}{n}\sum\{F_2(\hat{\xi}_n(T_i))-F_2(\xi_0(T_i))-(\hat{\xi}_n(T_i)-\xi_0(T_i))f_{20}(T_i)\}l_{20}(T_i)(1+\delta_0h_1(T_i)) \\ &\quad +\frac{1}{n}\sum\{F_2(\hat{\xi}_n(T_i))-F_2(\xi_0(T_i))\}(\hat{l}_{2n}(T_i)-l_{20}(T_i))(1+\delta_0h_1(T_i))+o_{\mathbf{P}}(n^{-1/2}) \\ &:= F_{Ia}+F_{Ib}+F_{Ic}+o_{\mathbf{P}}(n^{-1/2}). \end{aligned}$$

By the same argument as before, we can show that

$$\begin{aligned} |F_{Ib}| &= o_{\mathbf{P}}(n^{-1/2}), \\ |F_{Ic}| &= o_{\mathbf{P}}(n^{-1/2}). \end{aligned}$$

Furthermore,

$$\begin{aligned} \hat{\xi}_n(T_i)-\xi_0(T_i) &= \hat{\alpha}_n-\alpha_0+(\hat{\delta}_n\hat{\eta}_n(T_i)-\delta_0\eta_0(T_i)) \\ &= (\hat{\alpha}_n-\alpha_0)+(\hat{\delta}_n-\delta_0)\eta_0(T_i)+\delta_0(\hat{\eta}_n(T_i)-\eta_0(T_i)) \\ &\quad +(\hat{\delta}_n-\delta_0)(\hat{\eta}_n(T_i)-\eta_0(T_i)). \end{aligned}$$

Again, it follows from (10) and (11) that

$$\frac{1}{n}\sum(\hat{\delta}_n-\delta_0)(\hat{\eta}_n(T_i)-\eta_0(T_i))f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i))=o_{\mathbf{P}}(n^{-1/2}).$$

Therefore,

$$\begin{aligned} F_I &= (\hat{\alpha}_n-\alpha_0)\frac{1}{n}\sum f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i)) \\ &\quad +(\hat{\delta}_n-\delta_0)\frac{1}{n}\sum\eta_0(T_i)f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i)) \\ &\quad +\frac{1}{n}\sum(\hat{\eta}_n(T_i)-\eta_0(T_i))\delta_0f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i))+o_{\mathbf{P}}(n^{-1/2}). \end{aligned}$$

Note that

$$\begin{aligned} B_I &= \frac{1}{n}\sum R_il_{20}(T_i)(1+\delta_0h_1(T_i)) \\ &\quad -(\hat{\alpha}_n-\alpha_0)\frac{1}{n}\sum f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i)) \\ &\quad -(\hat{\delta}_n-\delta_0)\frac{1}{n}\sum\eta_0(T_i)f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i)) \\ &\quad -\frac{1}{n}\sum(\hat{\eta}_n(T_i)-\eta_0(T_i))\delta_0f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i))+o_{\mathbf{P}}(n^{-1/2}). \end{aligned}$$

Hence,

$$\begin{aligned}\frac{d}{ds}Q(\widehat{\boldsymbol{\theta}}_{ns}^{I})\Big|_{s=0} &= \frac{1}{n}\sum\left\{Z_iW_il_{10}(T_i)h_1(T_i)+R_il_{20}(T_i)(1+\delta_0h_1(T_i))\right\}\\ &\quad -(\hat{\alpha}_n-\alpha_0)\frac{1}{n}\sum f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i))\\ &\quad -(\hat{\delta}_n-\delta_0)\frac{1}{n}\sum \eta_0(T_i)f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i))\\ &\quad -\frac{1}{n}\sum\left(\hat{\eta}_n(T_i)-\eta_0(T_i)\right)\left\{Z_if_{10}(T_i)l_{10}(T_i)h_1(T_i)+\delta_0f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i))\right\}\\ &\quad +o_{\mathbf{P}}(n^{-1/2}).\end{aligned}$$

Recall

$$h_1(t)=-\frac{\delta_0f_{20}(t)l_{20}(t)}{p_0(t)f_{10}(t)l_{10}(t)+\delta_0^2f_{20}(t)l_{20}(t)}, \qquad (27)$$

where $p_0(t)=F_2(\xi_0(t))=E_0(Z|T=t)$, so an application of Theorem 2.4 in Mammen and van de Geer (1997) yields

$$\begin{aligned}\frac{d}{ds}Q(\widehat{\boldsymbol{\theta}}_{ns}^{I})\Big|_{s=0} &= \frac{1}{n}\sum\left\{Z_iW_il_{10}(T_i)h_1(T_i)+R_il_{20}(T_i)(1+\delta_0h_1(T_i))\right\}\\ &\quad -(\hat{\alpha}_n-\alpha_0)\frac{1}{n}\sum f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i))\\ &\quad -(\hat{\delta}_n-\delta_0)\frac{1}{n}\sum \eta_0(T_i)f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i))\\ &\quad +o_{\mathbf{P}}(n^{-1/2}).\end{aligned}$$

On the other hand, under the assumption that $J(\hat{\eta}_n)=O_{\mathbf{P}}(1)$, $J(h_1)<\infty$ and $\lambda_n=o_{\mathbf{P}}(n^{-1/4})$,

$$\frac{d}{ds}\lambda_n^2J^2(\hat{\eta}_{ns}^{I})\Big|_{s=0}\le 2\lambda_n^2J(\hat{\eta}_n)J(h_1)=o_{\mathbf{P}}(n^{-1/2}).$$

We obtain

$$\begin{aligned}0 &= \frac{1}{n}\sum\left\{Z_iW_il_{10}(T_i)h_1(T_i)+R_il_{20}(T_i)(1+\delta_0h_1(T_i))\right\}\\ &\quad -(\hat{\alpha}_n-\alpha_0)\frac{1}{n}\sum f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i))\\ &\quad -(\hat{\delta}_n-\delta_0)\frac{1}{n}\sum \eta_0(T_i)f_{20}(T_i)l_{20}(T_i)(1+\delta_0h_1(T_i))\\ &\quad +o_{\mathbf{P}}(n^{-1/2}),\end{aligned}$$

which is Equation (19).

Now let us work on the second path of perturbation,

$$\frac{d}{ds}Q(\widehat{\boldsymbol{\theta}}_{ns}^{II})\Big|_{s=0}=\frac{d}{ds}\bar{Q}_{1n}(F_1(\hat{\eta}_{ns}^{II}))\Big|_{s=0}+\frac{d}{ds}\bar{Q}_{2n}(F_2(\hat{\xi}_{ns}^{II}))\Big|_{s=0}:=A_{II}+B_{II},$$

where $\hat{\xi}_{ns}^{II} = \hat{\alpha}_n + (\hat{\delta}_n + s)(\hat{\eta}_n + sh_2)$. By similar procedure as A_I, we can write

$$\begin{aligned} A_{II} &= \frac{1}{n}\sum Z_i\{Y_i - F_1(\hat{\eta}_n(T_i))\}\hat{l}_{1n}(T_i)h_2(T_i) \\ &= \frac{1}{n}\sum Z_iW_il_{10}(T_i)h_2(T_i) \\ &\quad -\frac{1}{n}\sum Z_i(\hat{\eta}_n(T_i) - \eta_0(T_i))f_{10}(T_i)l_{10}(T_i)h_2(T_i) + o_{\mathbf{P}}(n^{-1/2}). \end{aligned}$$

Also,

$$\begin{aligned} B_{II} &= \frac{1}{n}\sum \{Z_i - F_2(\hat{\xi}_n(T_i))\}l_2(\hat{\xi}_n(T_i))(\hat{\eta}_n(T_i) + \hat{\delta}_nh_2(T_i)) \\ &= \frac{1}{n}\sum R_il_{20}(T_i)(\eta_0(T_i) + \delta_0h_2(T_i)) \\ &\quad -(\hat{\alpha}_n - \alpha_0)\frac{1}{n}\sum f_{20}(T_i)l_{20}(T_i)(\eta_0(T_i) + \delta_0h_2(T_i)) \\ &\quad -(\hat{\delta}_n - \delta_0)\frac{1}{n}\sum \eta_0(T_i)f_{20}(T_i)l_{20}(T_i)(\eta_0(T_i) + \delta_0h_2(T_i)) \\ &\quad -\frac{1}{n}\sum (\hat{\eta}_n(T_i) - \eta_0(T_i))\delta_0f_{20}(T_i)l_{20}(T_i)(\eta_0(T_i) + \delta_0h_2(T_i)) + o_{\mathbf{P}}(n^{-1/2}). \end{aligned}$$

Recall

$$h_2(t) = -\frac{\delta_0\eta_0(t)f_{20}(t)l_{20}(t)}{p_0(t)f_{10}(t)l_{10}(t) + \delta_0^2f_{20}(t)l_{20}(t)}, \tag{28}$$

and

$$\frac{d}{ds}\lambda_n^2J^2(\hat{\eta}_{ns}^{II})\Big|_{s=0} \leq 2\lambda_n^2J(\hat{\eta}_n)J(h_2) = o_{\mathbf{P}}(n^{-1/2}),$$

Therefore, we have

$$\begin{aligned} 0 &= \frac{1}{n}\sum \{Z_iW_il_{10}(T_i)h_2(T_i) + R_il_{20}(T_i)(\eta_0(T_i) + \delta_0h_2(T_i))\} \\ &\quad -(\hat{\alpha}_n - \alpha_0)\frac{1}{n}\sum f_{20}(T_i)l_{20}(T_i)(\eta_0(T_i) + \delta_0h_2(T_i)) \\ &\quad -(\hat{\delta}_n - \delta_0)\frac{1}{n}\sum \eta_0(T_i)f_{20}(T_i)l_{20}(T_i)(\eta_0(T_i) + \delta_0h_2(T_i)) \\ &\quad +o_{\mathbf{P}}(n^{-1/2}), \end{aligned}$$

which is Equation (20). Write the two score equations (19) and (20) in matrix form as

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} \sqrt{n}(\hat{\alpha}_n - \alpha_0) \\ \sqrt{n}(\hat{\delta}_n - \delta_0) \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}. \tag{29}$$

By (27) and (28), and using the law of large numbers, we obtain

$$\begin{aligned}
a_{11} &= \frac{1}{n}\sum \frac{p_0(T_i)f_{10}(T_i)l_{10}(T_i)f_{20}(T_i)l_{20}(T_i)}{p_0(T_i)f_{10}(T_i)l_{10}(T_i)+\delta_0^2 f_{20}(T_i)l_{20}(T_i)} = \left\| \left(\frac{p_0 f_{10} l_{10} f_{20} l_{20}}{p_0 f_{10} l_{10} + \delta_0^2 f_{20} l_{20}} \right)^{1/2} \right\|^2 + o_{\mathbf{P}}(1), \\
a_{12} &= \frac{1}{n}\sum \frac{p_0(T_i)\eta_0(T_i)f_{10}(T_i)l_{10}(T_i)f_{20}(T_i)l_{20}(T_i)}{p_0(T_i)f_{10}(T_i)l_{10}(T_i)+\delta_0^2 f_{20}(T_i)l_{20}(T_i)} = \left\| \left(\frac{p_0 \eta_0 f_{10} l_{10} f_{20} l_{20}}{p_0 f_{10} l_{10} + \delta_0^2 f_{20} l_{20}} \right)^{1/2} \right\|^2 + o_{\mathbf{P}}(1), \\
a_{21} &= \frac{1}{n}\sum \frac{p_0(T_i)\eta_0(T_i)f_{10}(T_i)l_{10}(T_i)f_{20}(T_i)l_{20}(T_i)}{p_0(T_i)f_{10}(T_i)l_{10}(T_i)+\delta_0^2 f_{20}(T_i)l_{20}(T_i)} = \left\| \left(\frac{p_0 \eta_0 f_{10} l_{10} f_{20} l_{20}}{p_0 f_{10} l_{10} + \delta_0^2 f_{20} l_{20}} \right)^{1/2} \right\|^2 + o_{\mathbf{P}}(1), \\
a_{22} &= \frac{1}{n}\sum \frac{p_0(T_i)\eta_0^2(T_i)f_{10}(T_i)l_{10}(T_i)f_{20}(T_i)l_{20}(T_i)}{p_0(T_i)f_{10}(T_i)l_{10}(T_i)+\delta_0^2 f_{20}(T_i)l_{20}(T_i)} = \left\| \left(\frac{p_0 \eta_0^2 f_{10} l_{10} f_{20} l_{20}}{p_0 f_{10} l_{10} + \delta_0^2 f_{20} l_{20}} \right)^{1/2} \right\|^2 + o_{\mathbf{P}}(1).
\end{aligned}$$

Moreover,

$$\begin{aligned}
b_1 &= \frac{1}{\sqrt{n}}\sum \frac{l_{10}(T_i)l_{20}(T_i)\big\{R_i p_0(T_i)f_{10}(T_i) - \delta_0 W_i Z_i f_{20}(T_i)\big\}}{p_0(T_i)f_{10}(T_i)l_{10}(T_i)+\delta_0^2 f_{20}(T_i)l_{20}(T_i)} + o_{\mathbf{P}}(1), \\
b_2 &= \frac{1}{\sqrt{n}}\sum \frac{\eta_0(T_i)l_{10}(T_i)l_{20}(T_i)\big\{R_i p_0(T_i)f_{10}(T_i) - \delta_0 W_i Z_i f_{20}(T_i)\big\}}{p_0(T_i)f_{10}(T_i)l_{10}(T_i)+\delta_0^2 f_{20}(T_i)l_{20}(T_i)} + o_{\mathbf{P}}(1).
\end{aligned}$$

Letting

$$m_0(t) = \frac{l_{10}(t)l_{20}(t)}{p_0(t)f_{10}(t)l_{10}(t)+\delta_0^2 f_{20}(t)l_{20}(t)},$$

the above formulas can be further simplified as follows,

$$\begin{aligned}
a_{11} &= \left\| (m_0 p_0 f_{10} f_{20})^{1/2} \right\|^2 + o_{\mathbf{P}}(1), \\
a_{12} &= a_{21} = \left\| (m_0 p_0 \eta_0 f_{10} f_{20})^{1/2} \right\|^2 + o_{\mathbf{P}}(1), \\
a_{22} &= \left\| (m_0 p_0 f_{10} f_{20})^{1/2} \eta_0 \right\|^2 + o_{\mathbf{P}}(1),
\end{aligned}$$

and

$$\begin{aligned}
b_1 &= \frac{1}{\sqrt{n}}\sum m_0(T_i)\left\{R_i p_0(T_i)f_{10}(T_i) - \delta_0 W_i Z_i f_{20}(T_i)\right\} + o_{\mathbf{P}}(1), \\
b_2 &= \frac{1}{\sqrt{n}}\sum m_0(T_i)\eta_0(T_i)\left\{R_i p_0(T_i)f_{10}(T_i) - \delta_0 W_i Z_i f_{20}(T_i)\right\} + o_{\mathbf{P}}(1).
\end{aligned}$$

By checking the coefficient matrix on the left hand side of (29), it follows from the Cauchy-Schwartz inequality that it is nonnegative definite, and it is singular if and only if $\eta_0(T_i)$ is a constant for all $i = 1, \cdots, n$, which holds with probability $\to 0$ as $n \to \infty$, given assumptions (R1) and (R2). Hence, $\sqrt{n}(\hat{\alpha}_n - \alpha_0, \hat{\delta}_n - \delta_0)^T$ is asymptotically bivariate normal. The explicit form of the covariance matrix can be obtained by solving (29) and plug in the specific expressions for the coefficients.

Proof of Theorem 3.
The proof of Theorem 3 follows from the proof of Lemma 3.1 in Mammen and van de Geer (1997) with the parameter space being bounded, so it is omitted here.

Proof of Theorem 4.

Almost the same as the the proof of Theorem 2, but with the following directions of perturbation:

$$\begin{aligned} h_1^* &= -\frac{w_0 f_{20}\zeta_0}{p_0^* f_{10}^* l_{10}^* v_0 + w_0^2}, \\ h_2^* &= -\frac{w_0 \eta_0 f_{20}\zeta_0}{p_0^* f_{10}^* l_{10}^* v_0 + w_0^2}. \end{aligned}$$

The score equations then become

$$\begin{aligned} 0 &= \frac{1}{n}\sum\left\{E_i W_i^* l_{10}^*(T_i)h_1^*(T_i) + R_i^* \frac{f_{20}(T_i)\zeta_0(T_i)(1+\delta_0 h_1^*(T_i)) + F_{20}(T_i)\dot{\zeta}_0(T_i)h_1^*(T_i)}{v_0(T_i)}\right\} \\ &\quad -(\hat{\alpha}_n - \alpha_0)\frac{1}{n}\sum f_{20}(T_i)\zeta_0(T_i)\frac{f_{20}(T_i)\zeta_0(T_i)(1+\delta_0 h_1^*(T_i)) + F_{20}(T_i)\dot{\zeta}_0(T_i)h_1^*(T_i)}{v_0(T_i)} \\ &\quad -(\hat{\delta}_n - \delta_0)\frac{1}{n}\sum \eta_0(T_i) f_{20}(T_i)\zeta_0(T_i)\frac{f_{20}(T_i)\zeta_0(T_i)(1+\delta_0 h_1^*(T_i)) + F_{20}(T_i)\dot{\zeta}_0(T_i)h_1^*(T_i)}{v_0(T_i)} \\ &\quad +o_{\mathbf{P}}(n^{-1/2}), \end{aligned}$$

and

$$\begin{aligned} 0 &= \frac{1}{n}\sum\left\{E_i W_i^* l_{10}^*(T_i)h_2^*(T_i) + R_i^* \frac{f_{20}(T_i)\zeta_0(T_i)(\eta_0(T_i)+\delta_0 h_2^*(T_i)) + F_{20}(T_i)\dot{\zeta}_0(T_i)h_2^*(T_i)}{v_0(T_i)}\right\} \\ &\quad -(\hat{\alpha}_n - \alpha_0)\frac{1}{n}\sum f_{20}(T_i)\zeta_0(T_i)\frac{f_{20}(T_i)\zeta_0(T_i)(\eta_0(T_i)+\delta_0 h_2^*(T_i)) + F_{20}(T_i)\dot{\zeta}_0(T_i)h_2^*(T_i)}{v_0(T_i)} \\ &\quad -(\hat{\delta}_n - \delta_0)\frac{1}{n}\sum \eta_0(T_i) f_{20}(T_i)\zeta_0(T_i)\frac{f_{20}(T_i)\zeta_0(T_i)(\eta_0(T_i)+\delta_0 h_2^*(T_i)) + F_{20}(T_i)\dot{\zeta}_0(T_i)h_2^*(T_i)}{v_0(T_i)} \\ &\quad +o_{\mathbf{P}}(n^{-1/2}), \end{aligned}$$

leading to the required conclusion.

A Simple Justification of Computing the Asymptotic Variance by Inverting the Fisher Information

In Theorem 2, we proved the asymptotic normality of the estimators $\hat{\alpha}_n$ and $\hat{\delta}_n$, with a rather complex formula for the asymptotic covariance matrix. However, how does it compare with the results obtained from inverting the observed Fisher information? Here we will illustrate their asymptotic equivalence in a simplest case where the regular response follows an additive model and we use the logit link for the probability of the inflated zero atom.

For the additive model, suppose the dispersion parameter is known and, without loss of generality, equal to 1. Then we have: $F_1(\eta) = \eta, V = 1, f_1 = 1$ and $l_1 = 1$. If F_2 is the inverse logit function, then $f_2 = p(1-p)$ and $l_2 = 1$. Now the formulas could be simplified as follows:

$$\begin{aligned} m_0 &= \frac{1}{p_0(1+\delta_0^2(1-p_0))}, \\ a_{11} &\doteq E\left\{\frac{p_0(1-p_0)}{1+\delta_0^2(1-p_0)}\right\}, \\ a_{12} &\doteq E\left\{\frac{\eta_0 p_0(1-p_0)}{1+\delta_0^2(1-p_0)}\right\}, \\ a_{22} &\doteq E\left\{\frac{\eta_0^2 p_0(1-p_0)}{1+\delta_0^2(1-p_0)}\right\}. \end{aligned}$$

Also,

$$\begin{aligned} b_1 &\doteq \frac{1}{\sqrt{n}}\sum\left\{\frac{1}{1+\delta_0^2(1-p_0(T_i))}R_i - \frac{\delta_0(1-p_0(T_i))}{1+\delta_0^2(1-p_0(T_i))}W_iZ_i\right\}, \\ b_2 &\doteq \frac{1}{\sqrt{n}}\sum\left\{\frac{\eta_0(T_i)}{1+\delta_0^2(1-p_0(T_i))}R_i - \frac{\delta_0\eta_0(T_i)(1-p_0(T_i))}{1+\delta_0^2(1-p_0(T_i))}W_iZ_i\right\}. \end{aligned}$$

Solving Equation (29) we have

$$\begin{aligned} \sqrt{n}(\hat{\alpha}_n - \alpha_0) &= \frac{a_{22}b_1 - a_{12}b_2}{a_{11}a_{22} - a_{12}^2}, \\ \sqrt{n}(\hat{\delta}_n - \delta_0) &= \frac{a_{11}b_2 - a_{12}b_1}{a_{11}a_{22} - a_{12}^2}. \end{aligned}$$

Using the fact that $Var(R|T) = p_0(T)(1-p_0(T))$, $Var(ZW|T) = p_0(T)(1-p_0(T))$ and noting that R and W are conditionally uncorrelated given T, then by the central limit theorem the limiting distribution of $\sqrt{n}(\hat{\alpha}_n - \alpha_0)$ is $Normal(0, v_\alpha)$, where v_α is

$$E\left\{\frac{(a_{22} - a_{12}\eta_0)^2 p_0(1-p_0)}{1+\delta_0^2(1-p_0)}\right\}$$

multiplied by the constant $\left(a_{11}a_{22} - a_{12}^2\right)^{-2}$. Notice we have suppressed T from the notations, even though the expectation is taken with respect to T. Similarly, $\sqrt{n}(\hat{\delta}_n - \delta_0)$ is asymptotically $Normal(0, v_\delta)$, with v_δ equal to

$$E\left\{\frac{(a_{12} - a_{11}\eta_0)^2 p_0(1-p_0)}{1+\delta_0^2(1-p_0)}\right\}$$

multiplied by $\left(a_{11}a_{22} - a_{12}^2\right)^{-2}$. Considering the specific forms of a_{11}, a_{12} and a_{22}, we have the approximations

$$\begin{aligned} v_\alpha &\propto E\left(\eta_0^2 p_0(1-p_0)\right), \\ v_\delta &\propto E\big(p_0(1-p_0)\big). \end{aligned}$$

In comparison, the information matrix of α and δ equals

$$I_{\alpha\delta} = \begin{pmatrix} \sum p_i(1-p_i) & \sum \eta_i p_i(1-p_i) \\ \sum \eta_i p_i(1-p_i) & \sum \eta_i^2 p_i(1-p_i) \end{pmatrix},$$

hence the claimed asymptotic equivalence.