Estimation of Zero-inflated Count Time Series Models With and Without Covariates

by

Bartholomew Embir Ghanney

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Statistics
University of Manitoba
Winnipeg

Dedication

To my mother

Abstract

Zero inflation occurs when the proportion of zeros of a model is greater than the proportion of zeros of the corresponding Poisson model. This situation is very common in count data. In order to model zero inflated count time series data, we propose the zero inflated autoregressive conditional Poisson (ZIACP) model by the extending the autoregressive conditional poisson (ACP) model of Ghahramani and Thavaneswaran (2009). The stationarity conditions and the autocorrelation functions of the ZIACP model are provided. Based on the expectation maximization (EM) algorithm an estimation method is developed. A simulation study shows that the estimation method is accurate and reliable as long as the sample size is reasonably high. Three real data examples, syphilis data Yang (2012), arson data Zhu (2012) and polio data Kitromilidou and Fokianos (2015) are studied to compare the performance of the proposed model with other competitive models in the literature.

$Keywords$: count data, times series of counts, zero inflation, Poisson, negative binomial, EM algorithm, observation driven, parameter driven.

Acknowledgments

I am deeply indebted to my supervisors, Prof. A. Thavaneswaran and Dr. S. Hossain for their patience, guidance and financial support from the NSERC grant during my masters studies. I would also like to thank them for spending considerable amount of time, their infectious enthusiasm for statistical research and introducing me to an interesting area in statistics. I feel very fortunate to have had Prof. Thavaneswaran and Dr. Hossain as my M.Sc. supervisors.

I would also like to thank my Advisory Committee member, Prof. Alexandre Leblanc for reading my lengthy thesis with patience and his helpful suggestions, criticisms and useful discussions on my thesis. My thanks also go to the external examiner Prof. S. Srimantoorao Appadoo for careful reading of my thesis. I will want to show my sincerest gratitude to Prof. Konstatinos Fokainos for his helpful suggestions in the course of writing my thesis. Prof. Fokianos kindly provided me with an **R** code for analyzing the transactions data in Fokianos et al. (2009). This was really helpful for me when I was writing the simulation program for the zero inflated Poisson with time varying parameter in this thesis. I am also thankful to Dr. Ming Yang for providing me with assistance on how the response variables of the zero inflated autoregressive process was generated. A big thank you go to Prof. Fukang Zhu for helping me in accessing the arson data set.

A big thank you goes to the Department of Statistics for awarding me the Faculty of Science Studentship for the period of 2013-2014 and recommending me for the University of Manitoba Graduate Fellowship. I am also grateful to the department for giving me the opportunity to teach in the statistics lab. Many thanks also go to the departmental support staff, Ms. Margaret Smith and Ms. Stana Drobko.

Last but not least, I would like to thank my parents for their constant encouragement and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Count response variables that take non-negative integer values without explicit upper bound are frequently encountered in empirical data analysis. Zero inflated data cannot be modelled by the usual Poisson distribution. It became very popular over the last decade to model zero inflated data using a mixture of a count distribution with a degenerate distribution supported at zero. Real valued time series models, such as the generalized autoregressive conditional heteroscedastic (GARCH) model, introduced by Bollerslev (1986) have been used in many applications. The time varying mean of the Poisson process have been modelled by Fokianos et al. (2009) and Ghahramani and Thavaneswaran (2009). Modeling time varying parameters is very important in count time series analysis. When modeling these parameters we may want to fully understand their behavior with respect to time. Our interest is to estimate the fixed parameters that govern the behavior of these time varying parameters of interest. When the conditional distribution of a time series is a function of its past observations, then the series is said to be observation driven. For example consider the GARCH model below

$$y_t = \sqrt{h_t} Z_t$$

$$h_t = w + \sum\nolimits_{i=1}^{p} \alpha_i y_{t-i}^2 + \sum\nolimits_{j=1}^{q} \beta_j h_{t-j}, \tag{1.1}$$

where Z_t is a sequence of independent, identically distributed random variables with zero mean and unit variance. Here h_t is the time varying parameter which turns out to be the conditional variance of the GARCH model. Essentially h_t has been modeled as a function of squared values of past observations and past values of itself. In general the GARCH(p, q) model for a time series y_t has the form (1.1). The corresponding ARMA model for y_t^2 can be written as

$$\Phi(B) y_t^2 = w + \beta(B) u_t, \tag{1.2}$$

where $u_t = y_t^2 - h_t$ is the martingale difference. We denote the variance of u_t as σ_u^2. $\Phi(B) = 1 - \sum_{i=1}^{r} \Phi_i B^i,\ \Phi_i = (\alpha_i + \beta_i),\ \beta(B) = 1 - \sum_{i=1}^{q} \beta_j B^j$ and $r = max(p, q)$. B is the back shift operator such that $By_t = y_{t-1}$. This representation is used to obtain the high order moments of the GARCH process y_t.

Recently, a new class of models that appears to be more informative has been developed by Harvey (2013), the generalized autoregressive score (GAS) models. Here the time varying parameter is modeled as a function of past values of itself and past values of the score resulting from the conditional likelihood. This model is very important since it has most of the commonly used time series models as special cases. For example if $f(y_t|\mu_t;\theta)$ follows the normal or the $t-$distribution and μ_t models the scale, then the GAS model reduces to a normal GARCH or a $t-$GARCH, respectively. In the same way, if $f(y_t|\mu_t;\theta)$

is a Poisson or a zero inflated Poisson and μ_t models the conditional mean, then the GAS model reduces to the Poisson or a ZIP model with time varying parameter, respectively.

The Poisson and the negative binomial distribution are well noted for modelling count data. The Poisson distribution in particular has the fundamental property that the mean counts is equal to the variance of the counts. This property rarely holds in the case of a zero inflated data. As such the normal/ordinary Poisson distribution fails to account for the extra over dispersion in the data introduced by the inflation of the zero counts. This situation is the motivation for the use of zero inflated models. In these models, the Poisson and the negative binomial distributions are modified to be able to account for the extra over dispersion exhibited by the zero inflated data. Therefore, there is need to develop an efficient way of estimating model parameters to explain the zero inflated count time series data. Failure to account for the zero-inflation in the data may lead to misleading inference and unreliable predictions. There has also been concern whether zero inflated distributions are necessary especially considering their estimation complexity. Particularly, Allison (2012) argued that, the negative binomial distribution models zero inflated data well enough so that it is not worth using the zero inflated distributions. He justified this argument by examples on many different real datasets. The applications of zero-inflated models have been found in many practical situations where excess zero observations are generated. This thesis concentrates on efficient estimation strategies for analyzing zero inflated count time series data which are mostly encountered in many biomedical and public health applications. For example, when a rare infectious disease occurs overtime, public health officials may be interested in monitoring the observed counts. Diseases with low infection rates will normally exhibit a high incident of zeros (zero inflation). The opposite is true that observed counts recorded can be very high during an outbreak.

1.1.1 Literature Review

Developments in the field of time series saw an interesting turn when for the first time autoregressive conditional heteroskedasticity (ARCH) models were introduced by Engle (see Engle and Russell, 1982). This idea was later generalized by Bollerslev (1986) making the ARCH model a special case of the generalized autoregressive conditional heteroskedasticity (GARCH) model. The first order GARCH (1,1) model of the time varying volatility h_t has the form

$$y_t = \sqrt{h_t} Z_t \tag{1.3}$$

$$h_t = w + \alpha y_{t-1}^2 + \beta h_{t-1}, \;\; \omega > 0, \;\; \alpha \geq 0, \;\; \beta \geq 0,$$

where Z_t is a sequence of independent, identically distributed random variables with zero mean and unit variance. The condition imposed on the model parameters α and β ensures that the conditional variance h_t is positive. The model (1.3) simplifies to the ARCH model when $\beta = 0$. Here α and β are chosen in such a way that their sum is close to one in order to ensure stationarity of the ARMA model in terms of y_t^2 observations. The integrated GARCH ($IGARCH$) model is obtained when their sum is one. GARCH models have been the principal means of analyzing, modeling and monitoring volatility changes especially for financial returns. The GAS models Creal et al. (2013) which are similar to the GARCH models are observation driven models and form the latest family of models that has been developed. In this new approach, the mechanism to update the parameters over time is the scaled score of the likelihood function. This way of modeling provides a unified and consistent framework for introducing time varying parameters in a wide range of nonlinear models. The GARCH, autoregressive conditional duration (ACD), autoregressive conditional intensity (ACI), and Poisson count models with time varying mean are all special cases of the proposed GAS models . The conditional score drives the dynamics of the model. These models have interesting features which make them easy to deal with. The likelihood for these models are available in closed form thus allowing for estimation and inference of the parameters of interest. In the next section, we provide an outline of the Thesis.

1.1.2 Overview of the Thesis

The remainder of Chapter one summarizes some key concepts from Harvey's book on dynamic models for volatility and heavy tails (Harvey, 2013) to give the necessary background for the Thesis. A brief introduction to GAS models is then given, where we model the conditional mean or the conditional variance as a function of past values of the process and past values of the score which is based on the conditional likelihood. We then verify the asymptotic distribution of the fixed parameters governing the dynamics of the model as provided in Harvey (2013).

Chapter two provides a summary of the likelihood based inference for linear and nonlinear Poisson autoregression of Fokianos et al. (2009) and conduct a simulation study to demonstrate the efficiency of the modelling approach. We also describe the inference procedure for the zero inflated Poisson (ZIP) distribution and later extend the ideas of the Poisson autoregression as well as the ZIP distribution to estimate the parameters of the zero inflated autoregressive conditional Poisson model (ZIACP).

Chapter three considers estimation of the parameters in the zero inflated Poisson autoregression model for time series count data when there are many potential predictors and some of them may not have influence on the response of interest. In the context of two competing models where one model includes all covariates and the other restricts variable coefficients to a linear restriction based on auxilliary information or prior knowledge. We investigate the relative performances of shrinkage and pretest estimators with respect to the unrestricted maximum likelihood estimator (UMLE). The asymptotic properties of the pretest and shrinkage estimators including the derivation of asymptotic distributional biases and risks are established. A Monte Carlo simulation study is conducted to examine the relative performance of the shrinkage and pretest estimators with the UMLE.

Chapter four provides a conclusion to the Thesis and possible future reseach interests.

1.1.3 Generalized Autoregressive Score Models

If y_t has a conditional distribution $f(y_t|\mu_t;\theta)$ and assuming we are interested in modeling the conditional mean as a time varying parameter, then the GAS model has the form

$$y_t \sim f(y_t|\mu_t;\boldsymbol{\theta}) \tag{1.4}$$

$$\mu_{t+1} = \omega + \sum_{i=1}^{p} A_i u_{t-i+1} + \sum_{j=1}^{q} B_j \mu_{t-j+1}, \tag{1.5}$$

where μ_t is the time varying parameter, $\boldsymbol{\theta}$ is a vector of unknown fixed parameters. It is possible for the conditional distribution (1.4) to depend on additional covariates. However, for the sake of simplicity this will not be considered. If u_t is a scaled score function, then

$$u_t = k_t \times \frac{\partial \log f(y_t|\mu_t;\theta)}{\partial \mu_t}, \tag{1.6}$$

where k_t is a user defined scaling matrix like the Fisher information matrix,

$$k_t = -E_{t-1}\left(\frac{\partial \log f(y_t|\mu_t;\boldsymbol{\theta})}{\partial \mu_t}\right)\left(\frac{\partial \log f(y_t|\mu_t;\boldsymbol{\theta})}{\partial \mu_t}\right)^{\top}.$$

1.1.4 Gaussian GAS models

If y_t observations have conditional distribution such that the log-density is of the form

$$\log f(y_t|\sigma_t^2;\theta) = -\frac{\log(2\pi)}{2} - \frac{\log\sigma_t^2}{2} - \frac{y_t^2}{2\sigma_t^2},$$

and σ_t^2 is the conditional variance of y_t, then the score function can be obtained as

$$\frac{\partial \log f(y_t|\sigma_t^2;\theta)}{\partial\sigma_t^2} = -\frac{1}{2\sigma_t^2} + \frac{y_t^2}{2\sigma_t^4},$$

$$\implies \frac{\partial^2 \log f(y_t|\sigma_t^2;\theta)}{\partial(\sigma_t^2)^2} = \frac{1}{2\sigma_t^4} - \frac{y_t^2}{\sigma_t^6}.$$

Therefore

$$-E\left(\frac{\partial^2 \log f(y_t|\sigma_t^2;\theta)}{\partial(\sigma_t^2)^2}\right) = \frac{1}{2\sigma_t^4},$$

and $\left[-E\left(\frac{\partial^2 \log f(y_t|\sigma_t^2;\theta)}{\partial(\sigma_t^2)^2}\right)\right]^{-1} = 2\sigma_t^4.$

If $k_t = 2\sigma_t^4 \implies k_t \times \frac{\partial \log f(y_t|\sigma_t^2;\theta)}{\partial\sigma_t^2} = y_t^2 - \sigma_t^2 = u_t$. We can write (1.5) as

$$\sigma_{t+1}^2 = \omega + A(y_t^2 - \sigma_t^2) + B\sigma_t^2.$$

Therefore, when the observations follow a normal distribution and $\mu_t = \sigma_t^2$, the GAS(1,1) model reduces to the GARCH(1,1) model. However, typically the GARCH(1,1) is parameterized as

$$\mu_{t+1} = \omega + \alpha y_t^2 + \beta \mu_t.$$

where $\alpha = A$ and $\beta = B - A$.

It is interesting to note that depending on the choice of link function, the parameterization of the time varying parameter changes accordingly. For example, when $\mu_t = \log(\sigma_t^2)$, the score, the inverse of the fisher information matrix, and GAS(1,1) model for the conditional distribution are,

$$\frac{\partial \log f(y_t|\mu_t;\theta)}{\partial \mu_t} = \frac{y_t^2}{2exp(\mu_t)} - \frac{1}{2}$$

$$\left[-E\left(\frac{\partial^2 \log f(y_t|\mu_t;\theta)}{(\partial \sigma_t^2)^2}\right)\right]^{-1} = 2$$

$$\mu_{t+1} = \omega + A\left(\frac{y_t^2}{exp(\mu_t)} - 1\right) + B\mu_t.$$

Taking into account that financial returns typically exhibit heavy tails i.e., extreme values occur from time to time, the Dynamic Conditional Score models introduced by Harvey (2013) shows how a radical change in the way GARCH models are formulated leads to a resolution of many inherent problems associated with statistical theory.

1.1.5 Student's t Distribution

The probability density function (pdf) of the t-distribution is given by,

$$f(y, \mu, \varphi, \nu) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\left(1 + \frac{(y-\mu)^2}{\nu\varphi^2}\right)^{-(\nu+1)/2} \varphi, \nu > 0, \tag{1.7}$$

where ν, ϕ and μ are the degree of freedom, scale, and location parameters, respectively. Here $\Gamma(.)$ is the gamma function. Moments exists only up to and including order $\nu - 1$. Since the distribution is symmetric, the mean is equal to the median which always exists. The mean is finite when $\nu > 1$. For $\nu > 2$, the variance is

$$\sigma^2 = \left(\frac{\nu}{\nu - 2}\right)\varphi^2.$$

As a special case, the Cauchy distribution is a t- distribution with degree of freedom equal to one and has no moments. Its pdf is

$$f(y) = \frac{1}{\pi\varphi}\left(1 + \frac{(y-\mu)^2}{\varphi^2}\right)^{-1}.$$

Lemma 1. [*Lemma* 2 *of* Harvey (2013)] *The expectation of the absolute value of a standardized* t_ν *variate* ε_t*, raised to a power c is*

$$E(|\varepsilon_t|^c) = \nu^{c/2}\Gamma(c/2 + 1/2)\Gamma(-c/2 + \nu/2)/(\Gamma(1/2)\Gamma(\nu/2), \quad -1 < c < \nu.$$

Proof. Since the ratio of two chi-square distributions results in an F distribution, the problem reduces to finding the moment of the F distribution raised to the power $c/2$ Therefore referring to

$$E(F^k) = \frac{\Gamma(\frac{m+2k}{2})\Gamma(\frac{n-2k}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\left(\frac{n}{m}\right)^k. \tag{1.8}$$

For $k = c/2, m = 1, n = \nu$.

$E(|\varepsilon_t|^c) = E(\{\varepsilon_t^2\}^{c/2}) = E(\{\chi_1^2/(\chi_\nu^2/\nu)\}^{c/2}) = \Gamma((1+c)/2)\Gamma((\nu-c)/2)/(\Gamma(1/2)\Gamma(\nu/2))\nu^{c/2}$.

□

Lemma 2. [*Lemma* 3 *of* Harvey (2013)] *If* $a \sim gamma(\theta, \alpha)$ *and* $b \sim gamma(\theta, \beta)$ *with* a *and* b *independent of each other, then* $y = a/(b+a) \sim B(\alpha, \beta)$.

Corollary 2.1. *The variable* $(t^2/\nu)/(1+t^2/\nu)$ *has a* $B(1/2, \nu/2)$ *distribution whereas* $1/(1+t^2/\nu)$ *has* $B(\nu/2, 1/2)$ *distribution.*

Proof. $t^2/\nu = \chi_1^2/\chi_\nu^2 \implies (t^2/\nu)/(1+t^2/\nu) = \frac{\chi_1^2/\chi_\nu^2}{1+\chi_1^2/\chi_\nu^2} = \frac{\chi_1^2}{\chi_1^2+\chi_\nu^2}$, hence from lemma (2) we have, $(t^2/\nu)/(1+t^2/\nu) \sim B(1/2, \nu/2)$, similarly $(1)/(1+t^2/\nu) = 1-(t^2/\nu)/(1+t^2/\nu) = 1 - B(1/2, \nu/2) = B(\nu/2, 1/2)$. □

1.1.6 Maximum Likelihood Estimates

Suppose $y_t, t = 1, \cdots n$ is a set of independent observations, each from a distribution with pdf $f(y_t; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes a vector of parameters that is of interest. If the observations are independent and identically distributed, the likelihood function reduces to the product of the individual denstiy functions (Harvey, 2013). For practical and theoretical reasons it is tractable to work with the logarithm of the likelihood function.

$$\log \mathcal{L}(\boldsymbol{\theta}; y_1 \cdots, y_n) = \sum_{t=1}^{n} \log f(y_t; \boldsymbol{\theta}).$$

The maximum likelihood principle finds the value of $\boldsymbol{\theta}$ that makes the sample most likely. The global maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ maximizes $\log \mathcal{L}(\boldsymbol{\theta})$ over the full parameter space. The estimate $\hat{\boldsymbol{\theta}}$ can be obtained by solving the score equation below:

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial(\boldsymbol{\theta})} = \mathbf{0}.$$

We can write the information matrix for a single observation as

$$I(\boldsymbol{\theta_0}) = E_0\left(\frac{\partial \log f}{\partial \boldsymbol{\theta}}\frac{\partial \log f}{\partial \boldsymbol{\theta'}}\right) = -E_0\left(\frac{\partial^2 \log f}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta'}}\right),$$

where the expectation is taken at the true value of the parameter $\boldsymbol{\theta}$ denoted by $\boldsymbol{\theta}_0$. The full information matrix is $\{n \times \boldsymbol{I}(\boldsymbol{\theta}_0)\}$ for the independent case where n emphasizes the sample size. Under certain regularity conditions such as the uniqueness of the maximum likelihood estimate (MLE), the existence of moments to at least the third order, positive definiteness of the Fisher information matrix, the MLE $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$ and it is asymptotically normal in the sense that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a multivariate normal with a zero vector mean and covariance matrix $\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)$ which is positive definite, provided that the model is identifiable. In the case of a time series, by means of a clever conditioning we could use the Partial Likelihood (PL) to transport the inferential feature appropriate for independent data to dependent data. Consider a time series $y_t, t = 1, \cdots n$ with the joint density $f_{\boldsymbol{\theta}}(y_1, y_2, y_3, \cdots, y_n)$ where $\boldsymbol{\theta}$ constitutes a parameter vector. Suppose the existence of some auxiliary information $(\boldsymbol{AI})$ that is known throughout the period the time series was observed. Then we can write the likelihood as a function of $\boldsymbol{\theta}$ by the equation

$$f_{\boldsymbol{\theta}}(y_1, y_2, y_3, \cdots, y_n|\boldsymbol{AI}) = f_{\boldsymbol{\theta}}(y_1|\boldsymbol{AI})\prod_{t=2}^{n} f_{\boldsymbol{\theta}}(y_t|y_1, y_2, y_3, \cdots, y_{t-1}, \boldsymbol{AI}) \tag{1.9}$$

In the event that the auxiliary information is not available or irrelevant, it can be dropped from equation (1.9), simplifying it to

$$f_{\boldsymbol{\theta}}(y_1, y_2, y_3, \cdots, y_n) = f_{\boldsymbol{\theta}}(y_1)\prod_{t=2}^{n} f_{\boldsymbol{\theta}}(y_t|y_1, y_2, y_3, \cdots, y_{t-1}) \tag{1.10}$$

using the Markovian assumption (See details, Kedem and Fokianos, 2005) the joint density can be written as in

$$f_{\boldsymbol{\theta}}(y_1, y_2, y_3, \cdots, y_n) = f_{\boldsymbol{\theta}}(y_1) \prod_{t=2}^{n} f_{\boldsymbol{\theta}}(y_t | y_{t-1}) \tag{1.11}$$

We could ignore the first factor $f_{\boldsymbol{\theta}}(y_1)$, as it does not depend on n and inference can be made about $\boldsymbol{\theta}$ based on the product term in (1.11)

1.1.7 Maximum Likelihood Estimation of Dynamic Linear Models

We consider initially a static model with only one parameter θ such that, the scaled score u_t is given by a product of a scalar k and the derivative of the loglikelihood fuction as in,

$$u_t = k \frac{\partial \log f(y_t; \theta)}{\partial \theta}, \quad t = 1, \cdots, n,$$

where k is a finite constant. The derivative $\partial \log f(y_t; \theta)/\partial \theta$ is a random variable with zero mean at the true parameter value, θ_0 so as u_t. Let σ_u^2 denote the variance of u_t which is finite under standard regularity conditions. The information quantity for this model for a single observation can be written as below

$$I(\theta_0) = -E\left(\frac{\partial^2 \log f}{\partial \theta^2}\right) = E\left[\left(\frac{\partial \log f}{\partial \theta}\right)^2\right] = E(u_t^2)/k^2 < \infty. \tag{1.12}$$

Interestingly this information quantity does not depend on θ, the parameter of interest.
Condition 1 [Harvey (2013), page 32] *The variance of the score in the static model is finite and does not depend on* θ_0.
However, if θ is allowed to be a time varying parameter $\theta_{t|t-1}$ (i.e., adopting the notation

in Harvey (2013)) such that it evolves over time as a function of past observations and past values of the scaled score of the conditional distribution, and the conditional score depends on past observations through $\theta_{t|t-1}$, $\frac{\partial \log f(y_t|Y_{t-1};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$ can be broken down into two parts as in

$$\frac{\partial \log f(y_t|Y_{t-1};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \frac{\partial \log f(y_t;\theta_{t|t-1})}{\partial \theta_{t|t-1}} \frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}. \tag{1.13}$$

where $\boldsymbol{\psi}$ are fixed parameters of the time varying parameter $\theta_{t|t-1}$.

Lemma 3. [*Lemma* 5 of Harvey (2013)] *Consider a model with a single time-varying parameter, $\theta_{t|t-1}$, which satisfies an equation that depends on variables which are fixed at time $t-1$. The process is governed by a set of fixed parameters, $\boldsymbol{\psi}$. If Condition 1 holds, then the conditional score for the t-th observation, $\partial \log f_t(y_t|Y_{t-1};\boldsymbol{\psi})/\partial \boldsymbol{\psi}$ is a martingale difference at $\boldsymbol{\psi} = \boldsymbol{\psi}_0$, with conditional covariance matrix*

$$E_{t-1}\left(\frac{\partial \log f_t(y_t|Y_{t-1};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right)\left(\frac{\partial \log f_t(y_t|Y_{t-1};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right)^{\top} = \boldsymbol{I} \times \left(\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}} \frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}'}\right), \tag{1.14}$$

$t = 1, \cdots, n$ where the information quantity, $\boldsymbol{I}$, is constant over time and independent of $\boldsymbol{\psi}$.

Proof: Let $\theta = \theta_{t|t-1}$ evolve over time as a function of past observations and past values of the scaled score of the conditional distribution. Since the conditional score depends on past observations through $\theta_{t|t-1}$, it can be written in two parts as

$$\frac{\partial \log f_t(y_t|Y_{t-1};\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \frac{\partial \log f_t(y_t;\theta_{t|t-1})}{\partial \theta_{t|t-1}} \frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}. \tag{1.15}$$

Since the derivative of the time-varying parameter, i.e., $\partial \theta_{t|t-1}/\partial \boldsymbol{\psi}$ is fixed at time $t-1$ and the expected value of the score in the static model is zero, the score (1.15) is a martingale difference. The conditional covariance matrix is obtained by writing its outer product as

$$\left(\frac{\partial \log f_t(y_t;\theta_{t|t-1})}{\partial \theta_{t|t-1}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\right)\left(\frac{\partial \log f_t(y_t;\theta_{t|t-1})}{\partial \theta_{t|t-1}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\right)^\top \tag{1.16}$$

$$= \left(\frac{\partial \log f_t}{\partial \theta_{t|t-1}}\right)^2\left(\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}'}\right).$$

Since $\theta_{t|t-1}$ and its derivatives depend only on past information, the distribution of the score conditional on available information at time $t-1$ is the same as its unconditional distribution and so time invariant. Taking the expectation of (1.16) conditional on information available at time $t-1$, $E_{t-1}(\partial \log f_t/\partial\theta_{t|t-1})^2$ is as in the static model considered above and equal to the unconditional expectation in the static model, that is (1.12). Because $\theta_{t|t-1}$ is fixed at time $t-1$ and hence static,

$$E_{t-1}\left[\left(\frac{\partial \log f_t(y_t;\theta_{t|t-1})}{\partial \theta_{t|t-1}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\right)\left(\frac{\partial \log f_t(y_t;\theta_{t|t-1})}{\partial \theta_{t|t-1}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\right)'\right]$$

$$= \left[E\left(\frac{\partial \log f_t}{\partial \theta_{t|t-1}}\right)^2\right]\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}'}.$$

Corollary 3.1. [*Corollary 4 of Harvey* (2013)] *The information matrix in the context of Lemma 3 is*

$$I(\boldsymbol{\psi}) = I \times \boldsymbol{D}(\boldsymbol{\psi}),$$

where

$$\boldsymbol{D}(\boldsymbol{\psi}) = E\left(\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}}\frac{\partial \theta_{t|t-1}}{\partial \boldsymbol{\psi}'}\right).$$

Here we derive the information matrix at time t for the fixed parameters of the first order model

$$\theta_{t+1|t} = \delta + \phi\theta_{t|t-1} + \kappa u_t, \;\; |\phi| < 1, \;\; \kappa \neq 0, \;\; t = 1, \cdots, n. \tag{1.17}$$

Unless ϕ is known to be zero, the condition $\kappa \neq 0$ is necessary for model identifiability. The condition $|\phi| < 1$ ensures that the process is stationary and hence enables $\theta_{t|t-1}$ to be expressed as an infinite moving average in terms of u_t's. Because the u_t's are martingale differences and hence white noise, the process $\theta_{t|t-1}$ is weakly stationary, with an unconditional mean of $\omega = \delta/(1-\phi)$ and an unconditional variance $\kappa^2\sigma_u^2/(1-\phi^2)$.

Rewriting (1.17) as

$$\theta_{t+1|t} = \omega(1-\phi) + \phi\theta_{t|t-1} + \kappa u_t, \tag{1.18}$$

we have

$$\frac{\partial\theta_{t+1|t}}{\partial\kappa} = \phi\frac{\partial\theta_{t|t-1}}{\partial\kappa} + \kappa\frac{\partial u_t}{\partial\kappa} + u_t$$

$$\frac{\partial\theta_{t+1|t}}{\partial\phi} = \phi\frac{\partial\theta_{t|t-1}}{\partial\phi} + \kappa\frac{\partial u_t}{\partial\phi} + \theta_{t|t-1} - \omega$$

$$\frac{\partial\theta_{t+1|t}}{\partial\omega} = \phi\frac{\partial\theta_{t|t-1}}{\partial\omega} + \kappa\frac{\partial u_t}{\partial\omega} + 1 - \phi.$$

But we can express $\frac{\partial u_t}{\partial\kappa} = \frac{\partial u_t}{\partial\theta_{t|t-1}}\frac{\partial\theta_{t|t-1}}{\partial\kappa}$, $\frac{\partial u_t}{\partial\phi} = \frac{\partial u_t}{\partial\theta_{t|t-1}}\frac{\partial\theta_{t|t-1}}{\partial\phi}$, $\frac{\partial u_t}{\partial\omega} = \frac{\partial u_t}{\partial\theta_{t|t-1}}\frac{\partial\theta_{t|t-1}}{\partial\omega}$, so that

$$\frac{\partial\theta_{t+1|t}}{\partial\kappa} = x_t\frac{\partial\theta_{t|t-1}}{\partial\kappa} + u_t \tag{1.19}$$

$$\frac{\theta_{t+1|t}}{\partial\phi} = x_t\frac{\partial\theta_{t|t-1}}{\partial\phi} + \theta_{t|t-1} - \omega \tag{1.20}$$

$$\frac{\partial\theta_{t+1|t}}{\partial\omega} = x_t\frac{\partial\theta_{t|t-1}}{\partial\omega} + 1 - \phi, \tag{1.21}$$

where $x_t = \phi + \kappa \frac{\partial u_t}{\partial \theta_{t|t-1}}, \;\; t = 1, 2, \cdots, n.$

Condition 2 [Harvey (2013), page 35] *For the static model, the score and its first derivative, or equivalently* u_t *and* u_t'*, where* $u_t' = \partial u_t / \partial \theta$*, have finite second moments and covariance that are time-invariant and do not depend on* θ*, that is,* $E(u_t^{2-k} u_t'^k) < \infty, \;\; k = 0, 1, 2 \cdots$.

The implications of the preceding condition is that, $E(u_t u_t') < \infty$, $E(u_t'^2) < \infty$ as well as $E(u_t^2) < \infty$. In view of Condition 2, the expectations below are valid, in that they are time invariant, the unconditional expectations can replace the conditional ones.

$$a = E_{t-1}(x_t) = \phi + \kappa E_{t-1}\left(\frac{\partial u_t}{\partial \theta_{t|t-1}}\right) = \phi + \kappa E\left(\frac{\partial u_t}{\partial \theta}\right). \tag{1.22}$$

$$b = E_{t-1}(x_t^2) = \phi^2 + 2\phi\kappa E\left(\frac{\partial u_t}{\partial \theta}\right) + \kappa^2 E\left(\frac{\partial u_t}{\partial \theta}\right)^2 \geq 0.$$

$$c = E_{t-1}(u_t x_t) = \kappa E\left(u_t \frac{\partial u_t}{\partial \theta}\right).$$

Lemma 4. [*Lemma* $6 \;\; of \;\; Harvey$ (2013)] *When the process for* $\theta_{t|t-1}$ *starts in the infinite past and provided that* $|a| < 1$*, the equations in (1.19) can be viewed as an* $AR(1)$ *process. Hence*

$$E\left(\frac{\partial \theta_{t+1|t}}{\partial \kappa}\right) = 0. \;\;\; t = 1, 2, \cdots, n \tag{1.23}$$

$$E\left(\frac{\partial \theta_{t+1|t}}{\partial \phi}\right) = 0.$$

$$E\left(\frac{\partial \theta_{t+1|t}}{\partial \omega}\right) = \frac{1-\phi}{1-a}.$$

Theorem 5. [*Theorem 1 of Harvey (2013)*] *Assume that Condition 2 holds and that* $b < 1$. *Then the information matrix for a single observation is time-invariant and given by*

$$I(\boldsymbol{\psi}) = I \times \boldsymbol{D}(\boldsymbol{\psi}) = (\sigma_u^2/k^2)(\boldsymbol{D}(\boldsymbol{\psi})), \tag{1.24}$$

where

$$\boldsymbol{D}(\boldsymbol{\psi}) = \boldsymbol{D}\begin{pmatrix} \kappa \\ \phi \\ \omega \end{pmatrix} = \frac{1}{1-b}\begin{bmatrix} A & D & E \\ D & B & F \\ E & F & C \end{bmatrix}$$

with

$A = \sigma_u^2$, $B = \frac{k^2\sigma_u^2(1+a\phi)}{(1-\phi^2)(1-a\phi)}$, $C = \frac{(1-\phi)^2(1+a)}{1-a}$, $D = \frac{a\kappa\sigma_u^2}{1-a\phi}$, $E = \frac{c(1-\phi)}{1-a}$, *and* $F = \frac{ac\kappa(1-\phi)}{(1-a)(1-a\phi)}$.

Proof. From (1.19) above, by squaring and applying expectation we obtain

$$\frac{\partial\theta_{t+1|t}}{\partial\kappa} = x_t\frac{\partial\theta_{t|t-1}}{\partial\kappa} + u_t,$$

$$\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}\right)^2 = x_t^2\left(\frac{\partial\theta_{t|t-1}}{\partial\kappa}\right)^2 + u_t^2 + 2x_tu_t\frac{\partial\theta_{t|t-1}}{\partial\kappa},$$

$$E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}\right)^2 = b\left(\frac{\partial\theta_{t|t-1}}{\partial\kappa}\right)^2 + \sigma_u^2 + 2c\frac{\partial\theta_{t|t-1}}{\partial\kappa},$$

$$E_{t-2}E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}\right)^2 = bE_{t-2}\left(\frac{\partial\theta_{t|t-1}}{\partial\kappa}\right)^2 + \sigma_u^2 + 2cE_{t-2}\left(\frac{\partial\theta_{t|t-1}}{\partial\kappa}\right),$$

$$\Longrightarrow \quad \lim_{n\to\infty}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}\right)^2 = \frac{\sigma_u^2}{1-b}.$$

Similarly from (1.20) we apply expection to the terms after squaring

$$\frac{\partial\theta_{t+1|t}}{\partial\phi} = x_t\frac{\partial\theta_{t|t-1}}{\partial\phi} + \theta_{t|t-1} - \omega,$$

$$E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\phi}\right)^2 = b\left(\frac{\partial\theta_{t|t-1}}{\partial\phi}\right)^2 + (\theta_{t|t-1} - \omega)^2 + 2a\frac{\partial\theta_{t|t-1}}{\partial\kappa}(\theta_{t|t-1} - \omega),$$

$$E_{t-2}E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\phi}\right)^2 = bE_{t-2}\left(\frac{\partial\theta_{t|t-1}}{\partial\phi}\right)^2 + \frac{\sigma_u^2k^2}{1-\phi^2} + 2aE_{t-2}\left(\frac{\partial\theta_{t|t-1}}{\partial\phi}(\theta_{t|t-1} - \omega)\right).$$

here we evaluate the expectations of $\left(\frac{\partial\theta_{t|t-1}}{\partial\phi}\right)^2$ and $\left(\frac{\partial\theta_{t|t-1}}{\partial\phi}(\theta_{t|t-1} - \omega)\right)$ further and substitute them back afterwards by writing $\theta_{t|t-1}$ in $\frac{\partial\theta_{t|t-1}}{\partial\phi}$ like the defining equation in (1.17).

$$\theta_{t|t-1} = \omega(1-\phi) + \phi\theta_{t-1|t-2} + \kappa u_{t-1}$$

$$\theta_{t|t-1} - \omega = \phi(\theta_{t-1|t-2} - \omega) + \kappa u_{t-1}$$

$$\frac{\partial\theta_{t|t-1}}{\partial\phi} = x_{t-1}\frac{\partial\theta_{t-1|t-2}}{\partial\phi} + \theta_{t-1|t-2} - \omega.$$

Therefore,

$$\begin{aligned}E_{t-2}\left(\frac{\partial\theta_{t|t-1}}{\partial\phi}(\theta_{t|t-1} - \omega)\right) = E_{t-2}\Bigg[&\left(x_{t-1}\frac{\partial\theta_{t-1|t-2}}{\partial\phi} + \theta_{t-1|t-2} - \omega\right)\\ &\times\left(\phi(\theta_{t-1|t-2} - \omega) + \kappa u_{t-1}\right)\Bigg]\end{aligned}$$

$$\implies E_{t-2}\left(\frac{\partial\theta_{t|t-1}}{\partial\phi}(\theta_{t|t-1}-\omega)\right) = \phi E_{t-2}\left(x_{t-1}\frac{\partial\theta_{t-1|t-2}}{\partial\phi}(\theta_{t-1|t-2}-\omega)\right)$$

$$+ E_{t-2}\left(\kappa u_{t-1}x_{t-1}\frac{\partial\theta_{t-1|t-2}}{\partial\phi}\right) + \phi E_{t-2}[(\theta_{t-1|t-2}-\omega)^2]$$

$$+ E_{t-2}[\kappa u_{t-1}(\theta_{t-1|t-2}-\omega)]. \quad (1.25)$$

The last term is zero and the penultimate term will eventually be zero as well. Therefore (1.25) becomes

$$E_{t-2}\left(\frac{\partial\theta_{t|t-1}}{\partial\phi}(\theta_{t|t-1}-\omega)\right) = a\phi E_{t-2}\left(\frac{\partial\theta_{t-1|t-2}}{\partial\phi}(\theta_{t-1|t-2}-\omega)\right) + \frac{\phi\sigma_u^2\kappa^2}{1-\phi^2}.$$

Since

$$|a| < 1 \implies E\left(\frac{\partial\theta_{t|t-1}}{\partial\phi}(\theta_{t|t-1}-\omega)\right) = \frac{\phi\sigma_u^2\kappa^2}{(1-a\phi)(1-\phi^2)}.$$

Therefore,

$$E\left(\frac{\partial\theta_{t+1|t}}{\partial\phi}\right)^2 = \frac{1}{1-b}\left[\frac{\sigma_u^2\kappa^2}{1-\phi^2} + 2a\left(\frac{\phi\sigma_u^2\kappa^2}{(1-a\phi)(1-\phi^2)}\right)\right]$$

$$= \frac{1}{1-b}\left[\frac{k^2\sigma_u^2(1+a\phi)}{(1-\phi^2)(1-a\phi)}\right].$$

$$\frac{\partial\theta_{t+1|t}}{\partial\omega} = x_t\frac{\partial\theta_{t|t-1}}{\partial\omega} + 1 - \phi$$

$$\left(\frac{\partial\theta_{t+1|t}}{\partial\omega}\right)^2 = (1-\phi)^2 + x_t^2\left(\frac{\partial\theta_{t|t-1}}{\partial\omega}\right)^2 + 2(1-\phi)x_t\frac{\partial\theta_{t|t-1}}{\partial\omega}$$

$$E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\omega}\right)^2 = b\left(\frac{\partial\theta_{t|t-1}}{\partial\omega}\right)^2 + (1-\phi)^2 + 2a(1-\phi)\frac{\partial\theta_{t|t-1}}{\partial\omega}$$

$$\implies \quad E\left(\frac{\partial\theta_{t+1|t}}{\partial\omega}\right)^2 = \frac{(1-\phi)^2(1+a)}{1-a}.$$

Since $E\left(\frac{\partial\theta_{t+1|t}}{\partial\omega}\right) = \frac{1-\phi}{1-a}$ and now obtaining the cross terms of the information matrix

matrix

$$E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}.\frac{\partial\theta_{t+1|t}}{\partial\phi}\right) = E_{t-1}\left[\left(x_t\frac{\partial\theta_{t|t-1}}{\partial\kappa} + u_t\right)\left(x_t\frac{\partial\theta_{t|t-1}}{\partial\phi} + \theta_{t|t-1} - \omega\right)\right]$$

$$\begin{aligned} E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}.\frac{\partial\theta_{t+1|t}}{\partial\phi}\right) = E_{t-1}\Bigg[x_t^2\frac{\partial\theta_{t|t-1}}{\partial\kappa}\frac{\partial\theta_{t|t-1}}{\partial\phi} + u_t x_t\frac{\partial\theta_{t|t-1}}{\partial\phi} \\ + x_t(\theta_{t|t-1} - \omega)\frac{\partial\theta_{t|t-1}}{\partial\kappa} + u_t(\theta_{t|t-1} - \omega)\Bigg] \end{aligned}$$

$$E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}.\frac{\partial\theta_{t+1|t}}{\partial\phi}\right) = b\left(\frac{\partial\theta_{t|t-1}}{\partial\kappa}\frac{\partial\theta_{t|t-1}}{\partial\phi}\right) + c\frac{\partial\theta_{t|t-1}}{\partial\phi} + a(\theta_{t|t-1} - \omega)\frac{\partial\theta_{t|t-1}}{\partial\kappa}. \tag{1.26}$$

But,

$$E_{t-2}\left((\theta_{t|t-1} - \omega)\frac{\partial\theta_{t|t-1}}{\partial\kappa}\right) = E_{t-2}\left[\left(x_{t-1}\frac{\partial\theta_{t-1|t-2}}{\partial\kappa} + u_{t-1}\right)\left(\phi(\theta_{t-1|t-2} - \omega) + \kappa u_{t-1}\right)\right]$$

$$\begin{aligned} E_{t-2}\left((\theta_{t|t-1} - \omega)\frac{\partial\theta_{t|t-1}}{\partial\kappa}\right) = E_{t-2}\Bigg[x_{t-1}\frac{\partial\theta_{t-1|t-2}}{\partial\kappa}\phi(\theta_{t-1|t-2} - \omega) \\ + x_{t-1}u_{t-1}\kappa\frac{\partial\theta_{t-1|t-2}}{\partial\kappa} + u_{t-1}\phi(\theta_{t-1|t-2} - \omega) + \kappa u_{t-1}^2\Bigg] \end{aligned}$$

$$\begin{aligned} \lim_{n\to\infty} E_{t-n}\left((\theta_{t|t-1} - \omega)\frac{\partial\theta_{t|t-1}}{\partial\kappa}\right) &= a\phi \lim_{n\to\infty}\left(E_{t-n}(\theta_{t-1|t-2} - \omega)\frac{\partial\theta_{t-1|t-2}}{\partial\kappa}\right) + \kappa\sigma_u^a \\ \implies \qquad E\left((\theta_{t|t-1} - \omega)\frac{\partial\theta_{t|t-1}}{\partial\kappa}\right) &= \frac{\kappa\sigma_u^2}{1 - a\phi}. \end{aligned}$$

Similarly from (1.26),

$$E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}.\frac{\partial\theta_{t+1|t}}{\partial\phi}\right) = \frac{1}{1-b}\left(\frac{\kappa\sigma_u^2}{1-a\phi}\right),$$

$$E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\omega}.\frac{\partial\theta_{t+1|t}}{\partial\phi}\right) = E_{t-1}\left[\left(x_t\frac{\partial\theta_{t|t-1}}{\partial\omega}+1-\phi\right)\left(x_t\frac{\partial\theta_{t|t-1}}{\partial\phi}+\theta_{t|t-1}-\omega\right)\right],$$

$$\implies E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\omega}.\frac{\partial\theta_{t+1|t}}{\partial\phi}\right) = E_{t-1}\left[x_t^2\frac{\partial\theta_{t|t-1}}{\partial\omega}\frac{\partial\theta_{t|t-1}}{\partial\phi}+(1-\phi)x_t\frac{\partial\theta_{t|t-1}}{\partial\phi}\right.$$
$$\left. + x_t(\theta_{t|t-1}-\omega)\frac{\partial\theta_{t|t-1}}{\partial\omega}+(1-\phi)(\theta_{t|t-1}-\omega)\right],$$

$$E_{t-1}\left(\frac{\partial\theta_{t+1|t}}{\partial\kappa}.\frac{\partial\theta_{t+1|t}}{\partial\phi}\right) = b\left(\frac{\partial\theta_{t|t-1}}{\partial\omega}\frac{\partial\theta_{t|t-1}}{\partial\phi}\right)+a(1-\phi)\frac{\partial\theta_{t|t-1}}{\partial\phi}$$
$$+ a(\theta_{t|t-1}-\omega)\frac{\partial\theta_{t|t-1}}{\partial\omega}. \quad (1.27)$$

Similarly as in above we can write,

$$E\left((\theta_{t|t-1}-\omega)\frac{\partial\theta_{t|t-1}}{\partial\kappa}\right) = \frac{\kappa c(1-\phi)}{(1-a\phi)(1-a)}, \quad (1.28)$$

$$\implies E\left(\frac{\partial\theta_{t+1|t}}{\partial\omega}.\frac{\partial\theta_{t+1|t}}{\partial\phi}\right) = \frac{1}{1-b}\left(\frac{a(1-\phi)\kappa c}{(1-a\phi)(1-a)}\right).$$

□

1.1.8 Asymptotic Distribution

Given that $\hat{\psi}$ is the maximum likelihood estimate (MLE) of the fixed parameters governing the model of the time varying parameter considered above, the maximum likelihood estimator is the global maximum of the likelihood function and $I^{-1}(\boldsymbol{\psi_0})$ is positive definite, by the central limit theorem this is consistent and the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})$ as $n \to \infty$ is a multivariate normal with mean vector zero and covariance matrix

$$Var(\hat{\boldsymbol{\psi}}) = I^{-1}(\boldsymbol{\psi_0}).$$

This implies that the asymptotic covariance matrix of $\hat{\boldsymbol{\psi}}$ is of the form $(1/n) \times Var(\hat{\boldsymbol{\psi}})$ where the asymptotic standard error of an estimate is the square root of the corresponding diagonal element of $(1/n) \times Var(\hat{\boldsymbol{\psi}})$.

1.1.9 Dynamic Student's t Location Model

Here we consider the dynamic linear model, specifically the first order-case of the model which is of the form,

$$y_t = \mu_{t|t-1} + v_t = \mu_{t|t-1} + exp(\lambda)\varepsilon_t, \quad t = 1, \cdots, n. \tag{1.29}$$

$$\mu_{t+1|t} = \delta + \phi\mu_{t|t-1} + \kappa u_t.$$

The prediction errors, v_t in (1.29) by construction are independently and identically distributed as t variates with mean zero and scale $exp(\lambda)$. The error term ε_t is defined as a serially independent standard t-variate. When the location is being considered as a time varying parameter, it may be captured by a model in which the conditional distribution of the observed series has a t_ν distribution with conditional median $\mu_{t|t-1}$ (in this case since the t_ν is a symmetric distribution, the mean and median are same). Here the log-likelihood, is given by

$$\log f(y_t; \mu_{t|t-1}, \varphi, \nu) = \log \Gamma((\nu+1)/2) - \frac{1}{2}\log \pi - \log \Gamma(\nu/2)$$
$$- \frac{1}{2}\log \nu - \log \varphi - \frac{\nu+1}{2}\log\left(1 + \frac{(y_t - \mu_{t|t-1})^2}{\nu\varphi^2}\right) \quad (1.30)$$

$$\frac{\partial \log f_t}{\partial \mu_{t|t-1}} = (\nu+1)(\nu\varphi^2)^{-1}\left(1 + \frac{(y_t - \mu_{t|t-1})^2}{\nu e^{2\lambda}}\right)^{-1}(y_t - \mu_{t|t-1})$$

$\implies k = \frac{\nu}{\nu+1}\varphi^2$. The scaled score function that drives the dynamics of the model has the form

$$u_t = \left(1 + \frac{(y_t - \mu_{t|t-1})^2}{\nu e^{2\lambda}}\right)^{-1} v_t, \quad t = 1, 2, \cdots n, \quad (1.31)$$

where $v_t = y_t - \mu_{t|t-1}$ is the prediction error and $\varphi = exp(\lambda)$ is the (time-invariant) scale. The model requires that the degree of freedom ν be positive. Similarly the u_t's are also iid as it is essentially a function of v_t which is in itself random. Since the mechanism of updating the parameters overtime depends on u_t, it is worth elaborating on the properties of u_t that follow from its relationship with the beta distribution.

Proposition 6. $[Proposition\ 7\ of\ Harvey$ (2013)$]$ *The variable u_t can be written*

$$u_t = (1 - b_t)(y_t - \mu_{t|t-1}), \quad (1.32)$$

where $b_t = \frac{(y_t - \mu_{t|t-1})^2/ve^{2\lambda}}{1+(y_t - \mu_{t|t-1})^2/ve^{2\lambda}}$, $\quad 0 \leq b_t \leq 1, \quad 0 < v < \infty$ *is distributed as a beta with shape parameters* $1/2$ *and* $\nu/2$*, we donote this as* $B(1/2, \nu/2)$*. The* u_t*'s are* $iid(0, \sigma_u^2)$

and symmetrically distributed. Even moments of all orders exist and are given by

$$E(u_t^m) = \nu^{m/2} e^{m\lambda} \frac{B((1+m)/2), (\nu+m)/2))}{B(1/2, \nu/2)}, \quad m = 2, 4, \cdots . \tag{1.33}$$

Proof: Since $(y_t - \mu_{t|t-1})/exp(\lambda) = \varepsilon_t$ has a standardized distribution, the distribution of u_t does not change with time and does not depend on $\mu_{t|t-1}$. Because the distribution of $y_t - \mu_{t|t-1}$ is symmetric about zero, the same applies to the distribution of u_t as b_t does not depend on the sign of $y_t - \mu_{t|t-1}$.

The fact that b_t follows a beta distribution is shown below, the term in the bracket in (1.31) is equal to $(1 - b_t)$. Now,

$$b_t = \frac{(y_t - \mu_{t|t-1})^2/\nu e^{2\lambda}}{1 + (y_t - \mu_{t|t-1})^2/\nu e^{2\lambda}}, \quad (1 - b_t) = \frac{1}{\left(1 + (y_t - \mu_{t|t-1})^2/\nu e^{2\lambda}\right)}.$$

$$u_t = \nu^{1/2} e^{\lambda}(1 - b_t)(y_t - \mu_{t|t-1})\nu^{-1/2} e^{-\lambda} = \nu^{1/2} e^{\lambda}(1 - b_t)\varepsilon_t/\nu^{1/2}.$$

$$u_t^2 = \nu e^{2\lambda}(1 - b_t)^2 \varepsilon_t^2/\nu.$$

Since $\varepsilon_t \sim t_\nu \implies \varepsilon_t^2 = \chi_1^2/(\chi_\nu^2/\nu) = \nu.\chi_1^2/\chi_\nu^2$, therefore $\varepsilon_t^2/\nu = \chi_1^2/\chi_\nu^2$

$$\begin{aligned}
u_t^2 &= \nu e^{2\lambda}(1 - b_t)^2 \chi_1^2/\chi_\nu^2 \\
&= \nu e^{2\lambda}\left(\frac{1}{1 + (\chi_1^2/\chi_\nu^2)}\right)^2 \times (\chi_1^2/\chi_\nu^2) = \nu e^{2\lambda}\left(\frac{\chi_\nu^2}{\chi_1^2 + \chi_\nu^2}\right)^2 \times (\chi_1^2/\chi_\nu^2) \\
&= \nu e^{2\lambda}\left(\frac{\chi_\nu^2}{\chi_1^2 + \chi_\nu^2}\right)\left(\frac{\chi_1^2}{\chi_1^2 + \chi_\nu^2}\right) \\
&= \nu e^{2\lambda} b_t(1 - b_t) \\
u_t^2 &= \nu^{2/2} e^{2\lambda} b_t^{2/2}(1 - b_t)^{2/2} \\
u_t^m &= \nu^{m/2} e^{m\lambda} b_t^{m/2}(1 - b_t)^{m/2}, \quad m = 2, 4, \cdots .
\end{aligned}$$

Hence from

$$E(b^h(1-b)^k) = \frac{B(a+h, b+k)}{B(a,b)}, \;\; h > -a, \;\; k > -b. \tag{1.34}$$

$E(u_t^m) = \nu^{m/2} e^{m\lambda} \frac{B((1+m)/2),(\nu+m)/2))}{B(1/2,\nu/2)}, \;\; m = 2, 4, \cdots.$

Corollary 6.1. [*Corollary* 10 *of* Harvey (2013)] *The variance of* u_t *is*

$$Var(u_t) = \sigma_u^2 = ve^{2\lambda} E(b_t(1-b_t)) = \frac{\nu^2 e^{2\lambda}}{(\nu+1)(\nu+3)}. \tag{1.35}$$

and its fourth moment

$$E(u_t^4) = \nu^2 e^{4\lambda} E(b_t^2(1-b_t)^2) = \frac{3v^3(\nu+2)e^{4\lambda}}{(\nu+1)(\nu+3)(\nu+5)(\nu+7)}. \tag{1.36}$$

Hence the kurtosis of u_t *is*

$$kurtosis(u_t) = \frac{2(\nu+2)(\nu+3)(\nu+1)}{\nu(\nu+5)(\nu+7)}.$$

Proof: Since $E(u_t) = 0 \implies Var(u_t) = E(u_t^2)$, therefore

$$E(u_t^2) = ve^{2\lambda} E(b_t(1-b_t)) \tag{1.37}$$

$$= \frac{\nu^2 e^{2\lambda}}{(\nu+1)(\nu+3)}.$$

From (1.34), $E(u_t^4)$ follows similarly. However, the kurtosis is,

$$kurtosis(u_t) = \frac{E(u_t^4)}{(E(u_t^2))^2} = \frac{2(\nu+2)(\nu+3)(\nu+1)}{\nu(\nu+5)(\nu+7)}.$$

The prediction error is $v_t = y_t - \mu_{t|t-1} = \epsilon_t e^{\lambda}$. The variance of the prediction error can be written as

$$Var(v_t) = Var(e^{\lambda}\varepsilon_t) = \nu/(\nu-2)e^{2\lambda}, \;\; \nu > 2.$$

Since $\varepsilon_t \sim t_v \implies \varepsilon_t^2 = \chi_1^2/(\chi_\nu^2/\nu) \implies \varepsilon^2/\nu = \chi_1^2/\chi_\nu^2$.

Also $(1-b_t) = \left(1 + \frac{(y_t-\mu_{t|t-1})^2}{\nu e^{2\lambda}}\right)^{-1} \implies (1-b_t) = \left(1+\frac{\varepsilon_t^2}{\nu}\right)^{-1} \implies (1-b_t) = \left(\frac{\chi_\nu^2}{\chi_1^2+\chi_\nu^2}\right)$. Using similar arguments we can also establish that $b_t = \left(\frac{\chi_1^2}{\chi_1^2+\chi_\nu^2}\right)$

It follows therefore that, $u_t v_t = (1-b_t)(y_t - \mu_{t|t-1})^2 \implies u_t v_t = \nu e^{2\lambda}(1-b_t)\varepsilon_t^2/v \implies u_t v_t = \nu e^{2\lambda}\left(\frac{\chi_\nu^2}{\chi_1^2+\chi_\nu^2}\right)\frac{\chi_1^2}{\chi_\nu^2} = \nu e^{2\lambda} b_t$ and hence from (1.34) we obtain

$$E(u_t v_t) = E\left(v e^{2\lambda} b_t\right) = (\nu/(\nu+1))e^{2\lambda}.$$

Chapter 2

Poisson Autoregression

2.1 Introduction

Several authors have considered models for count time series eg., see Kedem and Fokianos (2005). Very popular among these models that are discussed by the authors is the loglinear model. If we consider Y_t is conditionally Poisson distributed with mean λ_t, then for most of the existing models $\log \lambda_t$ is regressed on past values of the response and/or covariates. The loglinear model guarantees the positivity of the intensity parameter λ_t, which is a necessary condition to be satisfied in the case of the Poisson distribution. These models fall within the broad class of generalized linear time series models and their analysis is based on partial likelihood inference. The estimation, diagnostics, assessment, and forecasting based on these models are implemented in a straightforward manner with the computation carried out in various existing statistical computing environments. Although impressive gains have been made in this field of time series, an element that is largely missing in these developments has been the possibility of an autoregressive feedback mechanism in $\{\lambda_t\}$. A feedback of this nature is a key feature in state-space models such as the GARCH model for volatility. We expect these models to be parsimonious generally. In this chapter, we study autoregeressive models of λ_t, both linear and nonlinear. More specifically we regress λ_t on past values of the observed process and past values of itself. Processes like this have been considered by Rydberg and Shephard (2000) and Streett (2000). In this chapter, we summarize two classes of models. The first class is the simple linear model that postulates that the conditional mean of the Poisson observed time series is a linear function of its past values and lagged values of observed process. The second class of models generalize the linear model by imposing a nonlinear stucture on both past values of λ_t and lagged values of Y_t. The main focus of this chapter is to understand the likelihood based inference procedure of Fokianos et al. (2009) and how this procedure can be extended to the zero inflated autoregressive conditional Poisson (ZIACP) model. The chapter begins with a brief introduction to Poisson autoregression; writing its likelihood, score function and information matrix. We then conduct a simulation study to evaluate the finite sample performance of the MLE estimates of the parameters in the linear model. In the linear case of the Poisson autoregression, the conditional mean is modeled linearly as a function of its past values and past values of the observed Poisson process.

2.1.1 Linear Model

Consider the linear model

$$Y_t|\mathcal{F}_{t-1}^{Y,\lambda} \sim Poisson(\lambda_t), \quad \lambda_t = \gamma + \alpha\lambda_{t-1} + \beta Y_{t-1}, \tag{2.1}$$

for $t \geq 1$ and where the parameters γ, α, β are assumed to be positive. This model may be viewed as a special case of ACP model, specifically the ACP(1,1) model. It is tempting for this model to be viewed as an integer valued GARCH model. This is because for the Poisson distribution, the conditional variance equals the conditional mean, that is, $E[Y_t|\mathcal{F}_{t-1}^{Y,\lambda}] = Var[Y_t|\mathcal{F}_{t-1}^{Y,\lambda}] = \lambda_t$. However, the proposed modeling is based on the evolution of the mean of the Poisson, not on its variance see Fokianos et al. (2009). Even though the vector of time-dependent covariates that influences the evolution of (2.1) contains the unobserved process λ_t, the linear model still belongs to the class of observation driven models defined by Cox et al. (1981). This is true because the unobserved process λ_t can be expressed as a function of past values of the observed process Y_t, after repeated substitution. In particular, if $Y_t|\mathcal{F}_{t-1}^{Y,\lambda} \sim Poisson(\lambda_t)$ then by iterated expectation $E(Y_t) = E(E(Y_t|\mathcal{F}_{t-1}^{Y,\lambda})) = E(\lambda_t)$ hence from (2.1) it follows that

$$E(Y_t) = E(\lambda_t) \equiv \mu = \frac{\gamma}{1-\alpha-\beta}.$$

Here we write (2.1) in the form of an ARMA(1,1) by defining the martingale difference $u_t = Y_t - E(Y_t|\mathcal{F}_{t-1}) = Y_t - \lambda_t$, then from (2.1), $\lambda_t = \gamma + \alpha\lambda_{t-1} + \beta Y_{t-1} \implies Y_t - (\alpha+\beta)Y_{t-1} = \gamma + u_t - \alpha u_{t-1}$. The ARMA(1,1) has an MA representation with ψ weights as $\psi_j = (\alpha+\beta)^{j-1}\beta$ for $j \geq 1$ and $\psi_0 = 1$. Hence the autocovariance in terms of ψ weights has the form

$$\begin{aligned} Cov(Y_t, Y_{t+k}) &= \mu\{\psi_0\psi_k + \psi_1\psi_{k+1} + \psi_2\psi_{k+2} + \cdots\} \\ &= \mu\{\beta(\alpha+\beta)^{k-1} + \beta^2(\alpha+\beta)^k + \beta^2(\alpha+\beta)(\alpha+\beta)^{k+1} + \cdots\} \\ &= \begin{cases} \frac{(1-(\alpha+\beta)^2+\beta^2)\mu}{1-(\alpha+\beta)^2}, \ k=0 \\ \frac{\beta(1-\alpha(\alpha+\beta))(\alpha+\beta)^{k-1}\mu}{1-(\alpha+\beta)^2}, \ k \geq 1 \end{cases} \end{aligned}$$

The variance is given by the expression below,

$$Var(Y_t) = \mu\left(1 + \frac{\beta^2}{1-(\alpha+\beta)^2}\right),$$

this implies that $Var(Y_t) \geq E(Y_t)$ (overdispersion) with equality when $\beta = 0$. This can be explained in the following way: including the past values of Y_t in the evolution of λ_t leads to overdispersion, which is a frequent phenomenon in count time series data.

2.1.2 Conditional Least Squares Estimate (CLSE) for the Linear Model

By defining the martingale difference $u_t = Y_t - E(Y_t|\mathcal{F}_{t-1}) = Y_t - \lambda_t$, we can write (2.1) in the form of an ARMA(1,1) model as $Y_t - (\alpha + \beta)Y_{t-1} = \gamma + u_t - \alpha u_{t-1}$. However, the ARMA(1,1) model in **R** by default uses a different parameterization where the MA parameter is positive as in $Y_t - \phi Y_{t-1} = \gamma + a_t + \theta a_{t-1}$. Therefore comparing this fit to the theoretical ARMA(1,1) model, we obtain $\hat{\alpha} = -\hat{\theta}$ and $\hat{\beta} = \hat{\phi} + \hat{\theta}$. Also since $\alpha + \beta < 1$, the Y_t process is stationary. Hence the mean of Y_t can be written as $\mu = \frac{\gamma}{1-\phi} \implies \hat{\gamma} = \hat{\mu}(1 - \hat{\phi})$. Thus the CLSE is $\hat{\gamma}_{CLSE} = \hat{\mu}(1 - \hat{\phi}),\ \hat{\alpha}_{CLSE} = \hat{\phi} + \hat{\theta}$, $\hat{\beta}_{CLSE} = -\hat{\theta}$ for $\gamma,\ \alpha$ and β, respectively.

2.1.3 Nonlinear Model

Consider the nonlinear model from Fokianos et al. (2009) below,

$$Y_t|\mathcal{F}_{t-1}^{Y,\lambda} \sim Poisson(\lambda_t), \quad \lambda_t = f(\lambda_{t-1}) + b(Y_{t-1}) \tag{2.2}$$

for $t \geq 1$, where $f(.)$ and $b(.)$ are known functions up to an unknown finite-dimensional parameter vector and $f, b : R^+ \to R^+$. The initial values Y_0 and λ_0 are fixed. Equation (2.2) represents a general defination of which equation (2.1) forms a special case. We can obtain (2.1) from (2.2) by defining $f(x) = \gamma + \alpha x$ and $b(x) = bx$ with $\gamma, a, b > 0$ and $x \geq 0$. We consider the nonlinear model,

$$Y_t|\mathcal{F}_{t-1}^{Y,\lambda} \sim Poisson(\lambda_t), \quad \lambda_t = (a + c\exp(-\gamma\lambda_{t-1}^2))\lambda_{t-1} + bY_{t-1}, \tag{2.3}$$

which parallels the structure of the traditional exponential AR model (see Haggan and Ozaki, 1981).

2.1.4 Likelihood Inference

Let $\boldsymbol{\theta} = (\gamma, \alpha, \beta)'$ be a three dimensional vector of unknown parameters and $\boldsymbol{\theta}_0 = (\gamma_0, \alpha_0, \beta_0)$ be the true value of the parameters. Then, the conditional likelihood function for $\boldsymbol{\theta}$ based on (2.1), given the starting value λ_0 in terms of the observations $Y_1, \cdots, Y_n$ is given by

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{t=1}^{n} \frac{\exp(-\lambda_t(\boldsymbol{\theta}))\lambda_t^{Y_t}(\boldsymbol{\theta})}{Y_t!}$$

using $\lambda_t(\boldsymbol{\theta}) = \gamma + \alpha\lambda_{t-1}(\boldsymbol{\theta}) + \beta Y_{t-1}$ and $\lambda_t = \lambda_t(\boldsymbol{\theta}_0)$. Therefore the log likelihood function is

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(\boldsymbol{\theta}) = \sum_{t=1}^{n} (Y_t \log \lambda_t(\boldsymbol{\theta}) - \lambda_t(\boldsymbol{\theta})). \tag{2.4}$$

and the score function is defined by

$$S_n(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\boldsymbol{\theta})} - 1 \right) \frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{2.5}$$

where $\partial\lambda_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is a three dimensional vector with components given by

$$\tfrac{\partial \lambda_t}{\partial \gamma} = 1 + \alpha \tfrac{\partial \lambda_{t-1}}{\partial \gamma}, \quad \tfrac{\partial \lambda_t}{\partial \alpha} = \lambda_{t-1} + \alpha \tfrac{\partial \lambda_{t-1}}{\partial \alpha}, \quad \tfrac{\partial \lambda_t}{\partial \beta} = Y_{t-1} + \alpha \tfrac{\partial \lambda_{t-1}}{\partial \beta}.$$

The solution of $S_n(\boldsymbol{\theta}) = 0$ if it exists yields the conditional MLE of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$. However, since the score equation cannot be solved explicitly for the parameters of interest, we resort to a direct numerical optimization using the **optim** function in **R** to solve for the estimates of the parameters. Also the Hessian matrix is obtained by further differentiation of the score function (2.5),

$$H_n(\boldsymbol{\theta}) = -\sum_{t=1}^{n} \frac{\partial^2 \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

$$= \sum_{t=1}^{n} \frac{Y_t}{\lambda_t^2(\boldsymbol{\theta})} \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)' - \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\boldsymbol{\theta})} - 1 \right) \frac{\partial^2 \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}. \qquad (2.6)$$

Theorem 7. [*Theorem* 3.1 *of* Fokianoset al. (2009)] *Consider model (2.1) and assume that at the true value $\theta_0, \alpha_0 + \beta_0 < 1$. Then, there exists a fixed open neighborhood $\mathcal{O} = \mathcal{O}(\boldsymbol{\theta}_0)$, with probability tending to 1, as $n \to \infty$, where the log-likelihood function (2.4) has a unique maximum point $\hat{\theta}$. $\hat{\theta}$ is consistent and asymptotically normal,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}(0, \boldsymbol{G}^{-1}),$$

where the matrix $\boldsymbol{G}$ is defined as $G(\boldsymbol{\theta}) = E(\frac{1}{\lambda_t}(\frac{\partial \lambda_t}{\partial \boldsymbol{\theta}})(\frac{\partial \lambda_t}{\partial \boldsymbol{\theta}})')$. A consistent estimator of G is given by $G_n(\hat{\boldsymbol{\theta}})$ where

$$G_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} Var\left[\frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} | \mathcal{F}_{t-1} \right] = \sum_{t=1}^{n} \frac{1}{\lambda_t(\boldsymbol{\theta})} \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)'.$$

For the nonlinear model (2.3) with $\boldsymbol{\theta} = (\alpha, \psi, \beta, \gamma)$, the recursions for calculating the score are given by

$$\frac{\partial \lambda_t}{\partial a} = \left(1 - 2\gamma c \lambda_{t-1} \exp(-\gamma \lambda_{t-1}^2) \frac{\partial \lambda_{t-1}}{\partial a} \right) \lambda_{t-1} + \left(a + c \exp(-\gamma \lambda_{t-1}^2) \right) \frac{\partial \lambda_{t-1}}{\partial a},$$

$$\frac{\partial\lambda_t}{\partial c} = \left(1 - 2\gamma c\lambda_{t-1}\frac{\partial\lambda_{t-1}}{\partial c}\right)\exp(-\gamma\lambda_{t-1}^2)\lambda_{t-1} + \left(\alpha + c\exp(-\gamma\lambda_{t-1}^2)\right)\frac{\partial\lambda_{t-1}}{\partial c},$$

$$\frac{\partial\lambda_t}{\partial b} = a\frac{\partial\lambda_{t-1}}{\partial b} + \left(1 - 2\gamma\lambda_{t-1}^2\right)c\exp(-\gamma\lambda_{t-1}^2)\frac{\partial\lambda_{t-1}}{\partial b} + Y_{t-1},$$

$$\frac{\partial\lambda_t}{\partial \gamma} = -c\exp(-\gamma\lambda_{t-1}^2)\lambda_{t-1}^2\left(\lambda_{t-1} + 2\gamma\frac{\partial\lambda_{t-1}}{\partial\gamma}\right) + (a + c\exp(-\gamma\lambda_{t-1}^2))\frac{\partial\lambda_{t-1}}{\partial\gamma}.$$

2.1.5 Simulation for the linear model

We conduct a simulation study to illustrate the performance of the MLE with respect to the conditional least squares estimates (CLSE). Data is generated from the following model

$$Y_t|\mathcal{F}_{t-1}^{Y,\lambda} \sim Poisson(\lambda_t), \quad \lambda_t = \gamma + \alpha\lambda_{t-1} + \beta Y_{t-1}.$$

We consider the true parameters $(\gamma, \alpha, \beta) = (0.3, 0.4, 0.5)$ with different sample sizes of $n = 200$, 500, and 1000. We chose the model parameters α and β such that $\alpha + \beta < 1$. The CLSE is obtained by fitting the ARMA(1,1) model using the simulated data. Next we use this CLSE estimates as an initial value to calculate the MLE using the same simulated data. In order to obtain the MLE estimates, CLSE is used as an initial value for the **optim()** function in **R** package. This **optim()** function requires the score function, the information matrix, and the likelihood function as objects. It may be sensitive to initial values thus the initial values for the parameters γ, α and β are chosen to be the CLSE. We evaluate the performance of both the MLE and the CLSE by using the MSE criteria. The above process is iterated 1000 times and in every iteration the CLSE and MLE are stored to calculate the average estimates. The performance of MLE with respect to CLSE is evaluated using the relative mean squared error (RMSE) criteria. The RMSE is defined as

$$RMSE(\hat{\boldsymbol{\theta}}_{CLSE}, \hat{\boldsymbol{\theta}}_{MLE}) = \frac{MSE(\hat{\boldsymbol{\theta}}_{CLSE})}{MSE(\hat{\boldsymbol{\theta}}_{MLE})},$$

where

$$MSE(\hat{\boldsymbol{\theta}}) = VAR(\hat{\boldsymbol{\theta}}) + (BIAS(\hat{\boldsymbol{\theta}}))^2.$$

Table 2.1 reports the CLSE, MLE, RMSE, skewness, kurtosis and $p-$value of the Kolmogorov-Smirnov test. This table suggests that when the sample size is small ($n = 200$) the estimates are not very close to the true parameter values but when the sample size is increased to 500 and 1000 the estimates become very close to the true parameters. In all cases the MSE of the MLE is lower than that of the CLSE. That is, MLE outperforms the CLSE. The skewness for a normal distribution is zero and since we expect the sampling distribution of the estimates to be approximately normal, as the sample size increases the skewness approaches zero (sixth column). Similarly in column seven, as the sample size increases, the kurtosis approaches three, which is the kurtosis for the normal distribution. The last column of the table reports p-values which is based on the Kolmogorov- Smirnov test that compares the normality of the estimates based on 1000 simulations to a reference distribution. In this case the reference distribution is the normal distribution. The null hypothesis for the test is that the estimates are normally distributed.

Table 2.1: Simulation results for model (2.1) when (γ,α_0,β_0)=(0.3,0.4,0.5). Parameters are adapted as in Fokianos et al. (2009)

Parameters	Sample size	MLE	CLSE	RMSE	Skewness	Kurtosis	$p-$value
γ	200	0.3729	0.3870	1.3510	0.9771	4.2308	0.0793
α		0.3716	0.3769	1.2086	−0.4746	4.3317	0.5616
β		0.4984	0.4869	1.2281	−0.0120	3.2101	0.9815
γ	500	0.3309	0.3359	1.4710	0.6701	3.9512	0.0521
α		0.3878	0.3910	1.3599	−0.0028	3.3476	0.8688
β		0.4996	0.4953	1.3955	−0.0211	3.0733	0.9623
γ	1000	0.3148	0.3166	1.5519	0.4662	3.3707	0.0468
α		0.3955	0.3957	1.2582	−0.1251	3.0360	0.8809
β		0.4987	0.4975	1.4005	−0.0738	2.8402	0.3935

In Table 2.1, the third and fourth columns report the means of the MLE and CLSE. The fifth column reports the ratio of the MSE of CLSE to the MSE of the MLE. The other three columns report sample skewness, sample kurtosis, and p-values of the Kolmogorov-Smirnov test for normality.

2.1.6 Simulation for the nonlinear model

In this subsection, we report the results of a simulation study for the nonlinear model. Data is generated from the model

$$Y_t|\mathcal{F}_{t-1}^{Y,\lambda} \sim Poisson(\lambda_t), \quad \lambda_t = (a + c\exp(-\gamma\lambda_{t-1}^2))\lambda_{t-1} + bY_{t-1},$$

with the true parameters $(a, c, b, \gamma) = (0.25, 1.0, 0.65, \gamma)$ where $\gamma = 0.5, 1.0, 1.5$. In estimating the parameter vector (a, c, b, γ) we proceed as follows. We first fit a linear model to the simulated data to obtain starting values for both a and b, we then set the initial value of c to some constant. We generated a grid of values for γ and for each of these grid values we fit the nonlinear model with known γ. Finally, to maximize the log-likelihood function over all (a, c, b, γ), we use as a starting value the value of γ that yields the maximum likelihood from the previous step together with corresponding coefficients. The above process is iterated 500 times and in every iteration the estimates of a, c, b and γ are stored to calculate the average estimates. The MSE of the estimates is also calculated based on the 500 estimates by computing the variance of the 500 estimates and also the bias of each estimate with respect to its true parameter value. The MSE for each estimate is stored for each iteration to calculate the average MSE estimates.

Table 2.2: Simulation results for the nonlinear model when sample size $n = 500$ and where (a,c,b,γ)=(0.25,1,0.65,γ) as in Fokianos et al. (2009) with $\gamma = 0.5, 1.0, 1.5$

$\hat{a}$	$\hat{c}$	$\hat{b}$	$\hat{\gamma}$	True γ
$0.2326(0.0398)$	$1.1180(0.3088)$	$0.6656(0.0044)$	$0.5407(0.1174)$	0.5
$0.2606(0.0128)$	$1.0333(0.2545)$	$0.6604(0.0048)$	$1.0065(0.5543)$	1.0
$0.2513(0.0102)$	$1.0447(0.2676)$	$0.6602(0.0062)$	$1.5012(1.7605)$	1.5

In Table 2.2, we only considered 500 iterations since it takes a long time for the program to run and we found 500 iterations to give appropriate estimates.

2.1.7 Zero Inflated Poisson (ZIP) distribution

Suppose we are interested in the distribution of the number of insects on a leaf of a tree. The number of insects on a suitable leaf can be modeled by the Poisson distribution, see Nanjundan and Naika (2013). If a leaf has insect on it then it is suitable for feeding and if a leaf has no insect on it, then it may be due to its unsuitability or by chance variation due to the Poisson distribution. The probability function of the number of insects y on any observed leaf is

$$p(y;\lambda,\omega) = \begin{cases} \omega + (1-\omega)e^{-\lambda},\ y = 0 \\ (1-\omega)\frac{e^{-\lambda}\lambda^{y}}{y!},\ y = 1, 2, 3, \cdots \end{cases} \tag{2.7}$$

with $\lambda > 0$, $0 \leq \omega < 1$, ω is the so called zero inflation parameter and when $\omega = 0$, $p(y, \lambda, 0)$ turns out to be the usual Poisson distribution. Thus the distribution of Y is a convex combination of the distribution degenerate at zero and a Poisson distribution with mean λ. In the probability mass function (2.7), the zero inflation parameter ω can take negative values, given that $\omega \geq \frac{-e^{\lambda}}{(1-e^{-\lambda})}$, see Kharrati-Kopaei and Faghih (2011). In this case, the frequency of zeroes is less than the one accounted for under the ordinary Poisson distribution. This situation is described as a zero-deflated Poisson distribution (ZDP). However, the zero deflated case rarely occurs in practice.

2.1.8 Maximum Likelihood Estimation for the ZIP Model

Let $\boldsymbol{Y} = (Y_1, Y_2, Y_3, \cdots, Y_n)$ be a random sample with the probability mass function specified in (2.7). Then the likelihood function is given by

$$\mathcal{L}(\lambda, \omega|\boldsymbol{y}) = \prod_{j=1}^{n}\{\omega + (1-\omega)e^{-\lambda}\}^{1-t(y_j)}\left\{(1-\omega)\frac{e^{-\lambda}\lambda^{y_j}}{y_j!}\right\}^{t(y_j)}, \tag{2.8}$$

where

$$t(y_j) = \begin{cases} 0, & \text{if } y_j = 0 \\ 1, & \text{if } y_j \geq 1 \end{cases}$$

It is obvious that the above likelihood does not yield closed form expressions for the MLE's of λ and ω. This suggests that the MLE's of λ and ω have to be computed using a numerical procedure. In this case, the EM algorithm is considered over the Newton Raphson method as the latter may fail due to boundary problem, see McLeish and Small (1988) and Sprott (1980). The EM algorithm is an iterative procedure to estimate the parameters of a model which does not admit a closed form solution of the parameters of interest. This algorithm finds solutions of the log-likelihood function corresponding to the local maxima. Given that $Y = y$, the EM algorithm for maximizing $\ell_c(\theta|Y)$ is given by the following iterative procedure. If $\theta^{(j)}$ is the estimate of the EM algorithm at the jth iteration, then at the $j+1$ iteration the estimate is updated. This numerical procedure requires the likelihood to be rewritten to accommodate missing data. We could introduce a Bernoulli random variable Z_t such that $Z_t = 0$ when $Y_t = 0$ is from a Poisson distribution and $Z_t = 1$ when $Y_t = 0$ is from the degenerate distribution. Because just the observed data has no information on where zeroes are coming from, it is regarded as incomplete as such when $(y_1, y_2, \cdots, y_n)$ is augmented with $(z_1, z_2, \cdots, z_n)$ then $((y_1, z_1), (y_2, z_2), \cdots, (y_n, z_n))$ becomes a complete data set whose likelihood is given by

$$\mathcal{L}_c(\lambda, \omega|\boldsymbol{y}, \boldsymbol{u}) = \prod_{j=1}^{n}\omega^{u_j}\{(1-\omega)\frac{e^{-\lambda}\lambda^{y_j}}{y_j!}\}^{1-u_j},$$

where

$$u_j = \begin{cases} Z_j, \text{ if } y_j = 0, \\ 0, \text{ if } y_j \geq 1, \end{cases}$$

and is computable only if $y_1, y_2, \cdots, y_n$ and $z_1, z_2, \cdots, z_n$ are available. Hence the log of the complete data likelihood becomes

$$\log \mathcal{L}_c(\lambda, \omega | \boldsymbol{y}, \boldsymbol{u}) = \sum_{j=1}^{n} u_j \log \omega + \sum_{j=1}^{n} (1 - u_j) \log\{(1-\omega)\frac{e^{-\lambda}\lambda^{y_j}}{y!}\}, \tag{2.9}$$

We could split equation (2.9) according to the defination of u_j by substituting Z_j in the place of u_j when the observed data is zero and zero in place of u_j when the observed data is a non-zero. Substituting u_j in equation (2.9), we obtain,

$$\log \mathcal{L}_c(\lambda, \omega | \boldsymbol{y}, \boldsymbol{u}) = \sum_{j:y_j>0} \{\log(1-\omega) + \log \frac{e^{-\lambda}\lambda^{y_j}}{y!}\} + \sum_{j:y_j=0} Z_j \log \omega$$

$$+ \sum_{j:y_j=0} (1 - Z_j) \log\{(1-\omega)\frac{e^{-\lambda}\lambda^{y_j}}{y!}\}, \tag{2.10}$$

In order to obtain the expected value of $\log \mathcal{L}_c(\lambda, \omega | \boldsymbol{y}, \boldsymbol{u})$ i.e., $E\{\log \mathcal{L}_c(\lambda, \omega | \boldsymbol{y}, \boldsymbol{u})\}$, we apply the expectation function E through (2.10). The result (2.11), is shown below,

$$E\{\log \mathcal{L}_c(\lambda, \omega | \boldsymbol{y}, \boldsymbol{u})\} = \sum_{j:y_j>0} \{\log(1-\omega) + \log \frac{e^{-\lambda}\lambda^{y_j}}{y!}\} + \sum_{j:y_j=0} E(Z_j) \log \omega$$

$$+ \sum_{j:y_j=0} (1 - E(Z_j)) \log\left[(1-\omega)\frac{e^{-\lambda}\lambda^{y_j}}{y!}\right]. \tag{2.11}$$

In the EM algorithm, $E(Z_j)$ is replaced by the conditional expectation

$E(Z_j | \lambda_0, \omega_0, Y_j = 0)$ where λ_0 and ω_0 are the initial estimates of λ and ω, respectively.

Thus

$$E(Z_j|\lambda_0,\omega_0,Y_j=0) = 0\times P(Z_j=0|\lambda_0,\omega_0,Y_j=0) + 1\times P(Z_j=1|\lambda_0,\omega_0,Y_j=0)$$

$$= P(Z_j=1|\lambda_0,\omega_0,Y_j=0).$$

From Baye's theorem,

$$P(Z_j=1|\lambda_0,\omega_0,Y_j=0) = \frac{P(Y_j=0|\lambda_0,\omega_0,Z_j=1)P(Z_j=1|\lambda_0,\omega_0)}{\sum\limits_{z_j=0,1} P(Y_j=0|\lambda_0,\omega_0,Z_j=z_j)P(Z_j=z_j|\lambda_0,\omega_0)},$$

and so,

$$E(Z_j|\lambda_0,\omega_0,Y_j=0) = \frac{\omega_0}{\omega_0+(1-\omega_0)e^{-\lambda_0}} = \psi,$$

which is a constant i.e., independent of j. Therefore, (2.11) becomes

$$E\{\log\mathcal{L}_c(\lambda,\omega|\underline{y},\underline{u})\} = \sum_{j:y_j>0}\left[\log(1-\omega)+\log\frac{e^{-\lambda}\lambda^{y_j}}{y!}\right] + \sum_{j:y_j=0}\psi\log\omega$$

$$+\sum_{j:y_j=0}(1-\psi)\log\left[(1-\omega)\frac{e^{-\lambda}\lambda^{y_j}}{y!}\right].$$

Assuming ψ is fixed and known, the next step of the algorithm requires maximizing $E\{\log\mathcal{L}_c(\lambda,\omega|\boldsymbol{y},\boldsymbol{u})\}$ for λ and ω by differentiating it with respect to λ and ω and equating both to zero and solving for λ_1 and ω_1, the improved estimates of λ_0 and ω_0, respectively.

$$\frac{\partial E\{\log\mathcal{L}_c(\lambda,\omega|\boldsymbol{y},\boldsymbol{u})\}}{\partial\lambda} = \sum_{j:y_j>0}\left[-1+\frac{Y_j}{\lambda}\right] + \sum_{j:y_j=0}(1-\psi)\left(-1+\frac{Y_j}{\lambda}\right) = 0$$

$\Longrightarrow$

$$\lambda_1 = \frac{\sum\limits_{j:y_j>0} Y_j}{(1-\psi)n_0+n_g}.$$

Using $n_g = n - n_o$ we can write λ_1 as

$$\lambda_1 = \frac{\sum\limits_{j:y_j>0}}{n - \psi n_o},$$

which estimates the number of Poisson observations. Similarly,

$$\frac{\partial E\{\log \mathcal{L}_c(\lambda, \omega|\boldsymbol{y}, \boldsymbol{u})\}}{\partial \omega} = \sum_{j:y_j>0} \frac{-1}{1-\psi} + \sum_{j:y_j=0} \{\frac{\psi}{\omega} - \frac{(1-\psi)}{(1-\omega)}\} = 0$$

$\implies \omega_1 = \frac{n_0 \psi}{n}$. This expresses the proportion of zeroes times proportion of structural zeroes among all zeroes. The n_0's denote the number of zero observations and n_g the number of non-zero observations. The expectation step is repeated by taking $\lambda_0 = \lambda_1$ and $\omega_0 = \omega_1$. After each iteration, the value of the log-likelihood (2.11) is evaluated and the difference between this maximum value and the preceding maximum value is taken. The convergence criterion is met when the absolute difference of the likelihoods (i.e, succeeding and preceding) is ≤ 0.00001.

2.1.9 The Fisher Information Matrix

Let $\boldsymbol{\theta}_0 = (\omega_0, \lambda_0)'$ be the true parameter values for the model and $\hat{\boldsymbol{\theta}}$ as the corresponding MLE estimates then, under certain regularity conditions as the sample size increases the MLE $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$, and it is asymptotically normal in the sense that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges to a normal distribution with a zero vector mean and covariance matrix $\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)$ i.e.,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim N(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_0)),$$

where $I(\boldsymbol{\theta}_0)$ constitutes the Fisher information matrix. Therefore the asymptotic variance $Var(\hat{\omega})$ and $Var(\hat{\lambda})$ of the estimates of ω and λ, respectively are given by

$$Var(\hat{\omega}) = \frac{(1-\omega)\Big(\omega + (1-\omega)e^{-\lambda} - \lambda\omega e^{-\lambda}\Big)\Big(\omega + (1-\omega)e^{-\lambda}\Big)}{(1-e^{-\lambda})\Big(\omega + (1-\omega)e^{-\lambda} - \omega\lambda e^{-\lambda}\Big) - \lambda e^{-2\lambda}}. \tag{2.12}$$

$$Var(\hat{\lambda}) = \frac{\lambda\Big(\omega + (1-\omega)e^{-\lambda}\Big)(1-e^{-\lambda})}{(1-\omega)\Big((1-e^{-\lambda})[\omega + (1-\omega)e^{-\lambda} - \omega\lambda e^{-\lambda}] - \lambda e^{-2\lambda}\Big)}. \tag{2.13}$$

Proof: From (2.7) taking the logarithm on both sides

$$\log p(y;\lambda,\omega) = \begin{cases} \log\{\omega + (1-\omega)e^{-\lambda}\}, \ y = 0 \\ \log(1-\omega) - \lambda + y\log\lambda - \log y! \ , \ y = 1,2,3,\cdots \end{cases} \tag{2.14}$$

$$\frac{\partial \log p(y;\lambda,\omega)}{\partial\lambda} = \begin{cases} \frac{-(1-\omega)e^{-\lambda}}{\omega+(1-\omega)e^{-\lambda}}, \ y = 0 \\ (-1 + \frac{y}{\lambda}) \ , \ y = 1,2,3,\cdots \end{cases}$$

$$\frac{\partial \log p(y;\lambda,\omega)}{\partial\omega} = \begin{cases} \frac{(1-e^{-\lambda})}{\omega+(1-\omega)e^{-\lambda}}, \ y = 0 \\ \frac{-1}{(1-\omega)} \ , \ y = 1,2,3,\cdots \end{cases}$$

$$\frac{\partial^2 \log p(y;\lambda,\omega)}{\partial\omega\partial\lambda} = \begin{cases} \frac{e^{-\lambda}}{\{\omega+(1-\omega)e^{-\lambda}\}^2}, \ y = 0 \\ 0 \ , \ y = 1,2,3,\cdots \end{cases}$$

$$\frac{\partial^2 \log p(y;\lambda,\omega)}{\partial\lambda^2} = \begin{cases} \frac{\omega(1-\omega)e^{-\lambda}}{\{\omega+(1-\omega)e^{-\lambda}\}^2}, \ y = 0 \\ \frac{-y}{\lambda^2} \ , \ y = 1,2,3,\cdots \end{cases}$$

$$\frac{\partial^2 \log p(y;\lambda,\omega)}{\partial\omega^2} = \begin{cases} \frac{-(1-e^{-\lambda})^2}{\{\omega+(1-\omega)e^{-\lambda}\}^2}\ ,\ y=0 \\ \frac{-1}{(1-\omega)^2}\ ,\ y=1,2,3,\cdots \end{cases}$$

Hence, we can get

$$E\biggl(\frac{\partial \log p(y;\lambda,\omega)}{\partial\lambda}\biggr) = \frac{-(1-\omega)e^{-\lambda}}{\omega+(1-\omega)e^{-\lambda}} \times p(0;\lambda,\omega)$$

$$+\sum_{y=1}^{\infty}(-1+\frac{y}{\lambda}) \times (1-\omega)p(y;\lambda,\omega)$$

$$= -(1-\omega)e^{-\lambda} + (1-\omega) - (1-\omega)(1-e^{-\lambda}) = 0.$$

By similar argument $E\biggl(\frac{\partial \log p(y;\lambda,\omega)}{\partial\omega}\biggr) = 0$. Also

$$I_{\omega\omega} = E\biggl[\biggl(\frac{\partial \log p(y;\lambda,\omega)}{\partial\omega}\biggr)\biggr]^2 = -E\biggl(\frac{\partial^2 \log p(y;\lambda,\omega)}{\partial\omega^2}\biggr)$$

$$= \frac{(1-e^{-\lambda})}{(1-\omega)\biggl(\omega+(1-\omega)e^{-\lambda}\biggr)}.$$

$$I_{\lambda\lambda} = E\biggl[\biggl(\frac{\partial \log p(y;\lambda,\omega)}{\partial\lambda}\biggr)\biggr]^2$$

$$= -E\biggl(\frac{\partial^2 \log p(y;\lambda,\omega)}{\partial\lambda^2}\biggr) = \frac{(1-\omega)\biggl(\omega+(1-\omega)e^{-\lambda}-\omega\lambda e^{-\lambda}\biggr)}{\lambda\biggl(\omega+(1-\omega)e^{-\lambda}\biggr)}.$$

$$I_{\lambda\omega} = E\left(\frac{\partial^2 \log p(y;\lambda,\omega)}{\partial\lambda\partial\omega}\right) = \frac{e^{-\lambda}}{\omega+(1-\omega)e^{-\lambda}}.$$

Therefore the Fisher information matrix is given by,

$$I = \begin{pmatrix} I_{\omega\omega} & I_{\omega\lambda} \\ I_{\lambda\omega} & I_{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} \frac{(1-e^{-\lambda})}{(1-\omega)\{\omega+(1-\omega)e^{-\lambda}\}} & \frac{e^{-\lambda}}{\omega+(1-\omega)e^{-\lambda}} \\ \frac{e^{-\lambda}}{\omega+(1-\omega)e^{-\lambda}} & \frac{(1-\omega)[\omega+(1-\omega)e^{-\lambda}-\omega\lambda e^{-\lambda}]}{\lambda\{\omega+(1-\omega)e^{-\lambda}\}} \end{pmatrix}.$$

The inverse of the Fisher information matrix is given by

$$\boldsymbol{I}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where

$$\Sigma_{11} = \frac{(1-\omega)\Big(\omega+(1-\omega)e^{-\lambda}-\lambda\omega e^{-\lambda}\Big)\Big(\omega+(1-\omega)e^{-\lambda}\Big)}{(1-e^{-\lambda})\Big(\omega+(1-\omega)e^{-\lambda}-\omega\lambda e^{-\lambda}\Big)-\lambda e^{-2\lambda}},$$

$$\Sigma_{12} = \frac{-\lambda e^{-\lambda}\Big(\omega+(1-\omega)e^{-\lambda}\Big)}{(1-e^{-\lambda})\Big(\omega+(1-\omega)e^{-\lambda}-\omega\lambda e^{-\lambda}\Big)-\lambda e^{-2\lambda}},$$

$$\Sigma_{22} = \frac{\lambda\Big(\omega+(1-\omega)e^{-\lambda}\Big)(1-e^{-\lambda})}{(1-\omega)\Big((1-e^{-\lambda})[\omega+(1-\omega)e^{-\lambda}-\omega\lambda e^{-\lambda}]-\lambda e^{-2\lambda}\Big)}.$$

2.1.10 Method of Moments Estimators for the ZIP Distribution

The first and second moments of Y having the probability mass function (2.7) are, respectively

$$E(Y) = (1-\omega)\lambda$$

and

$$E(Y^2) = (1-\omega)\lambda(1+\lambda),$$

where $\boldsymbol{Y} = (Y_1, Y_2, Y_3, \cdots, Y_n)$ is a random sample with the probability mass function specified in (2.7), the MME's of λ and ω are given by the following simultaneous equations

$$M_{1n} = (1-\omega)\lambda$$

and

$$M_{2n} = (1-\omega)\lambda(1+\lambda)$$

with $M_{1n} = \frac{1}{n}\sum_{j=1}^{n} Y_j$ and $M_{2n} = \frac{1}{n}\sum_{j=1}^{n} Y_j^2$. Solving the simultaneous equations, the MMEs of λ and ω are, respectively $\hat{\lambda}_m = \frac{M_{2n}}{M_{1n}} - 1$ and $\hat{\omega}_m = 1 - \frac{M_{1n}^2}{M_{2n}-M_{1n}}$. It is true that $P(M_{1n} = 0) = \{\omega + (1-\omega)e^{-\lambda}\}^n \to 0$, as $n \to \infty$. In other words, with probability tending to one as n becomes large M_{1n} is not equal to zero. Similarly, $P(M_{1n} = M_{2n}) \to 0$ as $n \to \infty$. Hence the problem of division by zero in these MMEs doesn't arise when n is sufficiently large.

It can be easily verified that if $\hat{\omega}_m > 0$, then $S^2 > \bar{X} = M_{1n}$ (i.e an overdispersion case) where $S^2 = M_{2n} - M_{1n}^2$ is the sample variance. Similarly if $\hat{\omega}_m < 0$, then $S^2 < \bar{X} = M_{1n}$ (i.e an underdispersion case) and if $\hat{\omega}_m = 0$ then $S^2 = \bar{X} = M_{1n}$ (i.e an equidispersion case).

We require $\hat{\omega}_m$ to test the Poisson against the ZIP distribution. $\hat{\omega}_m$ can be used to make inferences based on its asymptotic distribution see Kharrati-Kopaei and Faghih (2011).

Theorem 8. *Suppose that $Y_1, Y_2, Y_3, \cdots, Y_n$ are a random sample from the ZIP(ω, λ); then*

$$\sqrt{n}\left(\begin{pmatrix} \hat{\omega} \\ \hat{\lambda} \end{pmatrix} - \begin{pmatrix} \omega \\ \lambda \end{pmatrix}\right) \xrightarrow{\mathbf{D}} N(\mathbf{0}, \mathbf{\Sigma})$$

,

where $\xrightarrow{D}$ means convergence in distribution and

$$\mathbf{\Sigma} = \begin{pmatrix} (1-\omega)(\omega\lambda^2+2)/\lambda^2 & 2/\lambda \\ 2/\lambda & (\lambda+2)/(1-\omega) \end{pmatrix}.$$

Proof: Let $M = \sum_{i=1}^{n} A_i$ where $A_i = \begin{pmatrix} Y_i \\ Y_i^2 \end{pmatrix}$.

A_i's are iid random vectors since the Y_i's are iid's. In that case, we can write that, $\frac{1}{n}M = \frac{1}{n}\sum_{i=1}^{n} A_i = \bar{A}_n$. From the multivariate version of the central limit theorem,

$$\sqrt{n}(\bar{\boldsymbol{A}}_{\boldsymbol{n}} - \boldsymbol{\mu}) \xrightarrow{D} N_2(\mathbf{0}, \mathbf{\Sigma}),$$

where $\boldsymbol{\mu} = E(A_i) = \begin{pmatrix} (1-\omega)\lambda \\ (1-\omega)(\lambda^2+\lambda) \end{pmatrix}$

and

$$\mathbf{\Sigma} = \begin{pmatrix} Var(Y_i) & Cov(Y_i, Y_i^2) \\ Cov(Y_i^2, Y_i) & Var(Y_i^2) \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix},$$

with the elements of the variance covariance matrix $\mathbf{\Sigma}$ given by

$\sigma_{11} = \lambda(1-\omega)(\lambda+1) - (1-\omega)^2\lambda^2,$

$\sigma_{12} = \sigma_{21} = \lambda(1-\omega)[\lambda^2 + 3\lambda + 1 - (1-\omega)(\lambda+\lambda^2)],$

$\sigma_{22} = \lambda(1-\omega)[\lambda^3 + 6\lambda^2 + 7\lambda + 1 - (1-\omega)\lambda(1-\lambda)^2].$

This implies that $\sqrt{n}\left(\begin{pmatrix} M_{1n} \\ M_{2n} \end{pmatrix} - \begin{pmatrix} (1-\omega)\lambda \\ (1-\omega)(\lambda^2+\lambda) \end{pmatrix}\right) \xrightarrow{\mathbf{D}} N(\mathbf{0}, \mathbf{\Sigma}).$

To complete the proof, we rely on the following lemma.

Lemma 9 (Multivariate Delta Method)**.** *Assume that* $\boldsymbol{Y} = (Y_{n1}, Y_{n2}, Y_{n3}, \cdots, Y_{np})^\top$ *is* $AN_P(\boldsymbol{\mu}, b_n^2\Sigma)$ *with* $b_n \to 0$ *as* $n \to \infty$*. Let* $\mathbf{g} : R^p \to R^m$ *be real valued differentiable at* $\boldsymbol{x} = \boldsymbol{\mu}$ *with* $\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \cdots, g_m(\boldsymbol{x}))'$ *and* $\boldsymbol{D} = \left(\frac{dg_i}{dx_j}|_{\boldsymbol{x}=\boldsymbol{\mu}}\right)$ *for* $i = 1, \cdots, m$ *and* $j = 1, \cdots, p$ *having some non-zero elements. Then,* $\boldsymbol{g}(\boldsymbol{x_n})$ *is* $AN_m(\boldsymbol{g}(\boldsymbol{\mu}), b_n^2\boldsymbol{D\Sigma D'})$

Using the multivariate delta method and defining **g** as

$$\mathbf{g}(x,y) = \begin{pmatrix} g_1(x,y) \\ g_2(x,y) \end{pmatrix},$$

where $g_1(x,y) = 1 - \frac{x^2}{y-x}$, is continuous on $\{(x,y) : x \neq y\}$ and $g_2(x,y) = \frac{y}{x} - 1$, continuous on $\{(x,y) : x \neq 0\}$.

Hence

$$\begin{pmatrix} g_1(x,y) \\ g_2(x,y) \end{pmatrix} \sim AN_2\Bigg(\mathbf{g}(E(Y_i), E(Y_i^2)), \frac{1}{n}\boldsymbol{D\Sigma D'}\Bigg).$$

According to the multivariate delta method, where $\boldsymbol{D} \neq$null matrix

$$\mathbf{g}(E(Y_i), E(Y_i^2)) = \begin{pmatrix} \omega \\ \lambda \end{pmatrix}$$

and

$$\boldsymbol{D} = \begin{pmatrix} \frac{dg_1}{dx} & \frac{dg_1}{dy} \\ \frac{dg_2}{dx} & \frac{dg_2}{dy} \end{pmatrix} = \begin{pmatrix} \frac{-2\lambda-1}{\lambda^2} & \frac{1}{\lambda^2} \\ \frac{-(\lambda+1)}{(1-\omega)\lambda} & \frac{1}{\lambda(1-\omega)} \end{pmatrix}.$$

Therefore

$$\boldsymbol{D\Sigma D'} = \begin{pmatrix} (1-\omega)(\omega\lambda^2+2)/\lambda^2 & 2/\lambda \\ 2/\lambda & (\lambda+2)/(1-\omega) \end{pmatrix}.$$

Hence the proof. Suppose we want to test $H_0 : \omega = 0$ (the Poisson model) against $H_1 : \omega > 0$ (the ZIP model). We can make use of the asymptotic distribution of $\hat{\omega}$ proposed by Kharrati-Kopaei and Faghih (2011). Given that

$\hat{\omega} \sim N(\omega, \frac{(1-\omega)(\omega\lambda^2+2)}{n\lambda^2})$, under H_0, $\hat{\omega} \sim N(0, \frac{2}{n\lambda^2})$

$$\Longrightarrow \frac{\hat{\omega}-0}{\sqrt{\frac{2}{n\lambda^2}}} \sim N(0,1).$$

Since λ is unknown in practice we use $\hat{\lambda}$ which is a consistent estimator of λ. By Slutsky's Theorem,

$$T = \hat{\lambda}\hat{\omega}\sqrt{\frac{n}{2}} \sim N(0,1).$$

Also we are only interested in positive values of ω, therefore we redefine ω as $\hat{\omega}^+ = \max\{0, \hat{\omega}\}$. The test statistic becomes $T = \hat{\lambda}\hat{\omega}^+\sqrt{\frac{n}{2}}$. It is reasonable to reject H_0 when T is large. When $\omega \to 1$, this means our data set has a lot of structural zeroes and as such $\bar{Y}$ and $\hat{\lambda}$ may be zero. In this case we reject H_0 and conclude that $\omega > 0$.

It is reasonable to reject H_0 when $T > Z_\alpha$ or when $\bar{Y} = 0$ or $\hat{\lambda} = 0$ where Z_α is the $(1-\alpha)^{th}$ quantile of the standard normal distribution.

Proof: Since $Z_\alpha > 0$, the probability of rejecting H_0 is bounded from below and above by $\{1 - P(T > Z_\alpha)\}$ and $\{1 - P(T > Z_\alpha) + P(\bar{Y} = 0) + P(\hat{\lambda} = 0)\}$, respectively. However, as n tends to infinity the last two expression of the upper bound tends to zero therefore under the null hypothesis, as $n \to \infty$

$$P\bigg(\{T > Z_\alpha\} \ \cup \ \{\bar{Y} = 0\} \ \cup \ \{\hat{\lambda} = 0\}\bigg) \to 1 - (1-\alpha) = \alpha.$$

2.1.11 Confidence Interval

This subsection provides a detailed procedure of constructing a confidence interval for $\omega \in [0, 1)$. If $Y_1, Y_2, \cdots, Y_n$ is a random sample from ZIP(ω, λ) and $\hat{\omega}^+ = \{0, \hat{\omega}\}$ then as $n \to \infty$, a $100(1-\alpha)\%$ confidence interval of ω is given by

$$\left(\frac{2\hat{\omega}^+ + \delta(1 - 2/\hat{\lambda}^2)}{2(1+\delta)}\right) \pm \sqrt{\frac{(2/\hat{\lambda}^2)\delta - (\hat{\omega}^+)^2}{(1+\delta)} + \left(\frac{2\hat{\omega}^+ + \delta(1 - 2/\hat{\lambda}^2)}{2(1+\delta)}\right)^2}$$

where $\delta = Z^2_{\alpha/2}/n$ (see Kharrati-Kopaei and Faghih, 2011)

Proof: we know that $\hat{\omega} \sim N(\omega, V_n(\omega))$ approximately for large n, where $V_n(\omega) = \frac{(1-\omega)(\omega\lambda^2+2)}{n\lambda^2}$. But from Slutsky's theorem

$$P\left(\frac{|\hat{\omega}^+ - \omega|}{\sqrt{V_n(\omega)}} < Z_{\alpha/2}\right) \geq P\left(\frac{|\hat{\omega} - \omega|}{\sqrt{V_n(\omega)}} < Z_{\alpha/2}\right)$$

Therefore we want to show that as $n \to \infty$

$$P\left(\frac{|\hat{\omega}^{+} - \omega|}{\sqrt{V_n(\omega)}} > Z_{\alpha/2}\right) \geq (1-\alpha).$$

By solving the inequality $(\hat{\omega}^{+} - \omega)^2 > Z^2_{\alpha/2} V_n(\omega)$, we obtain the result given above.

The confidence interval above is applicable when $\bar{Y} \neq 0$ and $\hat{\lambda} \neq 0$. However, if the observed value of $\hat{\lambda}$ or $\bar{Y}$ is zero, we define $\omega = 1$ and with these new conditions the confidence interval above still has a coverage probability of at least $(1-\alpha)$ for large n. Using the test statistic $W_n = \frac{|\hat{\omega}^{+} - \omega_0|}{\sqrt{V_n(\omega)}}$, we can test the null hypothesis $H_0 : \omega = \omega_0$ against the alternate $H_1 : \omega \neq \omega_0$. The null hypothesis is rejected at level α when $W_n \geq Z_{\alpha/2}$.

Below is a Q-Q plot obtained based on the zero inflated Poisson distribution with zero inflation parameter $\omega = 0.3$ and Poisson intensity parameter $\lambda = 2.5$. We generated 1000 different ZIP samples and the MLE obtained using the EM algorithm. The Q-Q plots below are based on 1000 estimates for sample sizes $n = 25, 50, 100$ and 250.

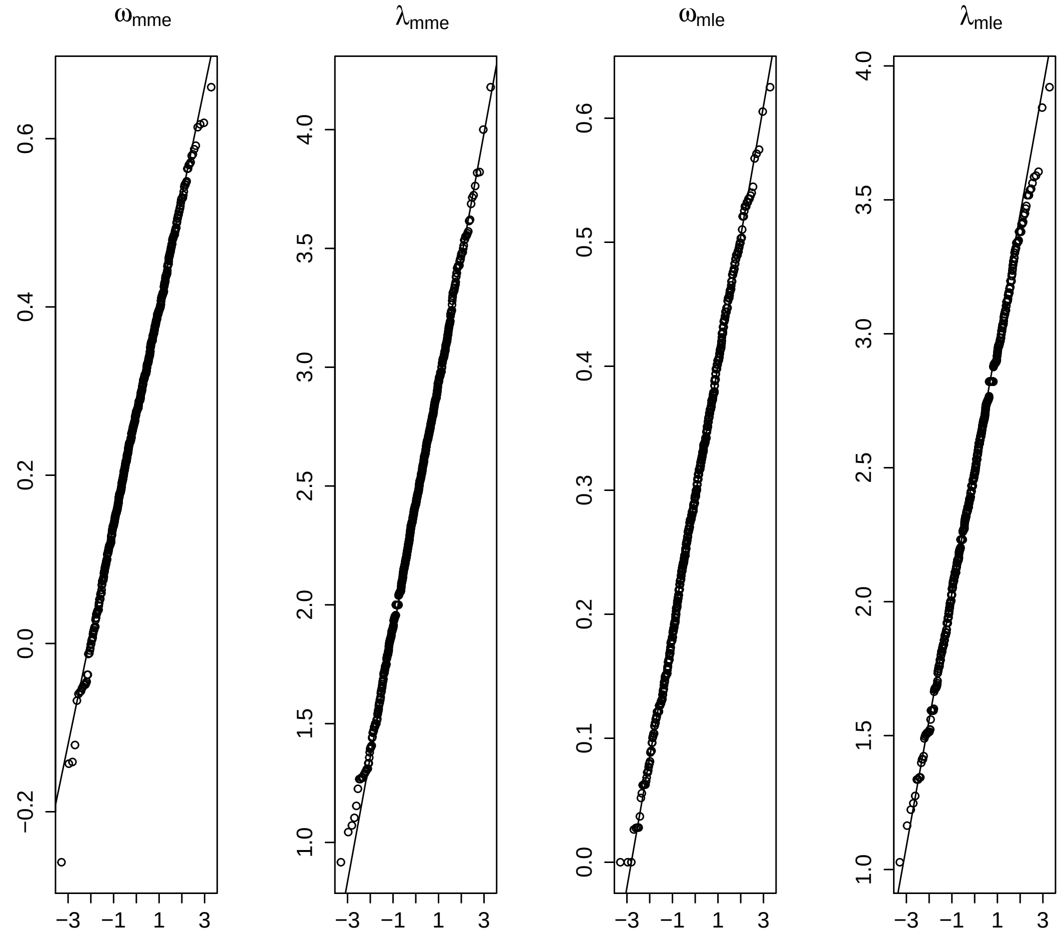

Figure 2.1: Q-Q plots of the MLEs and MMEs of λ and ω when $n = 25$ each is drawn from the ZIP distribution with $\lambda = 2.5$ and $\omega = 0.3$.

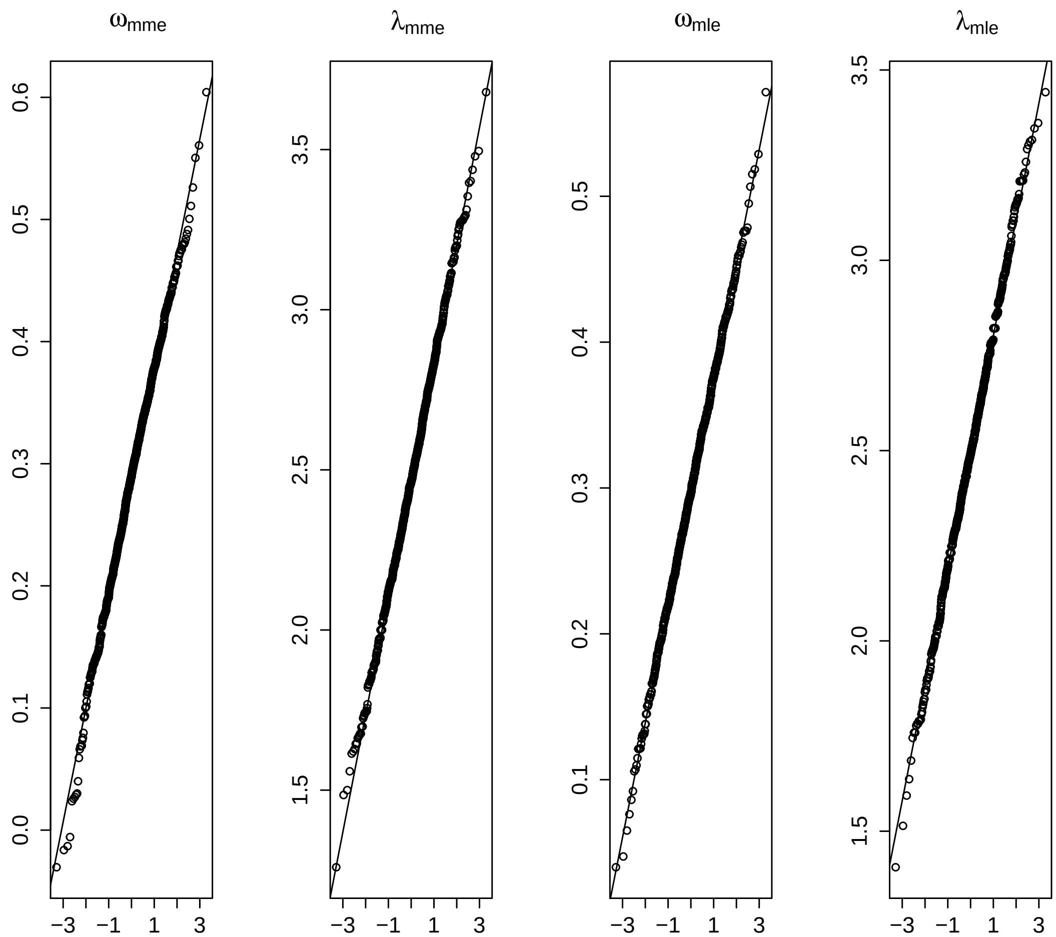

Figure 2.2: Q-Q plots of the MLEs and MMEs of λ and ω when $n = 50$ each is drawn from the ZIP distribution with $\lambda = 2.5$ and $\omega = 0.3$.

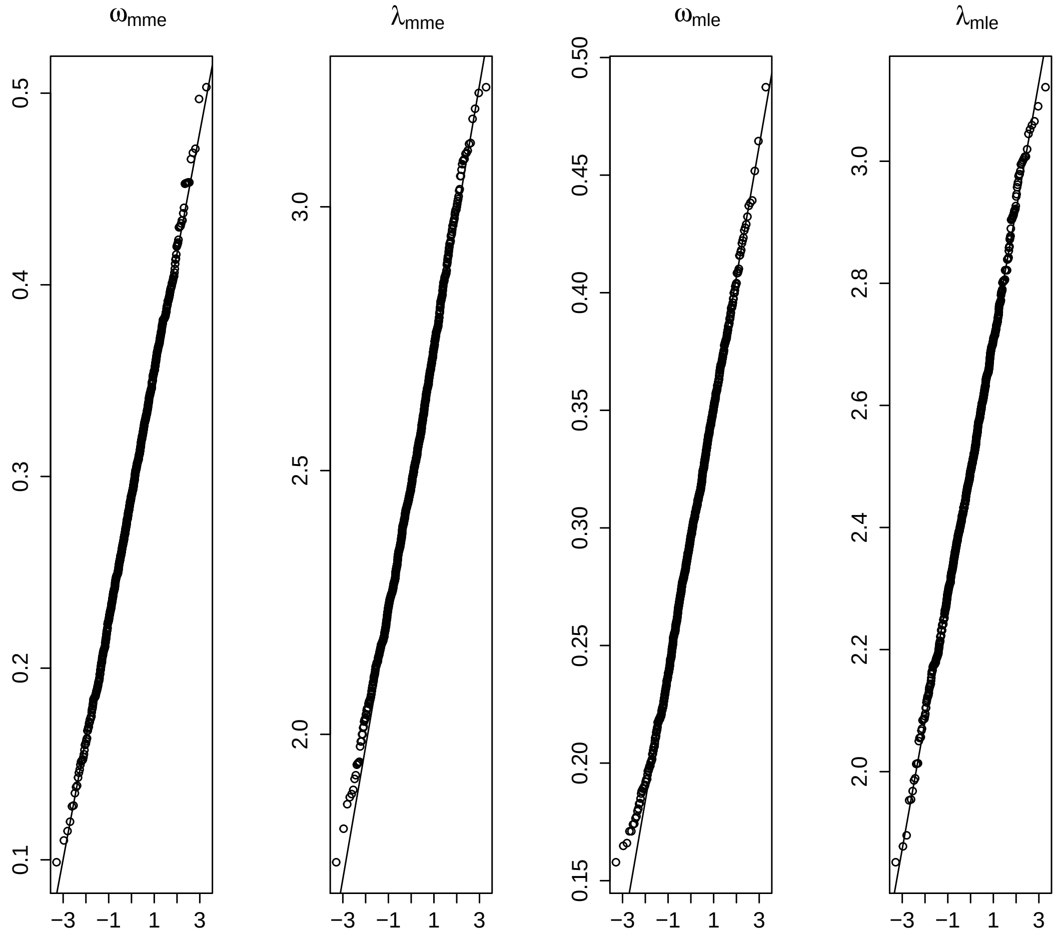

Figure 2.3: Q-Q plots of the MLEs and MMEs of λ and ω when $n = 100$ each is drawn from the ZIP distribution with $\lambda = 2.5$ and $\omega = 0.3$.

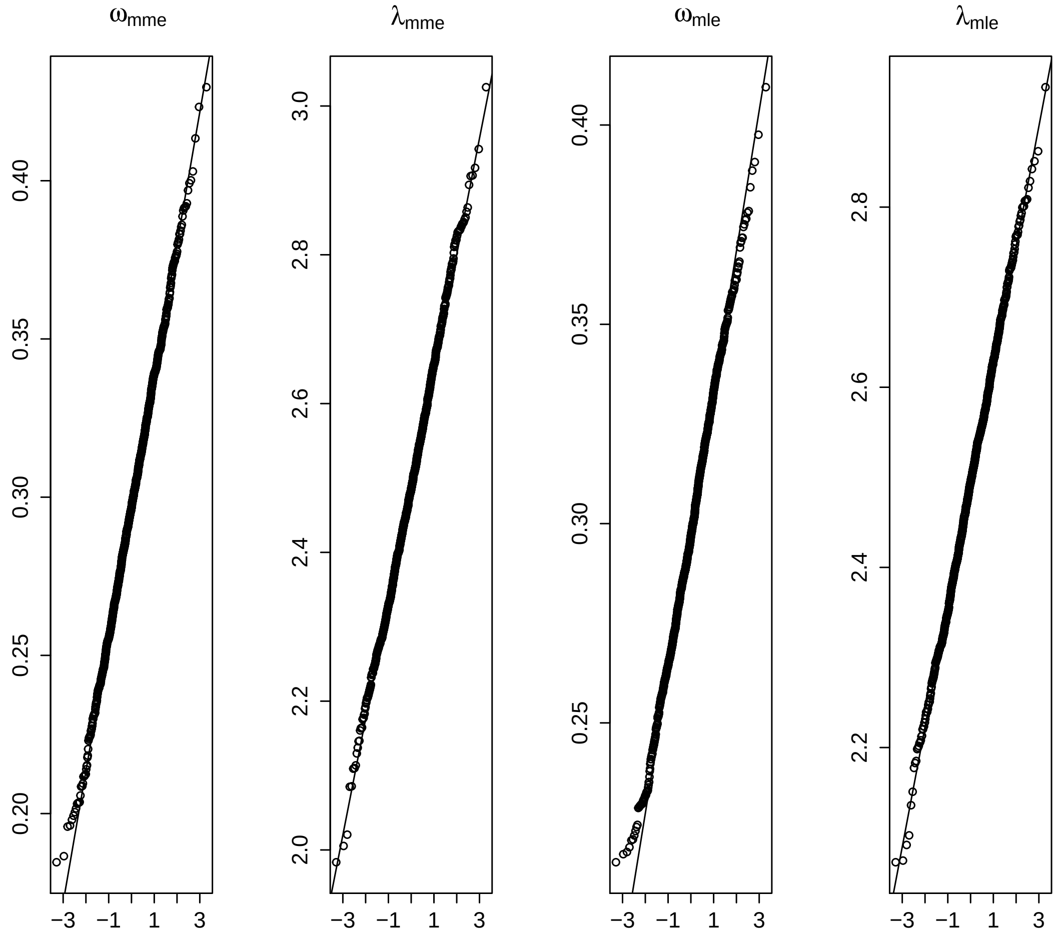

Figure 2.4: Q-Q plots of the MLEs and MMEs of λ and ω when $n = 250$ each is drawn from the ZIP distribution with $\lambda = 2.5$ and $\omega = 0.3$.

Table 2.3: Results of estimates of the mean square error (MSE), kurtosis and skewness of the parameter estimates associated with the ZIP distribution $\lambda = 2.5$ and $\omega = 0.3$.

Measure	λ_{mme}	ω_{mme}	λ_{mle}	ω_{mle}
Kurtosis	2.9114	3.1904	2.9240	2.7737
Skewness	0.0483	-0.1547	-0.0892	-0.0063
MSE	0.5337	0.0351	0.4144	0.0228

The results of Table 2.3 are based on a sample size of $n = 25$ and 1000 iterations. For normality, we expect kurtosis to be 3 and skewness to be zero. For a sample size of only $n = 25$, Table 2.3 suggests that the sampling distribution of the estimates are normal. This results agrees with the Q-Q plot in Figure (2.4) above. The tables for sample sizes $n = 50$, $n = 100$ and $n = 250$ are shown in the Appendix.

2.1.12 The Zero Inflated Autoregressive conditional Poisson (ZIACP)(p, q) Linear Model

Recently Zhu (2012) studied the zero inflated count time series model by extending the work of Fokianos et al. (2009). In the same year, Yang (2012) studied the autoregression for zero inflated count time series loglinear model.

Let $\{Y_t\}$ be a time series of counts. We assume that, conditional on $\mathcal{F}_{t-1}$, the random variables $Y_1, \cdots, Y_n$ are independent, and the conditional distribution of Y_t is specified by a ZIP distribution, that is,

$$Y_t|\mathcal{F}_{t-1} \sim ZIP(\lambda_t, w),\ \lambda_t = \gamma_0 + \sum\nolimits_{i=1}^{p} \alpha_i Y_{t-i} + \sum\nolimits_{j=1}^{q} \beta_j \lambda_{t-j}, \tag{2.15}$$

where $0 < w < 1,\ \ \gamma_0 > 0,\ \ \alpha_i \geq 0,\ \ \beta_j \geq 0, i = 1, \cdots, p\ \ , j = 1, \cdots, q\ \ , p \geq 0$. $\mathcal{F}_{t-1}$ is the σ-field generated by $\{Y_{t-1}, Y_{t-2}, \cdots\}$. The above model is denoted as ZIACP(p, q). The conditional mean and the conditional variance of Y_t are given by $E(Y_t|\mathcal{F}_{t-1}) = (1-\omega)\lambda_t$ and $Var(Y_t|\mathcal{F}_{t-1}) = (1-\omega)\lambda_t(1+\omega\lambda_t)$, respectively. Therefore, $Var(Y_t|\mathcal{F}_{t-1}) > E(Y_t|\mathcal{F}_{t-1})$. It can also be shown that

$$\begin{aligned} Var(Y_t) = E(Var(Y_t|\mathcal{F}_{t-1})) + Var(E(Y_t|\mathcal{F}_{t-1})) = E((1-\omega)\lambda_t(1+\omega\lambda_t)) \\ + \quad Var((1-\omega)\lambda_t) \end{aligned}$$

$$= (1-\omega)E(\lambda_t) + \omega(1-\omega)(E(\lambda_t)^2) + (1-\omega)^2 Var(\lambda_t) > (1-\omega)E(\lambda_t) = E(Y_t),$$

which means that model (2.15) can handle integer-valued time series with overdispersion.

2.1.13 The ZIACP(p, q) Model in an ARMA form

We consider the case when $p = 1$ and $q = 1$. If we define the martingale difference u_t as $u_t = Y_t - E(Y_t|\mathcal{F}_{t-1}) = Y_t - \lambda_t(1-\omega)$ then from $\lambda_t = \gamma_0 + \alpha_1 Y_{t-1} + \beta_1 \lambda_{t-1}$ we can write,

$$\begin{aligned} \lambda_t(1-\omega) &= (1-\omega)\gamma_0 + (1-\omega)\alpha_1 Y_{t-1} + (1-\omega)\beta_1 \lambda_{t-1} \\ Y_t - (1-\omega)\alpha_1 Y_{t-1} - \beta_1 Y_{t-1} &= (1-\omega)\gamma_0 + u_t - \beta_1 u_{t-1} \end{aligned} \tag{2.16}$$

$\implies Y_t - \phi_1 Y_{t-1} = (1-\omega)\gamma_0 + u_t - \beta_1 u_{t-1}$ where $\phi = \beta_1 + (1-\omega)\alpha_1$. Comparing this model to the standard ARMA(1,1) model of the form $Y_t - \phi Y_{t-1} = \gamma + a_t + \theta a_{t-1}$ (We use this form of the model as it is the same form that is used for fit in the standard **R** package which was used in the analysis) suggests that our initial estimate of $\hat{\beta}_1 = -\hat{\theta}$, $\hat{\alpha}_1 = \frac{\hat{\phi}+\hat{\theta}}{(1-\hat{\omega})}$ and since the ARMA(1,1) model has mean μ, $\mu = \frac{(1-\omega)\gamma_0}{1-\phi_1} \implies \hat{\gamma}_0 = \frac{(1-\hat{\phi})\hat{\mu}}{(1-\hat{\omega})}$.

For the case when $p = 1$ and $q = 0$, $\lambda_t = \gamma_0 + \alpha_1 Y_{t-1}$ we generate initial values for α_0 and α_1 by fitting an $AR(1)$ model. The details are shown below. We rewrite the model in the form of an AR(1) process by letting $u_t = Y_t - E(Y_t|\mathcal{F}) = Y_t - \lambda_t(1-\omega)$

$$\lambda_t(1-\omega) = (1-\omega)\gamma_0 + (1-\omega)\alpha_1 Y_{t-1}$$

$$Y_t - (1-\omega)\alpha_1 Y_{t-1} = (1-\omega)\gamma_0 + u_t$$

$$Y_t - \phi Y_{t-1} = (1-\omega)\gamma_0 + u_t \tag{2.17}$$

where $\phi = (1-\omega)\alpha_1$. Given that the Y_t process is stationary i.e., $|\phi_1| < 1$, we can write μ as $\mu = \frac{(1-\omega)\gamma_0}{1-(1-\omega)\alpha_1} \implies \hat{\alpha}_1 = \frac{\hat{\phi}}{1-\hat{\omega}}$ and $\hat{\gamma}_0 = \frac{\hat{\mu}(1-\hat{\phi})}{1-\hat{\omega}}$.

Theorem 10. *Let $\{Y_t\}$ be a time series of counts such that $Y_t \sim$ ZIACP(1,1) model, then the marginal mean, variance and auto correlations are given by*

$$\mu = E(Y_t) = \frac{(1-\omega)\gamma_0}{1-(1-\omega)\alpha_1 - \beta_1},$$

$$Var(Y_t) = \frac{1-2(1-\omega)\alpha_1\beta_1 - \beta_1^2}{1-(1-\omega)\alpha_1^2 - 2(1-\omega)\alpha_1\beta_1 - \beta_1^2}\left(\mu + \frac{\omega\mu^2}{1-\omega}\right),$$

$$\rho_y(k) = [(1-\omega)\alpha_1 + \beta_1]^{k-1}\frac{(1-\omega)\alpha_1[1-(1-\omega)\alpha_1\beta_1 - \beta_1^2]}{1-2(1-\omega)\alpha_1\beta_1 - \beta_1^2}, \quad k \geq 1,$$

respectively.

Proof: Given that $Y_t|\mathcal{F}_{t-1} \sim ZIACP(1,1)$, from (2.16) it has the form of an ARMA(1,1) model with the autoregressive parameter $\phi = \beta_1 + \alpha_1(1-\omega)$ and the moving average parameter $\theta = \beta_1$. After reparameterization, we have,

$$Y_t - \phi Y_{t-1} = \gamma + a_t - \theta a_{t-1}.$$

The expected value of Y_t is given by,

$$E(Y_t) = \frac{(1-\omega)\gamma_0}{1-(1-\omega)\alpha_i - \beta_1}.$$

Writing it in terms of $\psi_j' s,\ \psi_j = \phi^{j-1}(\phi - \theta) = [\beta_1 + \alpha_1(1-\omega)]^{j-1}\Bigg(\alpha_1(1-\omega)\Bigg).$

Also the variance is

$$Var(Y_t) = \sigma_u^2 \sum_{j=0}^{\infty} \psi_j^2,$$

where,

$$\begin{aligned}
\sum\nolimits_{j=0}^{\infty} \psi_j^2 &= 1 + \alpha_1^2(1-\omega)^2 + [\beta_1 + \alpha_1(1-\omega)]^2 \alpha_1^2 (1-\omega)^2 \\
&\quad + [\beta_1 + \alpha_1(1-\omega)]^4 \alpha_1^2 (1-\omega)^2 + \cdots \\
&= 1 + \alpha_1^2(1-\omega)^2\{1 + [\beta_1 + \alpha_1(1-\omega)]^2 + [\beta_1 + \alpha_1(1-\omega)]^4 + \cdots\} \\
&= 1 + \frac{\alpha_1^2(1-\omega)^2}{1-[\beta_1 + \alpha_1(1-\omega)]^2} = \frac{1 - 2\alpha_1\beta_1(1-\omega) - \beta_1^2}{1-(1-\omega)\alpha_1^2 - 2(1-\omega)\alpha_1\beta_1 - \beta_1^2}.
\end{aligned}$$

But

$$\sigma_u^2 = Var(Y_t|\mathcal{F}_{t-1}^{Y,\lambda}) = (1-\omega)\lambda_t(1+\omega\lambda_t).$$

Also from (2.15), $E(Y_t) = (1-\omega)E(\lambda_t)$. Here λ_t is a function of past observations and past values of itself. The distribution of λ_t conditional on information at time $t-1$ is the same as it's unconditional distribution and so time invariant $\implies \lambda_t \to \frac{\mu}{1-\omega}$. Therefore,

$$\sigma_u^2 = (1-\omega)\frac{\mu}{1-\omega}(1+\omega\frac{\mu}{1-\omega}) = \left(\mu + \frac{\omega\mu^2}{(1-\omega)}\right).$$

Hence,

$$Var(Y_t) = \frac{1-2(1-\omega)\alpha_1\beta_1 - \beta_1^2}{1-(1-\omega)\alpha_1^2 - 2(1-\omega)\alpha_1\beta_1 - \beta_1^2}\left(\mu + \frac{\omega\mu^2}{1-\omega}\right).$$

Finally, for the autocorelation function, we have

$$\rho_y(k) = \frac{\sum_{j=0}^{\infty}\psi_{j+k}\psi_j}{\sum_{j=0}^{\infty}\psi_j^2},$$

where

$$\sum_{j=0}^{\infty}\psi_{j+k}\psi_j = \alpha_1(1-\omega)[\beta+\alpha_1(1-\omega)]^{k-1} + \alpha_1^2(1-\omega)^2[\beta+\alpha_1(1-\omega)]^k$$

$$+\alpha_1^2(1-\omega)^2[\beta+\alpha_1(1-\omega)]^{k+2}\cdots$$

$$= \alpha_1(1-\omega)[\beta_1+\alpha_1(1-\omega)]^{k-1}$$

$$+\alpha_1^2(1-\omega)^2[\beta_1+\alpha_1(1-\omega)]^k\Bigg(1+[\beta_1+\alpha_1(1-\omega)]^2$$

$$+[\beta_1+\alpha_1(1-\omega)]^4+\cdots\Bigg)$$

$$= \frac{\alpha_1(1-\omega)[\beta_1+\alpha_1(1-\omega)]^{k-1}\left(1-\alpha_1\beta_1(1-\omega)-\beta_1^2\right)}{1-(1-\omega)\alpha_1^2-2(1-\omega)\alpha_1\beta_1-\beta_1^2}.$$

Therefore,

$$\rho_y(k) = [(1-\omega)\alpha_1 + \beta_1]^{k-1}\frac{(1-\omega)\alpha_1[1-(1-\omega)\alpha_1\beta_1 - \beta_1^2]}{1-2(1-\omega)\alpha_1\beta_1 - \beta_1^2}, \quad k \geq 1.$$

Hence the proof.

Similarly for the $ZIACP(2)$ model i.e., $p = 2$ and $q = 0$ which has λ_t modeled as $\lambda_t = \gamma_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2}$ we generate initial valuesfor γ_0 , α_1 and α_2 by fitting an $AR(2)$ model. The details are shown below.

Since $\lambda_t = \gamma_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2}$ and $u_t = Y_t - E(Y_t|\mathcal{F}) = Y_t - \lambda_t(1-\omega) \implies \lambda_t(1-\omega) = (1-\omega)\gamma_0 + (1-\omega)\alpha_1 Y_{t-1} + (1-\omega)\alpha_2 Y_{t-2}$.

Hence,

$$Y_t - (1-\omega)\alpha_1 Y_{t-1} - (1-\omega)\alpha_2 Y_{t-2} = (1-\omega)\gamma_0 + u_t$$

$$Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} = 1-\omega)\gamma_0 + u_t$$

where $\hat{\phi}_1 = (1-\hat{\omega})\hat{\alpha}_1$ and $\hat{\phi}_2 = (1-\hat{\omega})\hat{\alpha}_2$, $\hat{\mu} = \frac{(1-\hat{\omega})\hat{\gamma}_0}{1-\hat{\phi}_1-\hat{\phi}_2}$ given that the Y_t process is stationary i.e., $\phi_1 + \phi_2 < 1$, $\phi_1 - \phi_2 < 1$ and $-1 < \phi_2 < 1$.

$\implies \hat{\alpha}_1 = \frac{\hat{\phi}_1}{1-\hat{\omega}}$, $\hat{\alpha}_2 = \frac{\hat{\phi}_2}{1-\hat{\omega}}$ and $\hat{\gamma}_0 = \frac{\hat{\mu}(1-\hat{\phi}_1-\hat{\phi}_2)}{1-\hat{\omega}}$.

Theorem 11. *If $\{Y_t\}$ is a time series of counts such that $Y_t \sim$ ZIACP(1) model, then the marginal mean, variance and auto correlations are given by*

$$\mu = E(Y_t) = \frac{(1-\omega)\gamma_0}{1-(1-\omega)\alpha_i},$$

$$Var(Y_t) = \frac{(1-\omega)\gamma_0[1+\omega\gamma_0 - (1-\omega)\alpha_1]}{[1-(1-\omega)\alpha_1^2][1-(1-\omega)\alpha_1]^2},$$

$$\rho_y(k) = [(1-\omega)\alpha_1]^k, \quad k \geq 1,$$

respectively.

Theorem 12. *Let $\{Y_t\}$ be a time series of counts such that $Y_t \sim$ ZIACP(2) model, then the marginal mean and variance are given by*

$$\mu = E(Y_t) = \frac{(1-\omega)\gamma_0}{1-(1-\omega)(\alpha_1+\alpha_2)},$$

$$Var(Y_t) = \frac{(1-\omega)E(Y_t)+\omega[E(Y_t)]^2}{1-\omega-(1-\omega)^2[\frac{1+(1-\omega)\alpha_2}{1-(1-\omega)\alpha_2}\alpha_1^2+\alpha_2^2]},$$

respectively.

2.1.14 ZIACP Parameter Estimation

Let $\{Y_t\}_{i=1}^n$ be a time series of counts generated from model (2.15) with $p = 1$ and $q = 1$. Assume that zero observations come from a distribution with point mass at zero i.e., structural zeros or from the Poisson distribution. We introduce a Bernoulli random variable Z_t such that $Z_t = 1$ when $Y_t = 0$ is from a degenerate distribution and $Z_t = 0$ when $Y_t = 0$ is from a Poisson process. Practically, Z_t's are not observed, and thus, Z_t's are missing values. Let $\boldsymbol{Z} = (Z_1, Z_2, \cdots, Z_n)$, $\boldsymbol{\theta} = (\gamma_0, \alpha_1 \cdots, \alpha_p, \beta_1 \cdots, \beta_q)^T = (\theta_0, \theta_1, \cdots, \cdots, \theta_{p+q})$, and $\boldsymbol{\Theta} = (\omega, \boldsymbol{\theta}^T)^T$. The conditional likelihood function for the complete data is given by

$$\mathcal{L}(\boldsymbol{\Theta}) = \prod_{t=1}^{n} \omega^{Z_t} \left((1-\omega) \frac{\lambda_t^{Y_t} e^{-\lambda_t}}{Y_t!} \right)^{1-Z_t}.$$

Therefore the conditional log-likelihood is

$$\ell_c(\boldsymbol{\Theta}) = \sum_{t=1}^{n} \{Z_t \log \omega + (1-Z_t)[\log(1-\omega) + Y_t \log \lambda_t - \lambda_t - \log(Y_t!)]\}.$$

We can write the log-likelihood up to an additive constant as,

$$\implies \ell_c(\mathbf{\Theta}) = \sum_{t=1}^{n}(1-Z_t)(Y_t \log\lambda_t - \lambda_t) + \sum_{t=1}^{n}\{Z_t \log\omega + (1-Z_t)\log(1-\omega)\}. \quad (2.18)$$

The first derivative of the log-likelihood with respect to ω and θ_i are

$$\frac{\partial \ell_c}{\partial \omega} = \sum_{t=1}^{n}\left(\frac{Z_t}{\omega} - \frac{1-Z_t}{1-\omega}\right). \quad (2.19)$$

$$\frac{\partial \ell_c}{\partial \theta_i} = \sum_{t=1}^{n}(1-Z_t)\left(\frac{Y_t}{\lambda_t} - 1\right)\frac{\partial \lambda_t}{\partial \theta_i}. \quad (2.20)$$

Clearly, the likelihood above is a function of the missing Z_t values therefore ℓ_c cannot be maximized directly. We therefore use the EM algorithm to estimate the parameters since it will allow us to estimate the missing values. Obtaining an initial guess for the missing data Z_t is the basis for the E-step.

$$\begin{aligned}
E(Z_t|Y_t = y_t, \mathcal{F}_{t-1}) &= 1 \times P(Z_t = 1|Y_t = y_t, \mathcal{F}_{t-1}) + 0 \times P(Z_t = 0|Y_t = y_t, \mathcal{F}_{t-1}) \\
&= P(Z_t = 1|Y_t = y_t, \mathcal{F}_{t-1}) \\
&= \frac{P(Z_t = 1|\mathcal{F}_{t-1})P(Y_t = y_t|Z_t = 1, \mathcal{F}_{t-1})}{P(Y_t = y_t|\mathcal{F}_{t-1})} \\
&= \begin{cases} \frac{\omega}{w+(1-\omega)e^{-\lambda_t}}, \; y_t = 0 \\ 0, \quad y_t = 1, 2, 3, \cdots \end{cases} .
\end{aligned}$$

$E - step$: We compute the conditional expectation $Q(\theta|\theta^{(j)})$ of $\ell_c(\mathbf{\Theta})$:

$$\begin{aligned}
Q(\theta|\theta^{(j)}) &= E\{\ell_c(\Theta)|y, \theta^{(j)}\} \\
&= \sum_{t=1}^{n}(1 - \hat{Z_t}^{(j)})(Y_t \log\lambda_t - \lambda_t) + \sum_{t=1}^{n}\{\hat{Z_t}^{(j)} \log\omega + (1 - \hat{Z_t}^{(j)})\log(1-\omega)\},
\end{aligned}$$

where $\hat{Z}_t^{(j)}$ denotes the conditional expectation of Z_t at the j^{th} iteration.

$M - step$: In this step we find $\theta^{(j+1)}$ that maximizes $Q(\theta|\theta^{(j)})$. From (2.18) we can maximize the two terms independently to obtain the maximum likelihood estimates of the parameters.

$$\sum_{t=1}^{n}(1 - \hat{Z}_t^{(j)})(Y_t \log \lambda_t - \lambda_t) \tag{2.21}$$

$$\sum_{t=1}^{n}\{\hat{Z}_t^{(j)} \log \omega + (1 - \hat{Z}_t^{(j)}) \log(1 - \omega)\}. \tag{2.22}$$

Upon a careful look at the equation (2.21) we see a close relationship between this log-likelihood and log-likelihood of the Poisson process considered under Poisson autoregression above. Therefore in oder to obtain the parameter estimates we modify the likelihood and the conditional information matrix of the Poisson process by Fokianos et al. (2009). The maximum likelihood estimates of (2.21) are obtained by solving numerically using the **optim()** function in **R**. For (2.22) the maximum likelihood estimate of ω is obtained trivially by equating its derivative with respect to ω to zero and solving for $\hat{\omega}$ in terms of $\hat{Z}_t$.

2.1.15 Simulation studies

We conduct a simulation study to estimate the parameters of the ZIACP model. We generate the ZIP data using the model below,

$$Y_t|\mathcal{F}_{t-1} \sim ZIP(\lambda_t, w),\ \lambda_t = \gamma_0 + \alpha_1 Y_{t-1} + \beta_1 \lambda_{t-1}, \tag{2.23}$$

We consider true parameters of the model to be $(\omega, \gamma_0, \alpha_1, \beta_1) = (0.1, 1.0, 0.4, 0.3)$. The simulated data is generated by first generating λ_t's using its recursive formula above. The initial value of λ_t i.e., λ_1 was chosen to be γ_0. Next we generate data from $Uniform(0, 1) = U$ with sample size n. For all n sample data generated from the uniform distribution, we test the condition whether or not a data point is less than or equal to ω or greater than ω. When a data point is less than or equal to ω then our simulated data is zero (point mass zero distribution). On the other hand when the data point from the uniform distribution is greater than ω we generate the simulated data from the Poisson distribution with the corresponding intensity parameter. Once we have the simulated data, we fit the ZIACP model and estimate the parameters. In estimating the parameters of the model, we use the EM algorithm above. We first estimate the missing data Z_t's by computing its expected value according to the formula $E(Z_t|Y_t = y_t, \mathcal{F}_{t-1}) = \frac{\omega}{w+(1-\omega)e^{-\lambda_t}}$. We obtained the initial estimate of ω as the ratio of the number of zeros in the simulated data to the total sample size. Next we generated λ_t's using initial estimates of γ_0, α_1 and β_1. The initial estimates of γ_0, α_1 and β_1 was obtained by fitting an ARMA(1,1) model based on the simulated data. Based on the ARMA(1,1), initial values were calculated to be $\hat{\beta}_1 = -\hat{\theta}$, $\hat{\alpha}_1 = \frac{\hat{\phi}+\hat{\theta}}{(1-\hat{\omega})}$ and $\hat{\gamma}_0 = \frac{(1-\hat{\phi})\hat{\mu}}{(1-\hat{\omega})}$ where $\hat{\phi}$ and $\hat{\theta}$ is the estimate of the AR parameter and MA parameter of the ARMA(1,1) fit, respectively. With the missing data in place, we obtained the MLE of ω as suggested by (2.22). Next, we obtained the MLE's of γ_0, α_1 and β_1 by adjusting the likelihood, the score and information matrix for the missing data Z_t as in (2.20) and (2.21). The above process is iterated 1000 times and in every iteration the estimates of γ_0, α_1 and β_1 is stored to calculate the average estimates. We evaluate the estimates by computing the MSE and MADE as in Zhu (2012) .

$$MADE = 1/a \sum_{i=1}^{a} |\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}|$$

where a is the number of iterations. The simulation process was also carried out for different sets of parameters values as in (Zhu, 2012). The results are presented below.

The time series plot, histogram, ACF and PACF of the simulated data is shown in Figure 2.5 (data was simulated based on the sample size $n = 200$).

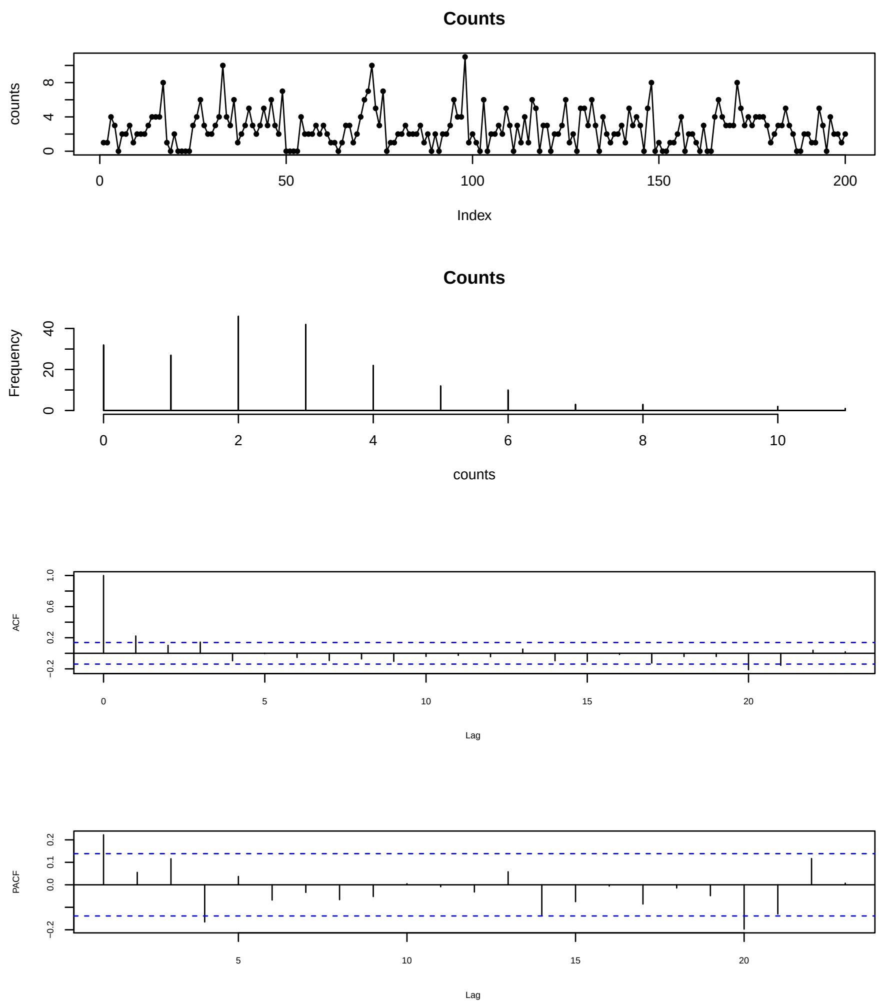

Figure 2.5: Time series Plot, ACF and PACF of ZIP data generated with the model parameters $\{\omega = 0.1, \gamma_0 = 1, \alpha_1 = 0.4, \beta_1 = 0.3\}$ based on sample size $n = 200$.

The dispersion index is a measure of over dispersion that gives us an initial impression of the zero inflation in the data set. For the simulated data the dispersion index is 1.739 indicating the disparity between the mean and the variance of the data. This means that the ordinary Poisson distribution will fail to model or explain such data accurately, hence we resort to the ZIP distribution to model the count time series data. The table below shows the maximum likelihood estimates (MLE) and its corresponding MADE and MSE for different sample sizes.

Table 2.4: Simulation results obtained for the $ZIACP(1,1)$ model for the parameter vectors (ω,γ_0,α_1,β_1)=(0.1,1.0,0.4,0.3). Parameters are adapted as in Zhu (2012).

Parameters	Sample size	MLE	MADE	MSE
ω	200	0.1013	0.0216	0.0015
γ_0		1.2635	0.3860	0.5025
α_1		0.4275	0.0789	0.0194
β_1		0.2460	0.1391	0.0638
ω	500	0.1000	0.0140	0.0006
γ_0		1.1435	0.2333	0.1618
α_1		0.4282	0.0537	0.0083
β_1		0.2847	0.0842	0.0224
ω	1000	0.0998	0.0102	0.0003
γ_0		1.1314	0.1794	0.0915
α_1		0.4315	0.0424	0.0046
β_1		0.2861	0.0582	0.0108

Table 2.5: Simulation results obtained for the $ZIACP(1,1)$ model for the parameter vectors (ω,γ_0,α_1,β_1)=(0.15,2,0.3,0.2). Parameters are adapted as in Zhu (2012).

Parameters	Sample size	MLE	MADE	MSE
ω	200	0.1374	0.0250	0.0017
γ_0		2.3811	0.7693	2.0222
α_1		0.3011	0.0736	0.0173
β_1		0.1213	0.2152	0.1554
ω	500	0.1380	0.0179	0.0008
γ_0		2.1974	0.4671	0.7078
α_1		0.3061	0.0465	0.0068
β_1		0.1673	0.1333	0.0568
ω	1000	0.1384	0.0141	0.0005
γ_0		2.1483	0.3411	0.3725
α_1		0.3079	0.0328	0.0034
β_1		0.1797	0.0928	0.0281

Tables 2.4 and 2.5 show the simulation results obtained for the $ZIACP(1,1)$ model for the parameter vectors (ω,γ_0,α_1,β_1)=(0.1,1,0.4,0.3) and (ω,γ_0,α_1,β_1)=(0.15,2,0.3,0.2), respectively . The Tables validate the ZIACP modelling procedure, in the sense that, the estimates obatained are very close to the true parameter values used in the data generation. The estimates become even better as the sample size increases. We also considered a simulation study for the ZIACP(1) and ZIACP(2) model with different true parameter vectors. The remainder of the simulation results have been provided in the Appendix. Figures (2.6) and (2.7) show a Q-Q plot to demonstrate the normality of the estimates obtained for the ZIACP(1,1) model.

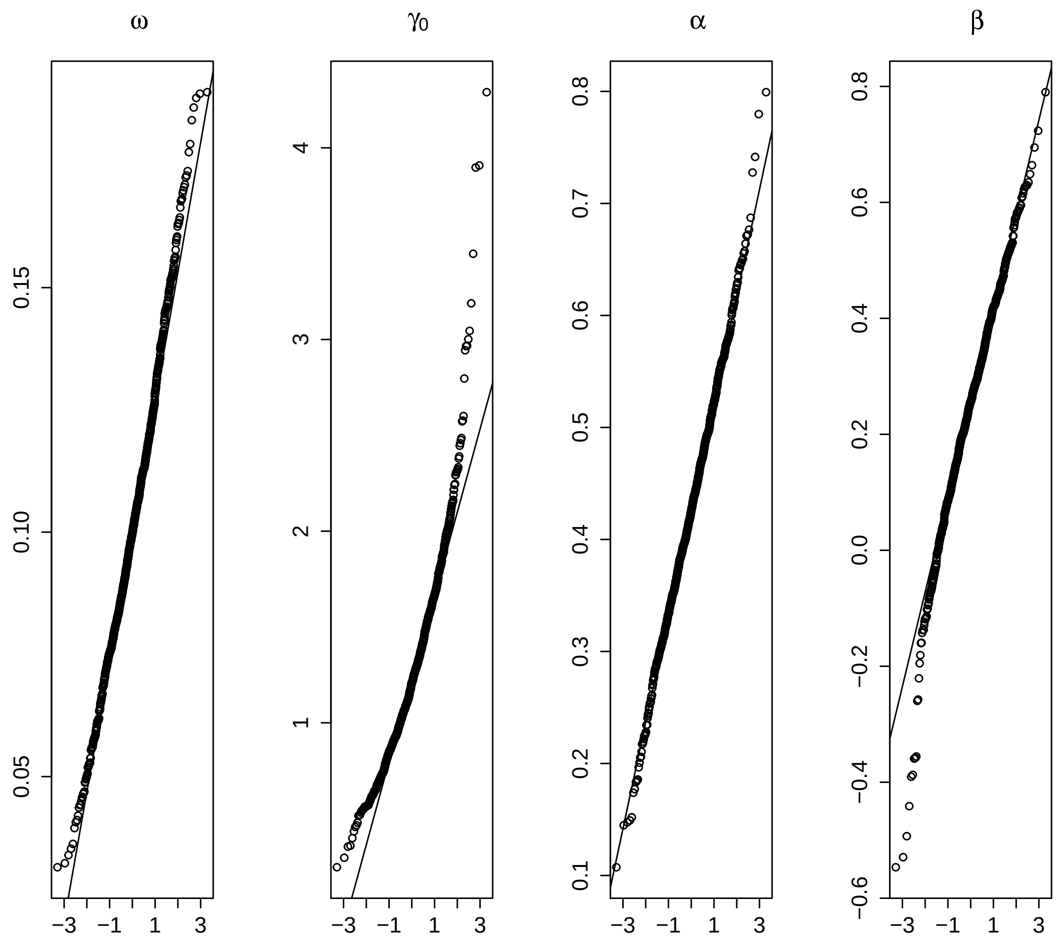

Figure 2.6: A Q-Q plot demonstrating normality of the ZIP estimates for the model parameters $\{\omega = 0.1, \gamma_0 = 1, \alpha_1 = 0.4, \beta_1 = 0.3\}$ based on sample size $n = 200$.

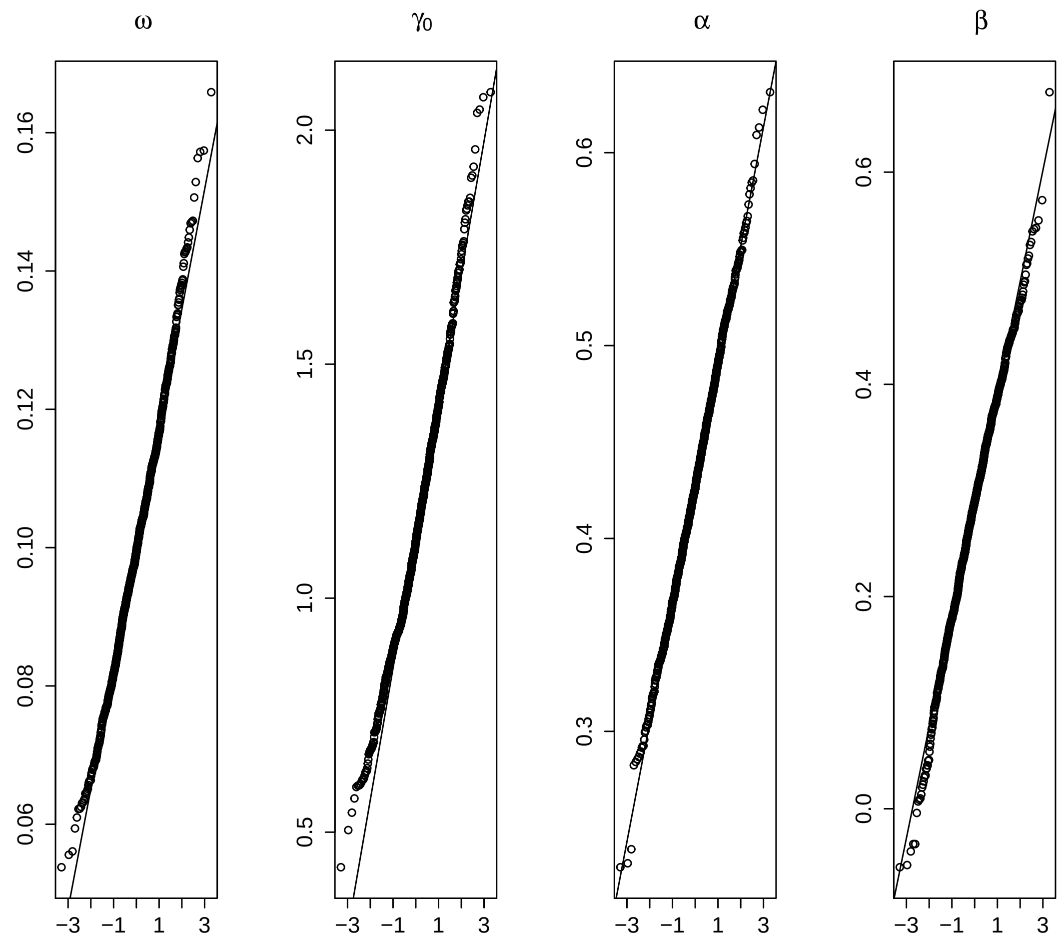

Figure 2.7: A Q-Q plot demonstrating normality of the ZIP estimates for the model parameters $\{\omega = 0.1, \gamma_0 = 1, \alpha_1 = 0.4, \beta_1 = 0.3\}$ based on sample size $n = 500$.

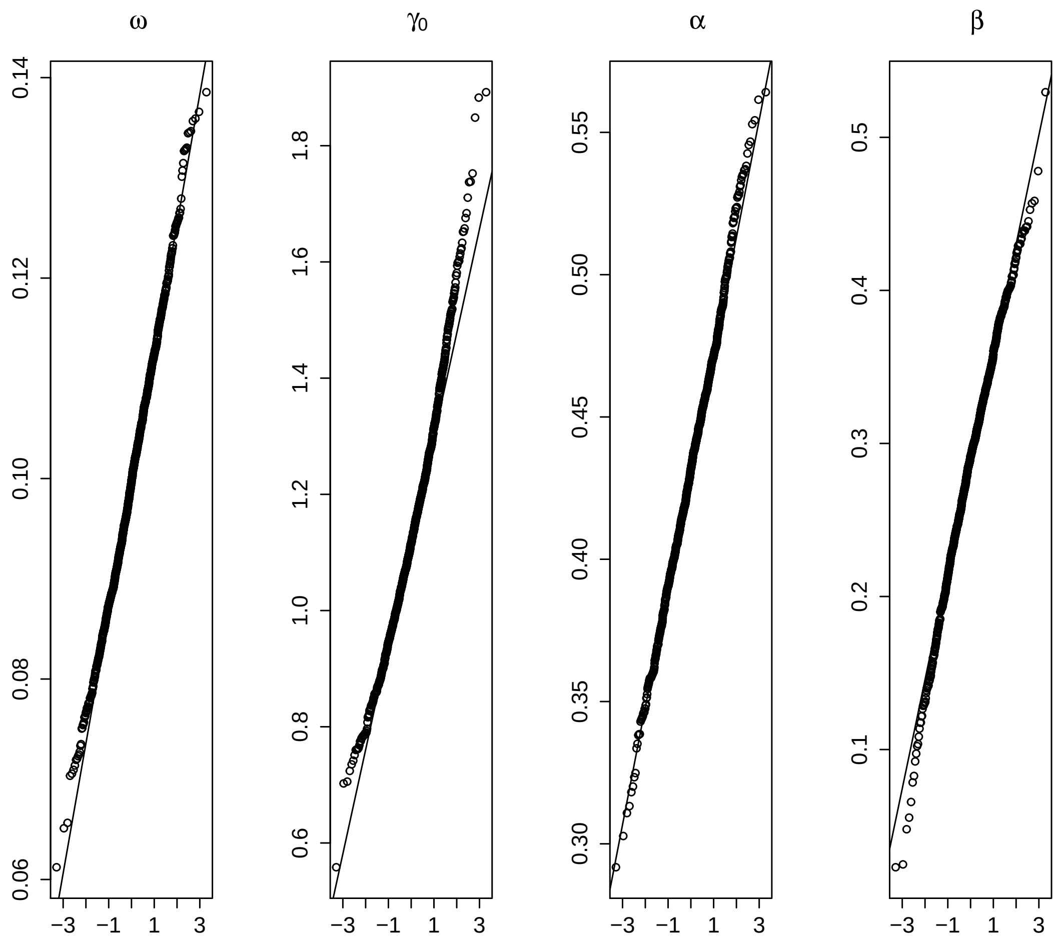

Figure 2.8: A Q-Q plot demonstrating normality of the ZIP estimates for the model parameters $\{\omega = 0.1, \gamma_0 = 1, \alpha_1 = 0.4, \beta_1 = 0.3\}$ based on sample size $n = 1000$.

2.1.16 The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

In a pool of competing models, our interest is to select the best model based on the information criterion. In the literature, the AIC and BIC has been used as a principal tool for model selection. The AIC and BIC measure the quality of a statistical model relative to other competing models. They estimate the information that is lost when a model is used to represent a physical process. This means that in a pool of competing models, we will select the model with the smallest AIC and BIC. It is necessay to emphasize here that, even though AIC and BIC helps us to select the best model based on the information criterion, it does not guarantee an optimum model. Given a statistical model, if $\mathcal{L}$ is the maximized value of the likelihood function and $\boldsymbol{\theta}$ the number of estimated parameters in the model, then we can calculate the AIC value of the model as below.

$$AIC = 2\boldsymbol{\theta} - 2\log(\hat{\mathcal{L}}).$$

The AIC is essentially a function of the likelihood. An important conclusion that can be drawn based on the structure of the formula is that, the AIC value increases as the number of estimated parameters in the model increase. This discourages overfitting as it is almost certain that we will eventually obtain a good fit by increasing the parameters that we have in our model. Below is a formula for the BIC

$$BIC = -2\log(\hat{\mathcal{L}}) + \boldsymbol{\theta}\log(\boldsymbol{n})$$

where n is the number of observations in the data being considered.

2.1.17 Real Data Example - Syphilis Data Analysis

In this subsection we apply the estimation strategy to a real data set. This data is based on public health surveillance for syphilis, a sexually transmitted disease that poses a major health challenge in the United States. Yang (2012) studied this data where she modeled the influence of autoregressive covariates and trend on the observed time series. The data consists of the weekly number of syphilis cases reported in Maryland from 2007-2010.

The series is extracted from the Centre for Disease Control (CDC) morbidity and mortality weekly report. There are 209 observations. The empirical mean and variance of the data are 3.4737 and 9.2794, respectively. The bar chart, ACF, and PACF for the number of syphilis cases are shown in Figure 2.9. This figure shows that a large number of zeros are observed over a total of 209 weeks.

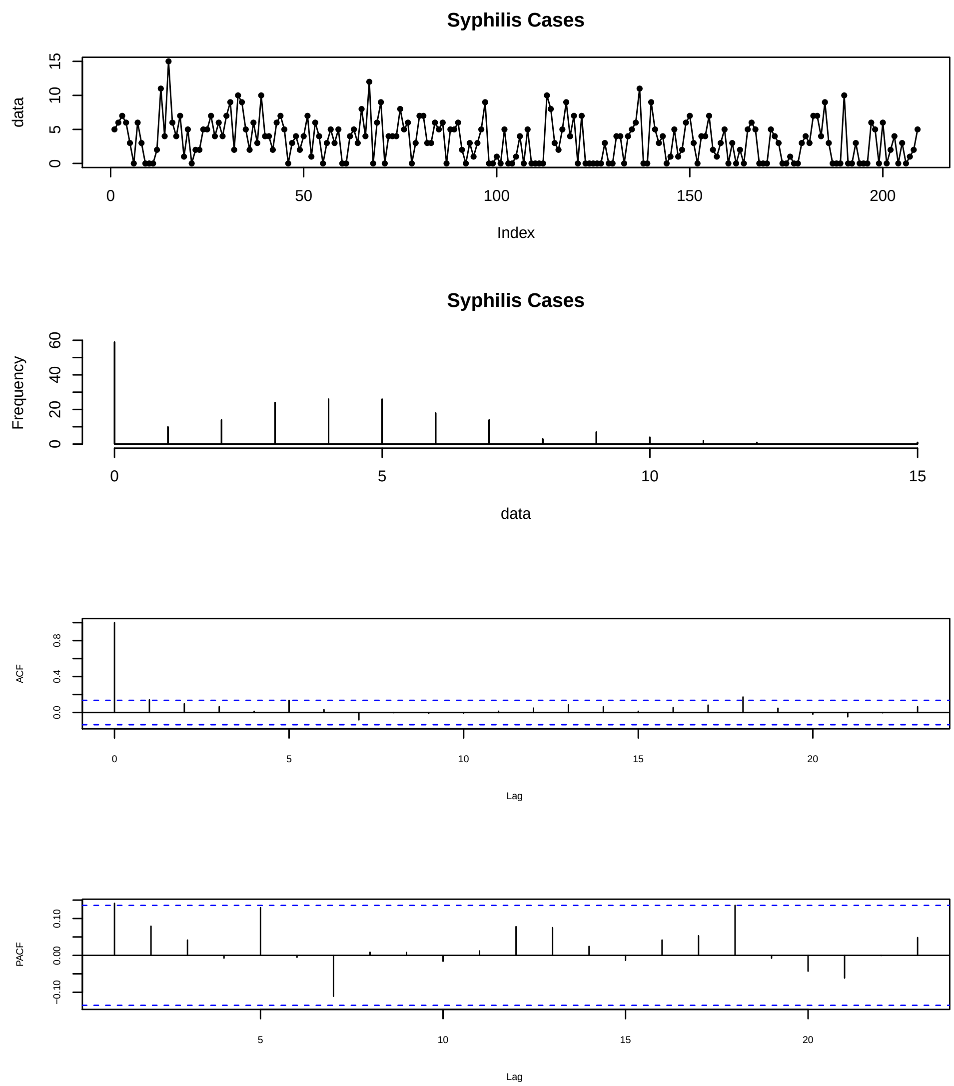

Figure 2.9: Plot of counts, Bar chart, ACF and PACF of syphilis cases

From the bar chart above, the syphilis data is inundated with zeros as the longest bar is for zero counts of syphilis cases. The empirical percentage of zeros in the series is 28.23%.

Table 2.6: Estimated parameters, AIC and BIC for syphilis counts

Model	$\hat{\omega}$	$\hat{\gamma}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\beta}_1$	AIC	BIC
$ZIACP(1,1)$	0.2704	1.2715	0.1613	–	0.6183	868.7936	855.4243
$ZIACP(1)$	0.2714	4.1464	0.1971	–	–	859.2112	849.1842
$ZIACP(2)$	0.2728	3.8094	0.1787	0.1109	–	890.4123	877.043
Poisson	–	2.0347	0.2419	–	0.1722	1156.438	1166.465

Table 2.6 shows that the model $ZIACP(1)$ has the lowest AIC and BIC values. Another criterion that we want to use to assess the adequacy of the modelling approah is by computing the mean and variance of the fitted model (i.e., using the formular for the theoretical mean and variance and the parameter estimates) and comparing to the empirical mean and variance of the data. The Empirical mean and variance of the data are 3.4737 and 9.2794, respectively. For the fitted $ZIACP(1,1)$ model, the mean $\widehat{E(Y_t)}$ based on the estimated parameters is given by

$$\widehat{E(Y_t)} = \hat{\mu} = \frac{(1-\hat{\omega})\hat{\gamma}_0}{1-(1-\hat{\omega})\hat{\alpha}_1-\hat{\beta}_1} = \frac{(1-0.2704)\times 1.2715}{1-(1-0.2704)\times 0.1613-0.6183} = 3.5137$$

This implies that the mean is over estimated by 0.04. Similarly we found the estimate of the variance,

$$\widehat{Var(Y_t)} = \frac{1 - 2(1-\hat{\omega})\hat{\alpha}_1\hat{\beta}_1 - \hat{\beta}_1^2}{1 - (1-\hat{\omega})\hat{\alpha}_1^2 - 2(1-\hat{\omega})\hat{\alpha}_1\hat{\beta}_1 - \hat{\beta}_1^2}\left(\hat{\mu} + \frac{\hat{\omega}\hat{\mu}^2}{1-\hat{\omega}}\right)$$

$$= 8.4282$$

$\widehat{Var(Y_t)}$ underestimates the variance by 0.8512. The Table 2.7 below reports the estimates of the moments from the fitted models.

Table 2.7: Estimates of the moments from the fitted models

Model	$\widehat{E(Y_t)}$	$\widehat{Var(Y_t)}$	$\widehat{E(Y_t)} - E(Y_t)$	$\lvert\widehat{Var(Y_t)} - Var(Y_t)\rvert$
$ZIACP(1,1)$	3.5137	8.4282	0.0400	0.8512
$ZIACP(1)$	3.5277	8.4009	0.0540	0.8785
$ZIACP(2)$	3.5092	8.4345	0.0355	0.8449
Poisson	3.4728	3.5971	-0.0009	5.6823

2.1.18 Analyzing Arson Data

We also consider a time series from the forecasting principles website (Zhu, 2012 also used this data) at http://www.forecastingprinciples.com in the crime data section. The data represents counts of arson cases in Pittsburgh. The data includes time series from January 1990 through December 2001, a total of 144 monthly observations. The empirical mean and variance of the data are 1.0417 and 1.3829, respectively.

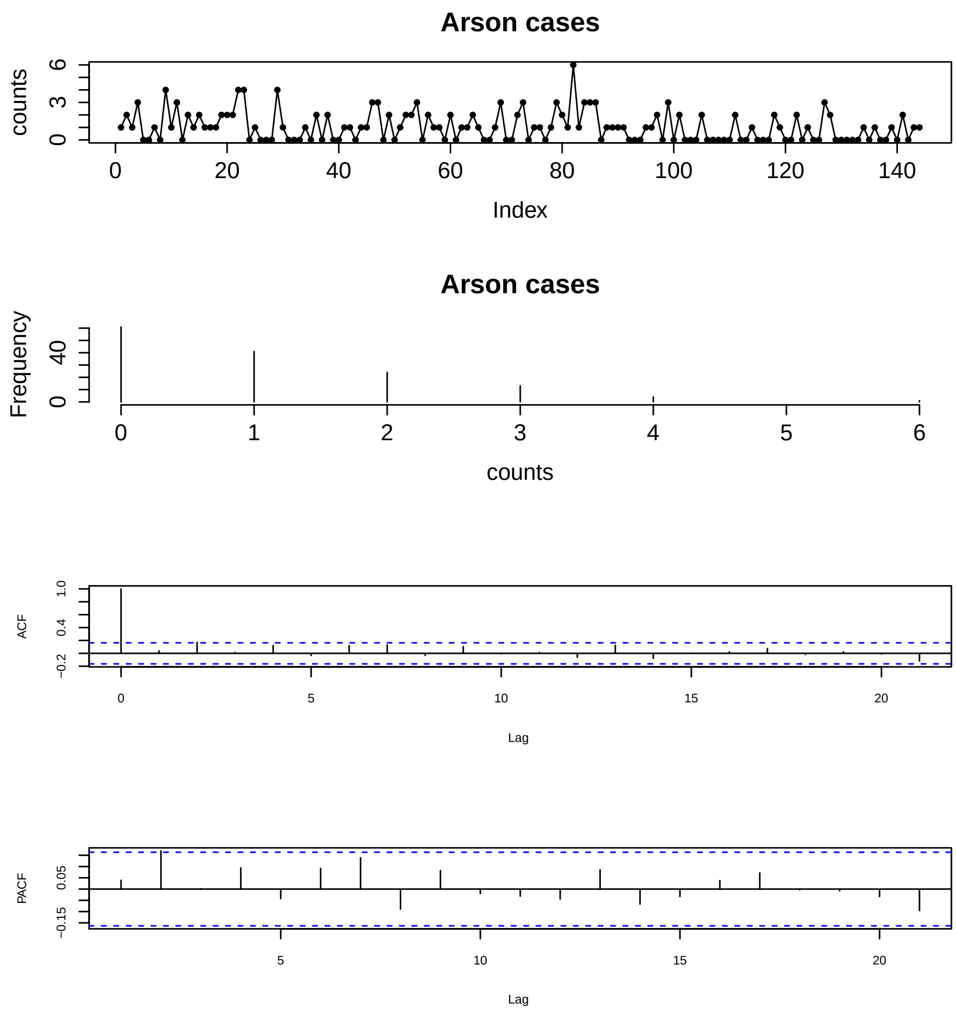

Figure 2.10: Plots of counts, Bar chart, ACF and PACF of arson counts

The bar chart in Figure (2.10) suggests the arson data is zero inflated with 42.36% zeros in the series. The figure also shows the ACF and PACF of the series. The estimated parameters, AIC and BIC of the various fitted ZIACP models are given in the Table 2.8

Table 2.8: ZIACP and Poisson models for arson counts data

Model	$\hat{\omega}$	$\hat{\gamma}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\beta}_1$	AIC	BIC
$ZIACP(1,1)$	0.3638	0.2155	0.1446	$-$	0.8216	394.9082	383.0289
$ZIACP(1)$	0.3417	1.7369	0.0680	$-$	$-$	390.8159	381.9065
$ZIACP(2)$	0.3259	1.4302	0.0568	0.2942	$-$	367.8608	355.9816
Poisson	$-$	0.5763	0.3594	$-$	0.0899	409.5677	418.4771

Based on the AIC, BIC criteria and similar arguments above, the $ZIACP(2)$ fits the arson data better. Our method improves on the AIC and the BIC obtained for all the models considered when compared to the results in Zhu (2012). Even though the same EM algorithm procedure was used by Zhu (2012) as well, perhaps extending the work of Fokianos et al. (2009) and obtaning results based on that made the difference.

Table 2.9: Estimates of the moments of the fitted models

Model	$\widehat{E(Y_t)}$	$\widehat{Var(Y_t)}$	$\widehat{E(Y_t)} - E(Y_t)$	$\lvert\widehat{Var(Y_t)} - Var(Y_t)\rvert$
$ZIACP(1,1)$	1.5867	3.2772	0.5450	1.8943
$ZIACP(1)$	1.1970	1.7798	0.1553	0.3969
$ZIACP(2)$	1.2629	2.1675	0.2212	0.7846
Poisson	1.0465	1.0571	0.0048	0.3258

2.1.19 Polio Data Analysis

The polio data consists of monthly number of incidents of poliomyelitis in the USA during the years 1970-1983. The data have been released by the US Centers for Disease and Control and there are a total of $n = 168$ observations. The empirical mean and variance of the dataset is 1.3333 and 3.5050, respectively.

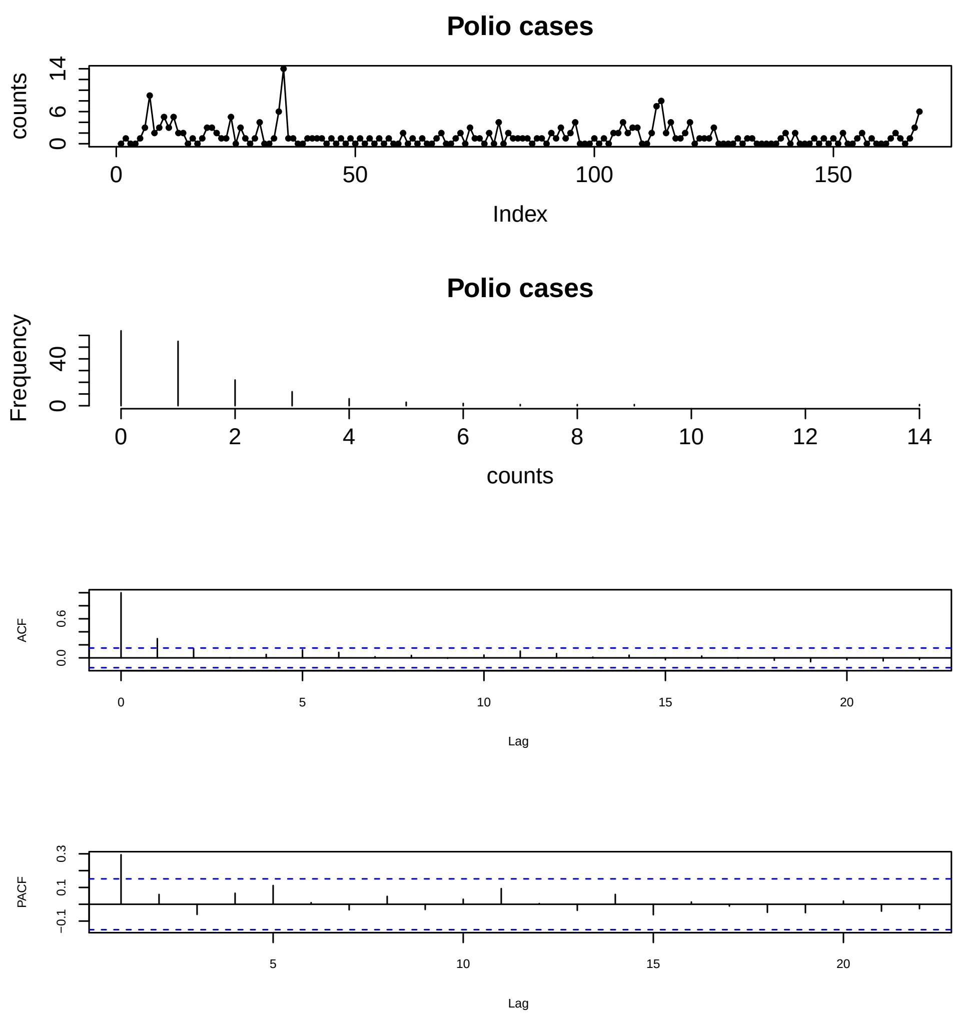

Figure 2.11: Plots of counts, Bar chart, ACF and PACF of Polio cases

Table 2.10: ZIACP and Poisson models for Polio counts data

Model	$\hat{\omega}$	$\hat{\gamma}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\beta}_1$	AIC	BIC
$ZIACP(1,1)$	0.2817	1.2174	0.4720	–	0.1551	463.7097	451.2138
$ZIACP(1)$	0.2811	1.5208	0.4948	–	–	465.7242	456.3523
$ZIACP(2)$	0.3026	1.4289	0.4657	0.1000	–	529.7672	517.2714
Poisson	–	0.6400	0.1837	–	0.3801	562.1346	571.5065

Table 2.11: Estimates of the moments of the fitted models

Model	$\widehat{E(Y_t)}$	$\widehat{Var(Y_t)}$	$\widehat{E(Y_t)} - E(Y_t)$	$\lvert\widehat{Var(Y_t)} - Var(Y_t)\rvert$
$ZIACP(1,1)$	1.7286	3.5535	0.3953	0.0485
$ZIACP(1)$	1.6969	3.4258	0.3636	0.0792
$ZIACP(2)$	1.6458	3.4441	0.3125	0.0609
Poisson	1.4672	1.7780	0.1339	1.7270

Table (2.10) gives a summary of the fitted model for the polio dataset. Based on the AIC and BIC criteria, the $ZIACP(1,1)$ provides the best fit. Kitromilidou and Fokianos (2015) considered a class of loglinear autoregressive model for the polio data. The AIC of their best fit model was 490.965. In that paper, the model (2.24) was chosen to provide the best fit for the data . Below is the model

$$\log \lambda_t = d + \sum_{j=1}^{5} b_j \log(1 + Y_{t-j}) + \beta t/n + \sum_{s=1}^{2} \{\beta_{1;s} sin(\omega_s t) + \beta_{2;s} cos(\omega_s t)\} \quad (2.24)$$

The data analysis of their paper was done without paying any attention to the fact that the polio data is zero inflated. This problem only suggests that analysing the polio dataset from the point of view as a zero inflated data is relevant and provides a better fit than when considered otherwise.

2.1.20 Exploring the ZIM and the pscl package in R

In this example we again model the number of syphilis counts as a function of the first lagged value of the response and a trend component i.e., $zim(counts \sim bshift(counts > 0)$ + $trend|trend)$. Here trend is defined by dividing each time point by 1000. The trend induces extra zeros in the model. The results in Table 2.12 is based on the model where we consider trend and lagged response as covariates .

Table 2.12: Estimates of parameters of the loglinear model of the ZIP autoregression

Parameter	Estimate(**ZIM**)	Estimate(**pscl**)	SE(**ZIM**)	SE(**pscl**)
Intercept	1.48942	1.4894	0.11995	0.1199
Lag autoregression	0.22111	0.2211	0.10072	0.1007
Trend	-1.01004	-1.0100	0.66687	0.6669

In Table 2.12, the associated $AIC = 918.7806$ and $BIC = 935.4683$ when using the **ZIM** package and $AIC = 916.7806$, $BIC = 930.1499$ for the **zeroinfl()** in the **pscl** package. p-value for the Intercept, lag autoregression and trend are, respectively $< 2e-16$, 0.02813 and 0.12987.

Table 2.13: Estimates of parameters of the logistic model of the ZIP autoregression

Parameter	Estimate(**ZIM**)	Estimate(**pscl**)	SE(ZIM)	SE(**pscl**)
Intercept	-1.93321	-1.933	0.37196	0.372
Trend	8.60517	8.604	2.80827	2.808

In Table 2.13 the logistic model has $AIC = 918.7806$ and $BIC = 935.4683$ when using the **ZIM** package and $AIC = 916.7806$, $BIC = 930.1499$ for the **zeroinfl()** in the **pscl** package. p-value for the Intercept and trend are, respectively $< 2.021e - 07$ and

0.002182. The two Tables (2.12) and (2.13) compare the estimates of the autoregressive parameter obtained by using the **R**-packages **ZIM** and **pscl**. The two tables also report the results when trend is considered as covariate. Even though the trend component is significant, the AIC and BIC that comes with this result is still higher than that obtained in the ZIACP modeling approach for the same real data set in Section 2.1.17.

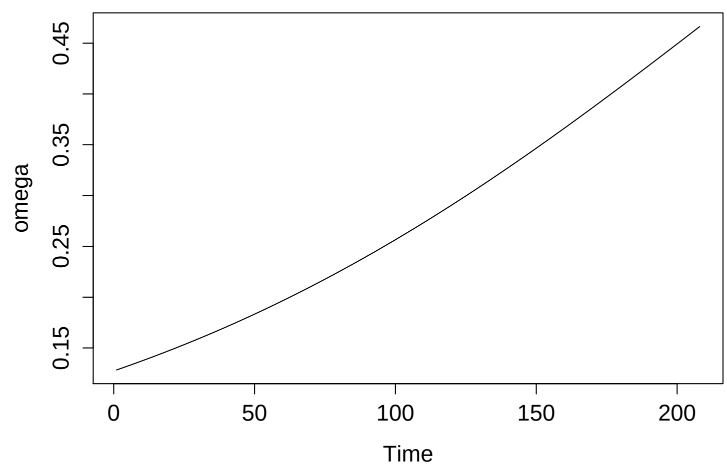

Figure 2.12: Plot of the zero inflation parameter ω_t over time for the ZIP autoregression.

Based on the AIC and BIC criteria, the ZIACP models considered in Table 2.6 are superior to the ZIP autoregressive model even though the trend component in the model is significant. That notwithstanding, the ZIP autoregressive model ensures no parameter restrictions, in the loglinear part of the model, we model the log of the intensity parameter. However this is not true for the ZIACP models where the parameter space has to be restricted in order to ensure that the intensity parameter is positive. In the autoregressive model, the zero inflation parameter is defined as a time varying parameter ω_t, a plot of ω_t is shown in Figure 2.12 . The mean of ω_t is 0.2756 which is very close to $\hat{\omega}$ obtained from the ZIACP models.

Table 2.14: Estimates of parameters of the loglinear model of the ZINB autoregression

Parameter	Estimate(**ZIM**)	Estimate(**pscl**)	SE(**ZIM**)	SE(**pscl**)
Intercept	1.47240	1.4725	0.13873	0.1387
Lag autoregression	0.23164	0.2316	0.11522	0.1152
Trend	-1.00364	-1.0038	0.77154	0.7714

In Table 2.14, the $AIC = 915.4927$ and $BIC = 935.5179$ when the **ZIM** package is used and $AIC = 911.4928$, $BIC = 924.8621$ when the **zeroinfl()** in the **pscl** package is used. p-value for the Intercept, lag autoregression and trend are, respectively $< 2e - 16$, 0.04438 and 0.19332.

Table 2.15: Estimates of the parameters of the Logistic Model of the ZINB autoregression

Parameter	Estimate(**ZIM**)	Estimate(**pscl**)	SE(**ZIM**)	SE(**pscl**)
Intercept	-1.97940	-1.976	0.38563	0.385
Trend	8.71684	8.690	2.88697	2.885

The associated AIC and BIC for the logistic model in Table 2.15 when the **ZIM** package is used is 915.4927 and 935.5179, respectively and $AIC = 911.4928$, $BIC = 924.8621$ when **zeroinfl()** in the **pscl** package is used. p-value for the Intercept and trend are, respectively $2.853e - 07$ and 0.002533.

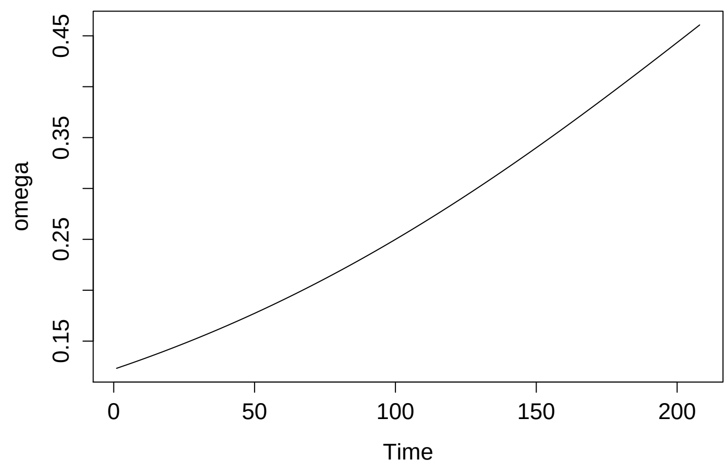

Figure 2.13: Plot of the zero inflation parameter ω_t over time for ZINB autoregression

In the case of ZINB autoregressive model, again the ZIACP models are superior. However, based on the AIC and BIC criteria, we found that the ZINB autoregressive model to be slightly better than the ZIP for the syphilis data. The ZIACP model particularly the ZIACP(1,1) provides a better fit than ZIP and ZINB autoregression models. Again, the mean of ω_t is 0.2694.

Table 2.16: Autoregressive model for syphilis data via **glm** function in the **pscl** package

Model	AIC	BIC
Poisson	1142.6174	1149.3021
Negative binomial	985.3994	992.0841

In Table (2.16) we fit the Poisson and the Negative binomial using the syphilis data via the **glm** function in the **pscl** package. Below is the **R** code:

```
fm_pois  <- glm(syph$a33 ~ bshift(syph$a33>0) ,
data = counts1, family = poisson)
summary(fm_pois)
fm_pois
fm_nb    <- glm.nb(syph$a33 ~ bshift(syph$a33>0) ,
data = counts1)
summary(fm_nb)
fm_nb
```

From Table 2.16 we see that the AIC and BIC of the negative binomial is very close to the zero inflated distributions considered in table 2.12. This is in line with the claim made in Allison (2012) which suggests that for most datasets the negative binomial is as good as the zero inflated distributions.

Chapter 3

The Zero Inflated Poisson Autoregression with Covariates (ZIPA) Model

3.1 Introduction

Unitl now, we have considered the ZIP model without incorporating the effects of covariates. In chapter 3, we consider the ZIP model with covariates and construct pretest, James-Stein shrinkage and positive shrinkage estimators for the vector of parameters. The shrinkage strategy is a method that allows the researcher to improve estimation strategies since it uses information from the insignificant covariates for estimating the coefficients of the significant covariates. For more details, see Ahmed et al. (2012), Thomson et al. (2014), and Hossain et al. (2009). In this situation, we may partition the regression parameter vector $\boldsymbol{\theta}$ into two sub-vectors as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are assumed to have dimensions $k_1 \times 1$ and $k_2 \times 1$, respectively, such that $k = k_1 + k_2$. We postulate a restriction, $\boldsymbol{\theta}_2 = \mathbf{0}$ which incorporates a variety of prior non-sample information about the parameters. This situation occurs frequently when there is over-modelling and one wishes to remove the irrelevant part of the model, which in turn will increase the efficiency of estimates of $\boldsymbol{\theta}$. On the other hand, when there is no prior non-sample information, one could apply to model selection strategies and identify some of the coefficients being practically zero. In this latter situation, as the final model need not be the true model, it is still safer to resort to estimation methods which take into account the restriction induced by the model selection criteria.

Based on the restriction, we build restricted MLE estimator for the restricted model. In the context of two models where one includes all covariates and the other includes a restriction. We optimally combine the estimates from the unrestricted and restricted models to define shrinkage estimator. The pretest estimator can be obtained from either the unrestricted model containing all coefficients or the restricted model stated by the null hypothesis. We investigate the relative performances of shrinkage and pretest estimators with respect to the unrestricted maximum likelihood estimator (UMLE). The asymptotic properties of these estimators including the derivation of asymptotic distributional biases and risks are established. A Monte Carlo simulation study is then undertaken in order to compare the performance of these estimators with respect to UMLE.

3.1.1 Models and the proposed estimators

Consider the ZIP autoregression model by the adding covariates in the model of the intensity parameter λ_t and the zero inflation parameter ω_t i.e.,

$$\eta_t = \log \lambda_t = \boldsymbol{x}_{t-1}^T \boldsymbol{\beta} \tag{3.1}$$

and

$$\xi_t = \textbf{logit}(\omega_t) = \boldsymbol{z}_{t-1}^T \boldsymbol{\gamma}, \tag{3.2}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_q)^T$ are the regression coefficients for the log-linear (3.1) and the logistic parts (3.2), respectively. For convenience, we let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ denote the $k = (p+q)$-dimensional vector of unknown parameters.

In models (3.1) and (3.2) we are often interested in testing the following hypothesis:

$$H_0 : \boldsymbol{\theta}_2 = \mathbf{0} \quad \text{versus} \quad H_A : \boldsymbol{\theta}_2 \neq \mathbf{0}.$$

For a random sample, $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$, the log-likelihood function is given by

$$\begin{aligned}
\ell(\boldsymbol{\lambda}, \boldsymbol{\omega}; \boldsymbol{y}) &= \sum_{t=1}^{n} \big\{ I_{(y_t=0)} \log[\omega_t + (1-\omega_t) exp(-\lambda_t)] \\
&+ I_{(y_t>0)} [\log(1-\omega_t) - \lambda_t + y_t \log \lambda_t - \log(y_t!)] \big\} \\
&= \sum_{t=1}^{n} \big\{ I_{(y_t=0)} \log[\exp(\boldsymbol{x}_{t-1}^{\top}\boldsymbol{\gamma}) + \exp(-\exp(\boldsymbol{x}_{t-1}^{\top}\boldsymbol{\beta}))] \\
&+ I_{(y_t>0)} [y_t \boldsymbol{x}_{t-1}^{\top}\boldsymbol{\beta} - \exp(\boldsymbol{x}_{t-1}^{\top}\boldsymbol{\beta}) - \log(y_t!)] \big\} \qquad (3.3) \\
&- \sum_{t=1}^{n} \log(1 + \exp(\boldsymbol{z}_{t-1}^{\top}\boldsymbol{\gamma})),
\end{aligned}$$

where $I_{(\cdot)}$ is an indicator function, which is equal to 1 if the event is true and 0 otherwise. The log-likelihood function (3.4) of the ZIPA model is quite complicated, especially as the first term involves both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Also, the responses are from a mixture distribution that includes both sets of the parameters ω_t and λ_t. The computation thus becomes quite challenging in terms of variance-covariance and accuracy when using the Newton-Raphson algorithm. To avoid this complication, we use the EM algorithm to maximize the log-likelihood function, see, Hall (2000) and Lambert (1992).

The EM algorithm is based on a latent variable U_t (we use this new notation instead of Z_t for the missing values in order not to cause conflict with the covariates z_t in the logistic model and the Chapter 2). We could observe U_t = 1, when Y_t is from the perfect zero state (or first process) and $U_t = 0$, when Y_t is from the Poisson state (or second process). To formulate the log-likelihood for the complete data, we use the conditional probability:

$$\begin{aligned}
& Pr\left(Y_t = y_t, U_t = u_t | \boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1}, \boldsymbol{\beta}, \boldsymbol{\gamma}\right) \\
= \;& Pr\left(Y_t = y_t | U_t = u_t, \boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1}, \boldsymbol{\beta}, \boldsymbol{\gamma}\right) \times Pr\left(U_t = u_t | \boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1}, \boldsymbol{\beta}, \boldsymbol{\gamma}\right) \\
= \;& \left(\frac{\exp(\boldsymbol{z}_{\boldsymbol{t-1}}^{\top}\boldsymbol{\gamma})}{1+\exp(\boldsymbol{z}_{t-1}^{\top}\boldsymbol{\gamma})}\right)^{u_t} \\
& \times \left(\frac{1}{(1+\exp(\boldsymbol{z}_{t-1}^{\top}\boldsymbol{\gamma}))} . \frac{\exp(-\exp(y_t\boldsymbol{x}_{t-1}^{\top}\boldsymbol{\beta}))\exp(\boldsymbol{x}_{t-1}^{\top}\boldsymbol{\beta}))}{y_t!}\right)^{1-u_t} .
\end{aligned}$$

Thus, the complete log-likelihood based on $(\boldsymbol{Y}, \boldsymbol{U})$ is

$$\begin{aligned}
\ell_c(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{U}, \boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1}) &= \log\left[\prod_{t=1}^{n} Pr\left(Y_t = y_t, U_t = u_t | \boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1}, \boldsymbol{\beta}, \boldsymbol{\gamma}\right)\right] \\
&= \sum_{t=1}^{n}\{u_t\boldsymbol{z}_{t-1}^{\top}\boldsymbol{\gamma} - \log(1+\exp(\boldsymbol{z}_{t-1}^{\top}\boldsymbol{\gamma}))\} \\
&+ \sum_{t=1}^{n}(1-u_t)(y_t\boldsymbol{x}_{t-1}^{\top}\boldsymbol{\beta} - \exp(\boldsymbol{x}_{t-1}^{\top}\boldsymbol{\beta}) - \log(y_i!)) \quad (3.4) \\
&= \ell_{c1}(\boldsymbol{\gamma}; \boldsymbol{U}, \boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1}) + \ell_{c2}(\boldsymbol{\beta}; \boldsymbol{U}, \boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1}) \\
&- \sum_{t=1}^{n}(1-u_t)\log(y_t!)),
\end{aligned}$$

where $\ell_{c1} = \sum_{t=1}^{n}\{u_t\boldsymbol{z}_{t-1}^{\top}\boldsymbol{\gamma} - \log(1+\exp(\boldsymbol{z}_{t-1}^{\top}\boldsymbol{\gamma}))\}$, $\ell_{c2} = \sum_{t=1}^{n}(1-u_t)(y_t\boldsymbol{x}_{t-1}^{\top}\boldsymbol{\beta} - \exp(\boldsymbol{x}_{t-1}^{\top}\boldsymbol{\beta}))$, and $U = \{u_t; t = 1, 2, \cdots, n\}$.

To implement the EM algorithm, we first initialize $(\boldsymbol{\beta}, \boldsymbol{\gamma})$. In the E-step, we use the initial values of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ to calculate the expectation of U_t and use it as an estimate of U_t. In the M-step, we use the estimate of U_t to maximize $\ell_c(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{U}, \boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1})$, which gives the unrestricted maximum likelihood estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The iteration l of the EM algorithm requires the following steps.

E-Step: Estimate $U_t^{(l)}$ by using the means given $\boldsymbol{\gamma}^{(l)}$ and $\boldsymbol{\beta}^{(l)}$,

$$\begin{aligned}
U_t^{(l)} &= E(U_t|y_t, \gamma^{(l)}, \boldsymbol{\beta}^{(l)}) \\
&= E(U_t = 1|y_t, \gamma^{(l)}, \boldsymbol{\beta}^{(l)}) \\
&= \frac{Pr(Y_t = y_t|U_t = 1)Pr(U_t = 1)}{Pr(Y_t = y_t|U_t = 1)Pr(U_t = 1) + Pr(Y_t = y_t|U_t = 0)Pr(U_t = 0)} \\
&= \begin{cases} \frac{\omega_t}{\omega_t + (1-\omega_t)exp(-\lambda_t)}, & y_t = 0 \\ 0, & y_t \geq 1, \end{cases} \\
&= \begin{cases} \left[1 + exp(-exp(\boldsymbol{x}_{t-1}^\top \boldsymbol{\beta}^{(l)}) - \boldsymbol{z}_{t-1}^\top \boldsymbol{\gamma}^{(l)})\right]^{-1}, & y_t = 0 \\ 0, & y_t \geq 1, \end{cases}
\end{aligned}$$

M-Step: Given $U_t = U_t^{(l)}$, maximize $\ell_{c1}(\boldsymbol{\gamma}; \boldsymbol{U}, \boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1})$ and $\ell_{c2}(\boldsymbol{\beta}; \boldsymbol{U}, \boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1})$ with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, respectively:

$$\begin{aligned}
\boldsymbol{\gamma}^{(l+1)} &= \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \{-\ell_{c1}(\boldsymbol{\gamma}; \boldsymbol{U}, \boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1})\} \\
\boldsymbol{\beta}^{(l+1)} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \{-\ell_{c2}(\boldsymbol{\beta}; \boldsymbol{U}, \boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1})\}
\end{aligned}$$

The iteration stops when $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ converges, and the final estimate is denoted as $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\gamma}}^\top)^\top$, the unrestricted maximum likelihood estimator (UMLE).

If the information matrix $\frac{1}{n}I_{(\boldsymbol{\beta},\boldsymbol{\gamma})}$ has a positive definite limit satisfying some regularity conditions, as in the work of McCullagh (1984), the quantity $\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)$ is asymptotically normally distributed with mean vector $\mathbf{0}$ and information matrix $I^{-1}_{(\boldsymbol{\beta},\boldsymbol{\gamma})}$ (Lambert, 1992). The matrix $I_{(\boldsymbol{\beta},\boldsymbol{\gamma})}$ can be partitioned as

$$\begin{pmatrix} I_{\boldsymbol{\beta},\boldsymbol{\beta}} & I_{\boldsymbol{\beta},\boldsymbol{\gamma}} \\ I_{\boldsymbol{\gamma},\boldsymbol{\beta}} & I_{\boldsymbol{\gamma},\boldsymbol{\gamma}} \end{pmatrix}$$

where the elements $I_{\boldsymbol{\beta},\boldsymbol{\beta}}$, $I^{\top}_{\boldsymbol{\beta},\boldsymbol{\gamma}} = I_{\boldsymbol{\gamma},\boldsymbol{\beta}}$, and $I_{\boldsymbol{\gamma},\boldsymbol{\gamma}}$ are, respectively,

$$-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\top}}\right], -\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}}\right], \text{ and } -\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^{\top}}\right]$$

with

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\top}} &= \sum_{t=1}^{n} \lambda_t \left\{1 - \frac{y_{0,t}\omega_t(\omega_t + (1-\omega_t)(1+\lambda_t)exp(-\lambda_t))}{p_{0,t}^2}\right\} x_{t-1}^{\top} x_{t-1}, \\
\frac{\partial^2 \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^{\top}} &= \sum_{t=1}^{n} \omega_t(1-\omega_t) \left\{1 - \frac{y_{0,t}exp(\lambda_t)}{p_{0,t}^2}\right\} z_{t-1}^{\top} z_{t-1}, \\
\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}} &= \frac{\partial^2 \ell}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}} = \sum_{t=1}^{n} \frac{-y_{0,t}\omega_t(1-\omega_t)\lambda_t exp(-\lambda_t)}{p_{0,t}^2} x_{t-1}^{\top} z_{t-1},
\end{aligned}$$

where $p_{0,t} = \omega_t + (1-\omega_t)exp(-\lambda_t)$ is the probability mass function of $Y_t|\mathcal{F}_{t-1}$ at zero and $\mathbf{1}_{\{y_t=0\}} = y_{0,t}$.

Suppose now that our interest is in estimating the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ from (3.4) under the restriction $\boldsymbol{\theta}_2 = \mathbf{0}$. The steps of the EM-algorithm for estimating the parameters using log-likelihood (3.4) under the above restriction are similar. The resulting estimator, $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}^{\top}, \tilde{\boldsymbol{\gamma}}^{\top})^{\top}$ is called the restricted maximum likelihood estimator (RMLE).

Theorem 13. [*Yang* (2012)] *For the ZIPA model defined above, the score function* $S_n(\boldsymbol{\theta})$ *is given by*

$$S_n(\boldsymbol{\theta}) = \frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \boldsymbol{C}_{t-1} v_t(\boldsymbol{\theta})$$

with $\boldsymbol{C}_t$ *and* $v_t(\boldsymbol{\theta})$ *defined as*

$$\boldsymbol{C}_{t-1} = \begin{bmatrix} \boldsymbol{x}_{t-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{z}_{t-1} \end{bmatrix}$$

and

$$v_t(\boldsymbol{\theta}) = \begin{bmatrix} v_{1,t}(\boldsymbol{\theta}) \\ v_{2,t}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} y_t - \lambda_t(1 - \frac{\omega_t y_{0,t}}{p_{0,t}}) \\ \omega_t(\frac{y_{0,t}}{p_{0,t}} - 1) \end{bmatrix}$$

where $y_{0,t} = \mathbf{1}_{\{y_t=0\}}$.

Theorem 14. [*Yang* (2012)]

The conditional information matrix of the ZIPA is given by,

$$G_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} Var\{\boldsymbol{C}_{t-1} v_t(\boldsymbol{\theta}) | \mathcal{F}_t\} = \sum_{t=1}^{n} \boldsymbol{C}_{t-1} \Sigma_t(\boldsymbol{\theta}) \boldsymbol{C}_{t-1}^T,$$

where $\Sigma_t(\boldsymbol{\theta}) = Var\{v_t(\boldsymbol{\theta}) | \mathcal{F}_{t-1}\}$ *is a symmetric* 2×2 *matrix with the elements.*

$$\Sigma_t(\boldsymbol{\theta}) = \begin{bmatrix} Var(v_{1,t}(\boldsymbol{\theta})) & Cov(v_{1,t}(\boldsymbol{\theta}), v_{2,t}(\boldsymbol{\theta})) \\ Cov(v_{2,t}(\boldsymbol{\theta}), v_{1,t}(\boldsymbol{\theta})) & Var(v_{1,t}(\boldsymbol{\theta})) \end{bmatrix}.$$

where

$$Var(v_{1,t}(\boldsymbol{\theta})) = \frac{(1-\omega_t)\lambda_t\{exp(-\lambda_t) + \omega_t(1-(1+\lambda_t)exp(-\lambda_t))\}}{p_{0,t}}.$$

$$Var(v_{2,t}(\boldsymbol{\theta})) = \frac{\omega_t^2(1-\omega_t)(1-exp(-\lambda_t))}{p_{0,t}}.$$

$$Cov(v_{1,t}(\boldsymbol{\theta}), v_{2,t}(\boldsymbol{\theta})) = \frac{-\lambda_t\omega_t(1-\omega_t)exp(-\lambda_t)}{p_{0,t}}.$$

A detailed proof is provided in the Appendix. For convenience, we can partition the information matrix as

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} \end{bmatrix},$$

where $\boldsymbol{G}_{ij}$ are positive-definite matrices when $i = j$.

The likelihood ratio test statistic will be used to test $H_0 : \boldsymbol{\theta}_2 = \mathbf{0}$. If $\tilde{\boldsymbol{\theta}}$ maximizes the log likelihood of the ZIPA model under H_0 of dimension $k - k_2$ and $\hat{\boldsymbol{\theta}}$ maximizes the log likelihood of the ZIPA model under a alternative hypothesis H_A of dimension k, then the test statistic D_n is

$$\begin{aligned} D_n &= 2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})] \\ &= n\hat{\boldsymbol{\theta}}_2^\top \boldsymbol{G}_{22.1}\hat{\boldsymbol{\theta}}_2. \end{aligned} \tag{3.5}$$

where $\boldsymbol{G}_{22.1} = \boldsymbol{G}_{22} - \boldsymbol{G}_{21}\boldsymbol{G}_{11}^{-1}\boldsymbol{G}_{12}$. Under the regularity conditions in Appendix and if $\boldsymbol{G}_n(\boldsymbol{\theta})$ is consistently estimated by $\boldsymbol{G}_n(\hat{\boldsymbol{\theta}})$, then

$$\hat{D}_n = n\hat{\boldsymbol{\theta}}_2^\top \hat{\boldsymbol{G}}_{22.1}\hat{\boldsymbol{\theta}}_2 + o_P(1),$$

is asymptotically χ^2-distributed with k_2 degrees of freedom when the null hypothesis $H_0 : \boldsymbol{\theta}_2 = \mathbf{0}$ is true. (Lambert, 1992).

3.1.2 The Pretest and Shrinkage Estimators

The pretest and shrinkage estimators are based on the test statistic D_n of (3.5) for testing $H_0 : \boldsymbol{\theta}_2 = \mathbf{0}$. Specifically, the pretest estimator (PT) of $\boldsymbol{\theta}$ is defined as

$$\hat{\boldsymbol{\theta}}_{PT} = \hat{\boldsymbol{\theta}} - (\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) I(\hat{D}_n \leq \chi^2_{k_2,\alpha}),$$

where $I(A)$ is an indicator function of a set A, and $\chi^2_{k_2,\alpha}$ is the α-level critical value of the approximate distribution of $\hat{D}_n$ under H_0. From the above definition, one can see that if the data yield $\hat{D}_n < \chi^2_{k_2,\alpha}$, then $\hat{\boldsymbol{\theta}}_{PT} = \tilde{\boldsymbol{\theta}}$, otherwise $\hat{\boldsymbol{\theta}}_{PT} = \hat{\boldsymbol{\theta}}$. So the PT is indeed a simple mixture of the UMLE and RMLE. In an ordinary two-step procedure, one would test the hypothesis $H_0 : \boldsymbol{\theta}_2 = \mathbf{0}$ first, then based on the test result decide which estimator should be adopted. The PT simply combines these two steps to form a single one. That is, testing and estimation are done simultaneously. It is important to note here that $\hat{\boldsymbol{\theta}}_{PT}$ performs better than $\hat{\boldsymbol{\theta}}$ in some important parts of the parameter space. For details, see Hossain et al. (2009) and Ahmed et al. (2006).

Because of extreme choices for either the UMLE or RMLE, the pretest procedures are not admissible for many models, even though they may improve upon UMLE, a well-documented fact in the literature (Judge and Bock, 1978). In view of this limitation we define a shrinkage estimator, which is a smoothed version of $\hat{\boldsymbol{\theta}}_{PT}$:

$$\hat{\boldsymbol{\theta}}_{SE} = \tilde{\boldsymbol{\theta}} + \left(1 - (k_2 - 2)\hat{D}_n^{-1}\right)(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}), \quad k_2 \geq 3. \tag{3.6}$$

This estimator is a weighted average of UMLE $\hat{\boldsymbol{\theta}}$ and RMLE $\tilde{\boldsymbol{\theta}}$, where the weights are a function of the test statistic for testing $H_0 : \boldsymbol{\theta}_2 = \mathbf{0}$.

We note that when the test statistic $\hat{D}_n$ is very small in comparison with $k_2 - 2$, i.e., when the ratio $(k_2 - 2)/\hat{D}_n$ is greater than one in absolute value, the shrinkage estimator $\hat{\boldsymbol{\theta}}_{SE}$ tends to shrink $\hat{\boldsymbol{\theta}}$ overly towards $\tilde{\boldsymbol{\theta}}$ and reversing the sign of $\hat{\boldsymbol{\theta}}$. By replacing $\left(1 - (k_2 - 2)\hat{D}_n^{-1}\right)$ by $\left(1 - (k_2 - 2)\hat{D}_n^{-1}\right)_+$ in (3.6), where $(x)_+ = x1_{(x \geq 0)}$, the positive-part shrinkage estimator, $\hat{\boldsymbol{\theta}}_{PSE}$ rectifies this problem. For details, see, Ahmed et al. (2012).

3.2 Asymptotic Results

In this section, we consider the asymptotic behavior of an estimator $\hat{\boldsymbol{\theta}}^*$, which could be any one of the five estimators considered in this paper: $\hat{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}_{PT}$, $\hat{\boldsymbol{\theta}}_{SE}$, and $\hat{\boldsymbol{\theta}}_{PSE}$. The main concern here is to evaluate the performance of these estimators when $\boldsymbol{\theta}_2$ is close to the null vector, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$. To derive any meaningful results we consider a sequence of local alternatives

$$K_n : \boldsymbol{\theta}_2 = \frac{\boldsymbol{\omega}}{\sqrt{n}}, \tag{3.7}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, \cdots, \omega_{k_2})^\top \in \Re^{k_2}$ is a given vector of real numbers. In this framework, $\boldsymbol{\theta}$= $(\boldsymbol{\theta}_1^\top, \mathbf{0}^\top)^\top$, and the quantity $\frac{\boldsymbol{\omega}}{\sqrt{n}}$ is the magnitude of the distance between the unrestricted model and the restricted model. For any fixed $\boldsymbol{\omega}$, this distance shrinks as the sample size increases.

To study the asymptotic distribution risks (ADR) of the estimators, we define a quadratic loss function by using a positive definite matrix $\boldsymbol{W}$, namely

$$\mathcal{L}(\hat{\boldsymbol{\theta}}^*; \boldsymbol{W}) = \left[\sqrt{n}(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta})\right]^\top \boldsymbol{W} \left[\sqrt{n}(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta})\right],$$

where $\hat{\boldsymbol{\theta}}^*$ is any one of the five estimators. The usual quadratic loss is defined when $\boldsymbol{W}$ is chosen as $\boldsymbol{I}$, the identity matrix. A general $\boldsymbol{W}$ gives a loss function that weighs differently for different $\boldsymbol{\theta}$'s.

We assume that the cumulative distribution function of $\hat{\boldsymbol{\theta}}^*$ under K_n exists and can be denoted as

$$F(\boldsymbol{x}) = \lim_{n\to\infty} P\left[\sqrt{n}(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}) \leq \boldsymbol{x} | K_n\right],$$

where $F(\boldsymbol{x})$ is nondegenerate. The ADR of $\hat{\boldsymbol{\theta}}^*$ is then defined as

$$\begin{aligned} ADR(\hat{\boldsymbol{\theta}}^*; \boldsymbol{W}) &= \int \cdots \int \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \, dF(\boldsymbol{x}) \\ &= \text{trace}(\boldsymbol{W}\boldsymbol{V}), \end{aligned} \tag{3.8}$$

where $\boldsymbol{V} = \int \cdots \int \boldsymbol{x}\boldsymbol{x}^\top dF(\boldsymbol{x})$ is the dispersion matrix for the distribution function $F(\boldsymbol{x})$.

The shrinkage estimators are, in general biased, the bias, however is accompanied by a reduction in variance. The asymptotic distributional bias (ADB) of an estimator $\hat{\boldsymbol{\theta}}^*$ is defined as

$$\text{ADB}(\boldsymbol{\theta}^*) = E\left\{\lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta})\right\}.$$

Under the local alternatives (3.7), the following theorems help the derivation and numerical computation of the ADB and the ADR of the estimators.

Theorem 15. *If $\boldsymbol{I}_{\beta,\gamma}$ is nonsingular and $\Delta = \boldsymbol{\omega}^\top \boldsymbol{G}_{22.1}\boldsymbol{\omega}$, then under the local alternatives K_n in (3.7) and regularity conditions in Appendix (see Thomson et al., 2014), we have as $n \to \infty$,*

1. $\sqrt{n}\boldsymbol{\theta}_2 \xrightarrow{\mathcal{L}} N(\boldsymbol{\omega}, \boldsymbol{G}_{22.1})$.

2. *The test statistic D_n in (3.5) converges to a non-central chi-squared distribution $\chi^2_{k_2}(\Delta)$ with k_2 degrees of freedom and non-centrality parameter Δ.*

Theorem 16. *Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \mathbf{0}^\top)^\top$. Under the local alternative (3.7) and the assumed regularity conditions in Appendix, the joint distributions are:*

$$(i). \quad \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} \sim N_{2k}\left(\begin{bmatrix} \mathbf{0} \\ \boldsymbol{\zeta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G}_{11.2}^{-1} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}^{**} \end{bmatrix}\right)$$

$$(ii). \quad \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_3 \end{bmatrix} \sim N_{2k}\left(\begin{bmatrix} \mathbf{0} \\ -\boldsymbol{\zeta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G}_{11.2}^{-1} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}^{**} \end{bmatrix}\right),$$

where $\boldsymbol{\eta}_1 = \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, $\boldsymbol{\eta}_2 = \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})$, $\boldsymbol{\eta}_3 = \lim_{n\to\infty} \sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$, $\boldsymbol{\zeta} = -\boldsymbol{G}_{11}^{-1}\boldsymbol{G}_{12}\boldsymbol{\delta}$, $\boldsymbol{\Omega}^{**} = \boldsymbol{\Omega}_{12} = \boldsymbol{\Omega}_{21}^\top = \boldsymbol{G}_{11}^{-1}\boldsymbol{G}_{12}\boldsymbol{G}_{22.1}^{-1}\boldsymbol{G}_{21}\boldsymbol{G}_{11}^{-1}$, $\boldsymbol{\Sigma}^{**} = \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\top = \boldsymbol{G}_{11.2}^{-1} - \boldsymbol{G}_{11}^{-1}\boldsymbol{G}_{12}\boldsymbol{G}_{22.1}^{-1}\boldsymbol{G}_{21}\boldsymbol{G}_{11}^{-1}$, and $\boldsymbol{G}_{11.2} = \boldsymbol{G}_{11} - \boldsymbol{G}_{12}\boldsymbol{G}_{22}^{-1}\boldsymbol{G}_{21}$.
The outline of the proof is given in Nkurunziza and Chen (2013).

Theorem 17. *Using the definition of ADB, Theorem 15, and regularity conditions in*

Appendix (see Ahmed et al., 2012), the ADBs of the estimators are,

$$
\begin{aligned}
ADB(\hat{\boldsymbol{\theta}}) &= \mathbf{0} \\
ADB(\tilde{\boldsymbol{\theta}}) &= \boldsymbol{\zeta} \\
ADB(\hat{\boldsymbol{\theta}}_{PT}) &= -\Psi_{k_2+2}\left(\chi^2_{k_2,\alpha}, \Delta\right)\boldsymbol{\zeta} \\
ADB(\hat{\boldsymbol{\theta}}_{SE}) &= -(k_2-2)E\left(\chi^{-2}_{k_2+2,\alpha}(\Delta)\right)\boldsymbol{\zeta} \\
ADB(\hat{\boldsymbol{\theta}}_{PSE}) &= \boldsymbol{ADB}\left(\hat{\boldsymbol{\theta}}_{SE}\right) - \boldsymbol{\zeta}\left[\Psi_{k_2+2}((k_2-2), \Delta)\right] \\
&+ (k_2-2)\boldsymbol{\zeta}E\left[\chi^{-2}_{k_2+2,\alpha}(\Delta)I\left(\chi^2_{k_2+2,\alpha}(\Delta) < (k_2-2)\right)\right],
\end{aligned}
$$

where $\Psi_\nu(\cdot, \Delta)$ is the distribution function of the $\chi^2_\nu(\Delta)$ distribution. Clearly, the bias of the estimators is a function of Δ. Therefore, for bias comparison, it suffices to compare the scalar factor Δ only. It is clear that the ADB of RMLE is an unbounded function of Δ. The ADB($\hat{\boldsymbol{\theta}}_{SE}$) and ADB($\hat{\boldsymbol{\theta}}_{PSE}$) start from the origin, and as Δ increases, they increase to a maximum and then decrease to 0. Note that, $E\left(\chi^{-2}_{k_2+2}(\Delta)\right)$ is a decreasing log-convex function of Δ and the ADB of $\hat{\boldsymbol{\theta}}_{PT}$ is a function of Δ and α. For a fixed α, ADB($\hat{\boldsymbol{\theta}}_{PT}$) starts at zero, increases to a point, then decreases gradually to zero. The proof of this theorem is given below:

Proof: It is obvious that $ADB(\hat{\boldsymbol{\theta}}) = 0$. The ADBs of the reduced, pretest, shrinkage, and positive shrinkage estimators are as follows:

$$
\begin{aligned}
\mathrm{ADB}(\tilde{\boldsymbol{\theta}}) &= \mathrm{E}\left[\lim_{n\to\infty} \sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\right] = \mathrm{E}(\boldsymbol{\eta}_3) = -\boldsymbol{G}_{11}^{-1}\boldsymbol{G}_{12}\boldsymbol{\delta} = \boldsymbol{\zeta}. \\
ADB(\hat{\boldsymbol{\theta}}_{PT}) &= \mathrm{E}\left[\lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\theta}}_{PT} - \boldsymbol{\theta})\right] \\
&= \mathrm{E}\left[\lim_{n\to\infty} \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} - I\left(\hat{D}_n \le \chi^2_{k_2,\alpha}\right)(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})\right)\right] \\
&= -\mathrm{E}\left[\lim_{n\to\infty} \sqrt{n} I\left(\hat{D}_n \le \chi^2_{k_2,\alpha}\right)\left(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\right)\right] \\
&= -\mathrm{E}\left[\lim_{n\to\infty} I\left(\hat{D}_n \le \chi^2_{k_2,\alpha}\right)\boldsymbol{\eta}_2\right] \\
&= -\mathrm{E}\left[I\left(\chi^2_{k_2+2,\alpha}(\Delta) \le \chi^2_{k_2,\alpha}\right)\right]\mathrm{E}(\boldsymbol{\eta}_2) \\
&= \Psi_{k_2+2}\left(\chi^2_{k_2,\alpha}, \Delta\right)\boldsymbol{\zeta}.
\end{aligned}
$$

$$
\begin{aligned}
ADB(\hat{\boldsymbol{\theta}}_{SE}) &= \mathrm{E}\left[\lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\theta}}_{SE} - \boldsymbol{\theta})\right] \\
&= \mathrm{E}\left[\lim_{n\to\infty} \sqrt{n}\left(\tilde{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} - (k_2 - 2)\hat{D}_n^{-1}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) - \boldsymbol{\theta}\right)\right] \\
&= -\mathrm{E}\left[\lim_{n\to\infty} \sqrt{n}((k_2 - 2)\hat{D}_n^{-1}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}))\right] \\
&= -(k_2 - 2)\mathrm{E}\left[\lim_{n\to\infty} \boldsymbol{\eta}_2 \hat{D}_n^{-1}\right] \\
&= -(k_2 - 2)\boldsymbol{\zeta}\mathrm{E}\left[\chi^{-2}_{k_2+2,\alpha}(\Delta)\right] \\
&= -(k_2 - 2)\mathrm{E}\left(\chi^{-2}_{k_2+2,\alpha}(\Delta)\right)\boldsymbol{\zeta}.
\end{aligned}
$$

Observe that we can rewrite $\hat{\boldsymbol{\theta}}_{PSE}$ as

$$
\begin{aligned}
ADB(\hat{\boldsymbol{\theta}}_{PSE}) &= \tilde{\boldsymbol{\theta}} + \left(1-(k_2-2)\hat{D}_n^{-1}\right)^+ (\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}) \\
&= \tilde{\boldsymbol{\theta}} + \left(1-(k_2-2)\hat{D}_n^{-1}\right) (\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}) \\
&- \left(1-(k_2-2)\hat{D}_n^{-1}\right) I\left(\hat{D}_n < (k_2-2)\right) (\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}) \\
&= \hat{\boldsymbol{\theta}}_{SE} - \left(1-(k_2-2)\hat{D}_n^{-1}\right) I\left(\hat{D}_n < (k_2-2)\right) (\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathrm{ADB}(\hat{\boldsymbol{\theta}}_{PSE}) &= \mathrm{E}\left[\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\theta}}_{PSE}-\boldsymbol{\theta})\right] \\
&= \mathrm{E}\Big[\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\theta}}_{SE}-\boldsymbol{\theta}) \\
&- \lim_{n\to\infty}\sqrt{n}\left(1-(k_2-2)\hat{D}_n^{-1}\right) I\left(\hat{D}_n < (k_2-2)\right)(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})\Big] \\
&= \mathrm{ADB}(\hat{\boldsymbol{\theta}}_{SE}) - \mathrm{E}\left[\lim_{n\to\infty}\boldsymbol{\eta}_2\left(1-(k_2-2)\hat{D}_n^{-1}\right) I\left(\hat{D}_n < (k_2-2)\right)\right] \\
&= \mathrm{ADB}(\hat{\boldsymbol{\theta}}_{SE}) \\
&- \boldsymbol{\zeta}\mathrm{E}\left[\left(1-(k_2-2)\chi^{-2}_{k_2+2,\alpha}(\Delta)\right) I\left(\chi^2_{k_2+2,\alpha}(\Delta) < (k_2-2)\right)\right] \\
&= \mathrm{ADB}\left(\hat{\boldsymbol{\theta}}_{SE}\right) - \boldsymbol{\zeta}\mathrm{E}\left[I\left(\chi^2_{k_2+2,\alpha}(\Delta) < (k_2-2)\right)\right] \\
&\quad +(k_2-2)\boldsymbol{\zeta}\mathrm{E}\left[\chi^{-2}_{k_2+2,\alpha}(\Delta) I\left(\chi^2_{k_2+2,\alpha}(\Delta) < (k_2-2)\right)\right] \\
&= \mathrm{ADB}\left(\hat{\boldsymbol{\theta}}_{SE}\right) - \boldsymbol{\zeta}\left[\Psi_{k_2+2}((k_2-2),\Delta)\right] \\
&+ (k_2-2)\boldsymbol{\zeta}\mathrm{E}\left[\chi^{-2}_{k_2+2,\alpha}(\Delta) I\left(\chi^2_{k_2+2,\alpha}(\Delta) < (k_2-2)\right)\right].
\end{aligned}
$$

Theorem 18. *Under the local alternatives K_n in (3.7) and regularity conditions in Ap-*

pendix (see Ahmed et al., 2012), the ADRs of the estimator are

$$
\begin{aligned}
ADR(\hat{\boldsymbol{\theta}};\boldsymbol{W}) &= tr(W\boldsymbol{G}_{11.2}^{-1}). \\
ADR(\tilde{\boldsymbol{\theta}};\boldsymbol{W}) &= ADR(\hat{\boldsymbol{\theta}};\boldsymbol{W})tr(\boldsymbol{W}\boldsymbol{\Sigma}^{**}) + \boldsymbol{\zeta}^\top\boldsymbol{W}\boldsymbol{\zeta}. \\
ADR(\hat{\boldsymbol{\theta}}_{PT};\boldsymbol{W}) &= ADR(\hat{\boldsymbol{\theta}};\boldsymbol{W}) - \Psi_{k_2+2}(\chi^2_{k_2,\alpha},\Delta)tr(\boldsymbol{W}\boldsymbol{\Omega}^{**}) \\
&+ \left[2\Psi_{k_2+2}\left(\chi^2_{k_2,\alpha},\Delta\right) - \Psi_{k_2+4}\left(\chi^2_{k_2,\alpha},\Delta\right)\right]\boldsymbol{\zeta}^\top\boldsymbol{W}\boldsymbol{\zeta}. \\
ADR(\hat{\boldsymbol{\theta}}_{SE};\boldsymbol{W}) &= ADR(\hat{\boldsymbol{\theta}};\boldsymbol{W}) + \left[(k_2-2)^2 E\left(\chi^{-4}_{k_2+2}(\Delta)\right)\right. \\
&- \left.2(k_2-2)E\left(\chi^{-2}_{k_2+2}(\Delta)\right)\right] tr\left(\boldsymbol{W}\boldsymbol{\Omega}^{**}\right) \\
&+ \left[(k_2-1)^2 E\left(\chi^{-4}_{k_2+4}(\Delta)\right) + 2(k_2-2)E\left(\chi^{-2}_{k_2+2}(\Delta)\right)\right. \\
&- \left.2(k_2-2)E\left(\chi^{-2}_{k_2+4}(\Delta)\right)\right]\left(\boldsymbol{\zeta}^\top\boldsymbol{W}\boldsymbol{\zeta}\right). \\
ADR(\hat{\boldsymbol{\theta}}_{PSE};\boldsymbol{W}) &= ADR(\hat{\boldsymbol{\theta}}_{SE};\boldsymbol{W}) \\
&- E\left[(1-(k_2-2)\chi^{-2}_{k_2+2}(\Delta))^2 I\left(\chi^2_{k_2+2}(\Delta) < (k_2-2)\right)\right] tr\left(\boldsymbol{W}\boldsymbol{\Omega}^{**}\right) \\
&+ \left[2\Psi_{k_2+2}((k_2-2),\Delta)\right. \\
&- 2(k_2-2)E\left(\chi^{-2}_{k_2+2}(\Delta) I\left(\chi^2_{k_2+2}(\Delta) < (k_2-2)\right)\right) \\
&- E\left(\left(1-(k_2-2)\chi^{-2}_{k_2+4}(\Delta)\right)^2\right. \\
&\quad \left.\left.\times\ I\left(\chi^2_{k_2+4}(\Delta) < (k_2-2)\right)\right)\right]\left(\boldsymbol{\zeta}^\top\boldsymbol{W}\boldsymbol{\zeta}\right).
\end{aligned}
$$

By comparing the risk of the estimators, we see that, as Δ moves away from 0, the risk of $\tilde{\boldsymbol{\theta}}$ becomes unbounded. That is, the RMLE $\tilde{\boldsymbol{\theta}}$ dominates the unrestricted estimator at and near $\Delta = 0$. The risk of $\hat{\boldsymbol{\theta}}_{PSE}$ is asymptotically superior to $\hat{\boldsymbol{\theta}}_{SE}$ for all values of Δ, with strict inequality holding for some Δ. Thus, not only does $\hat{\boldsymbol{\theta}}_{PSE}$ confirm the inadmissibility of $\hat{\boldsymbol{\theta}}_{SE}$, but it also provides a simple superior estimator. Further, the largest risk improvement of $\hat{\boldsymbol{\theta}}_{PSE}$ over $\hat{\boldsymbol{\theta}}_{SE}$ is at and near the null hypothesis. Also, by comparing the risks of $\hat{\boldsymbol{\theta}}_{SE}$, $\hat{\boldsymbol{\theta}}_{PSE}$, and $\hat{\boldsymbol{\theta}}$, it can be easily shown that, under certain conditions ADR($\hat{\boldsymbol{\theta}}_{PSE}, \boldsymbol{W}$) $\leq$ ADR($\hat{\boldsymbol{\theta}}_{SE}, \boldsymbol{W}$) $\leq$ ADR($\hat{\boldsymbol{\theta}}; \boldsymbol{W}$) for all $\Delta \geq 0$. For a given α, PT is not uniformly better than the unrestricted estimator near the null hypothesis. One may determine an α such that PT has a minimum guaranteed risk. If the minimum efficiency required is RE_0, then we can choose α by solving the equation $\min_{\lambda}\{\text{Relative Efficiency}(\alpha, \Delta)\} = RE_0$. The exact solution may not be available, but we can use a numerical method to search for the minimum. The proof of this theorem is given below:

Proof: To drive the ADR expressions, we first derive the asymptotic covariance matrices for all estimators. The covariance matrix $\boldsymbol{V}(\hat{\boldsymbol{\theta}}^*)$ of any estimator $\boldsymbol{\theta}^*$ is defined as:

$$\boldsymbol{V}(\boldsymbol{\theta}^*) = E\left[\lim_{n\to\infty} n(\boldsymbol{\theta}^* - \boldsymbol{\theta})(\boldsymbol{\theta}^* - \boldsymbol{\theta})^\top\right].$$

First, we derive the covariance matrices of the UMLE and RMLE:

$$\begin{aligned}
\boldsymbol{V}(\hat{\boldsymbol{\theta}}) &= E\left[\lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\right] \\
&= E(\boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top) = \text{Var}(\boldsymbol{\eta}_1) + E(\boldsymbol{\eta}_1)E(\boldsymbol{\eta}_1^\top) = \text{Var}(\boldsymbol{\eta}_1) = \boldsymbol{G}_{11.2}^{-1}. \\
\boldsymbol{V}(\tilde{\boldsymbol{\theta}}) &= E\left[\lim_{n\to\infty} \sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\right] \\
&= E(\boldsymbol{\eta}_3\boldsymbol{\eta}_3^\top) = \text{Var}(\boldsymbol{\eta}_3) + E(\boldsymbol{\eta}_3)E(\boldsymbol{\eta}_3^\top) = \boldsymbol{\Sigma}^{**} + \boldsymbol{\zeta}\boldsymbol{\zeta}^\top.
\end{aligned}$$

Second, we derive the covariance matrices of the pretest estimator:

$$\begin{aligned}
\boldsymbol{V}(\hat{\boldsymbol{\beta}}_{PT}) &= \mathrm{E}\left[\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\theta}}_{PT}-\boldsymbol{\theta})\sqrt{n}(\hat{\boldsymbol{\theta}}_{PT}-\boldsymbol{\theta})^{\top}\right] \\
&= \mathrm{E}\left(\boldsymbol{\eta}_1\boldsymbol{\eta}_1^{\top}+\boldsymbol{\eta}_2\boldsymbol{\eta}_2^{\top}\lim_{n\to\infty} I\left(\hat{D}_n\leq\chi^2_{k_2,\alpha}\right)-2\boldsymbol{\eta}_2\boldsymbol{\eta}_1^{\top}\lim_{n\to\infty} I\left(\hat{D}_n\leq\chi^2_{k_2,\alpha}\right)\right) \\
&= \mathrm{Var}\left(\boldsymbol{\eta}_1\right)+\mathrm{E}\left(\boldsymbol{\eta}_1\right)\mathrm{E}\left(\boldsymbol{\eta}_1^{\top}\right)+\mathrm{E}\left(\boldsymbol{\eta}_2\boldsymbol{\eta}_2^{\top}\lim_{n\to\infty} I\left(\hat{D}_n\leq\chi^2_{k_2,\alpha}\right)\right) \\
&\quad -2\mathrm{E}\left(\boldsymbol{\eta}_2\boldsymbol{\eta}_1^{\top}\lim_{n\to\infty} I\left(\hat{D}_n\leq\chi^2_{k_2,\alpha}\right)\right) \\
&= \boldsymbol{G}_{11.2}^{-1}+\boldsymbol{\Omega}^{**}\Psi_{k_2+2}\left(\chi^2_{k_2,\alpha},\Delta\right)+\boldsymbol{\zeta}\boldsymbol{\zeta}^{\top}\Psi_{k_2+4}\left(\chi^2_{k_2,\alpha}\right) \\
&- 2\mathrm{E}\left(\boldsymbol{\eta}_2\boldsymbol{\eta}_1^{\top}\lim_{n\to\infty} I\left(\hat{D}_n\leq\chi^2_{k_2,\alpha}\right)\right).
\end{aligned}$$

Consider the fourth term:

$$\begin{aligned}
&\mathrm{E}\left(\boldsymbol{\eta}_2\boldsymbol{\eta}_1^{\top}\lim_{n\to\infty} I\left(\hat{D}_n\leq\chi^2_{k_2,\alpha}\right)\right)=\mathrm{E}\left(\mathrm{E}\left(\boldsymbol{\eta}_2\boldsymbol{\eta}_1^{\top}\lim_{n\to\infty} I\left(\hat{D}_n\leq\chi^2_{k_2,\alpha}\right)|\boldsymbol{\eta}_2\right)\right) \\
=&\ \mathrm{E}\left(\boldsymbol{\eta}_2\left(\mathrm{E}\left(\boldsymbol{\eta}_1\right)+\mathrm{Cov}\left(\boldsymbol{\eta}_2,\boldsymbol{\eta}_1^{\top}\right)\boldsymbol{\Omega}^{**-1}\left(\boldsymbol{\eta}_2-\mathrm{E}(\boldsymbol{\eta}_2)\right)\right)^{\top}\lim_{n\to\infty} I\left(\hat{D}_n\leq\chi^2_{k_2,\alpha}\right)\right) \\
=&\ \mathrm{E}\left(\boldsymbol{\eta}_2\left(\boldsymbol{\eta}_2^{\top}-\mathrm{E}\left(\boldsymbol{\eta}_2\right)^{\top}\right)\boldsymbol{\Omega}^{**-1}\mathrm{Cov}\left(\boldsymbol{\eta}_2,\boldsymbol{\eta}_1\right)\lim_{n\to\infty} I\left(\hat{D}_n\leq\chi^2_{k_2,\alpha}\right)\right) \\
=&\ \mathrm{E}\left(\boldsymbol{\eta}_2\boldsymbol{\eta}_2^{\top}\lim_{n\to\infty} I\left(\hat{D}_n\leq\chi^2_{k_2,\alpha}\right)\right)\boldsymbol{\Omega}^{**-1}\boldsymbol{\Omega}_{12} \\
&\ -\mathrm{E}\left(\boldsymbol{\eta}_2\lim_{n\to\infty} I\left(\hat{D}_n\leq\chi^2_{k_2,\alpha}\right)\right)\mathrm{E}\left(\boldsymbol{\eta}_2\right)^{\top}\boldsymbol{\Omega}^{**-1}\boldsymbol{\Omega}_{12} \\
=&\ \boldsymbol{\Omega}^{**}\Psi_{k_2+2}\left(\chi^2_{k_2,\alpha},\Delta\right)\boldsymbol{\Omega}^{**-1}\boldsymbol{\Omega}_{12}+\boldsymbol{\zeta}\boldsymbol{\zeta}^{\top}\Psi_{k_2+4}\left(\chi^2_{k_2,\alpha},\Delta\right)\boldsymbol{\Omega}^{**-1}\boldsymbol{\Omega}_{12} \\
&\ -\boldsymbol{\zeta}\boldsymbol{\zeta}^{\top}\Psi_{k_2+2}\left(\chi^2_{k_2,\alpha},\Delta\right)\boldsymbol{\Omega}^{**-1}\boldsymbol{\Omega}_{12} \\
=&\ \boldsymbol{\Omega}^{**}\Psi_{k_2+2}\left(\chi^2_{k_2,\alpha},\Delta\right)+\boldsymbol{\zeta}\boldsymbol{\zeta}^{\top}\Psi_{k_2+4}\left(\chi^2_{k_2,\alpha},\Delta\right)-\boldsymbol{\zeta}\boldsymbol{\zeta}^{\top}\Psi_{k_2+2}\left(\chi^2_{k_2,\alpha},\Delta\right).
\end{aligned}$$

Hence

$$
\begin{aligned}
\boldsymbol{V}(\hat{\boldsymbol{\theta}}_{PT}) &= \boldsymbol{G}_{11.2}^{-1} + \lim_{n\to\infty} \mathrm{E}\left(\boldsymbol{\eta}_2\boldsymbol{\eta}_2^\top I\left(\hat{D}_n \le \chi^2_{k_2,\alpha}\right)\right) \\
&- 2\lim_{n\to\infty} \mathrm{E}\left(\boldsymbol{\eta}_2\boldsymbol{\eta}_1^\top I\left(\hat{D}_n \le \chi^2_{k_2,\alpha}\right)\right) \\
&= \boldsymbol{G}_{11.2}^{-1} + \boldsymbol{\Omega}^{**}\Psi_{k_2+2}\left(\chi^2_{k_2,\alpha},\Delta\right) + \boldsymbol{\zeta}\boldsymbol{\zeta}^\top\Psi_{k_2+4}\left(\chi^2_{k_2,\alpha},\Delta\right) \\
& \quad -2\left[\boldsymbol{\Omega}^{**}\Psi_{k_2+2}\left(\chi^2_{k_2,\alpha},\Delta\right) + \boldsymbol{\zeta}\boldsymbol{\zeta}^\top\Psi_{k_2+4}\left(\chi^2_{k_2,\alpha},\Delta\right)\right. \\
&- \left.\boldsymbol{\zeta}\boldsymbol{\zeta}^\top\Psi_{k_2+4}\left(\chi^2_{k_2,\alpha},\Delta\right)\right] \\
&= \boldsymbol{G}_{11.2}^{-1} - \boldsymbol{\Omega}^{**}\Psi_{k_2+2}\left(\chi^2_{k_2,\alpha},\Delta\right) - \boldsymbol{\zeta}\boldsymbol{\zeta}^\top\Psi_{k_2+4}\left(\chi^2_{k_2,\alpha},\Delta\right) \\
&+ 2\boldsymbol{\zeta}\boldsymbol{\zeta}^\top\Psi_{k_2+2}\left(\chi^2_{k_2,\alpha},\Delta\right).
\end{aligned}
$$

Third, we derive the covariance matrices of the shrinkage estimators:

$$
\begin{aligned}
\boldsymbol{V}(\hat{\boldsymbol{\theta}}_{SE}) &= \mathrm{E}\left[\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\theta}}_{SE}-\boldsymbol{\theta})\sqrt{n}(\hat{\boldsymbol{\theta}}_{SE}-\boldsymbol{\theta})^\top\right] \\
&= \mathrm{E}\left[\lim_{n\to\infty}\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}-(k_2-2)\hat{D}_n^{-1}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})\right)\right. \\
&= \left.\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}-(k_2-2)\hat{D}_n^{-1}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})\right)^\top\right] \\
&= \mathrm{E}(\boldsymbol{\eta}_1\boldsymbol{\eta}_1^\top) + (k_2-2)^2\mathrm{E}\left(\boldsymbol{\eta}_2\boldsymbol{\eta}_2^\top\lim_{n\to\infty}\hat{D}_n^{-2}\right) \\
&- 2(k_2-2)\mathrm{E}\left(\boldsymbol{\eta}_2\boldsymbol{\eta}_1^\top\lim_{n\to\infty}\hat{D}_n^{-1}\right) \\
&= \boldsymbol{G}_{11.2}^{-1} + \boldsymbol{\Omega}^{**}\mathrm{E}\left(\chi^{-4}_{k_2+2}(\Delta)\right) + \boldsymbol{\zeta}\boldsymbol{\zeta}^\top\mathrm{E}\left(\chi^{-4}_{k_2+4}(\Delta)\right) \\
&- 2(k_2-2)\mathrm{E}\left(\boldsymbol{\eta}_2\boldsymbol{\eta}_1^\top\lim_{n\to\infty}\hat{D}_n^{-1}\right).
\end{aligned}
$$

Consider the last term:

$$
\begin{aligned}
& \mathrm{E}(\boldsymbol{\eta}_2\boldsymbol{\eta}_1^\top \lim_{n\to\infty}\hat{D}_n^{-1}) = \mathrm{E}\left(\mathrm{E}(\boldsymbol{\eta}_2\boldsymbol{\eta}_1^\top \lim_{n\to\infty}\hat{D}_n^{-1}|\boldsymbol{\eta}_2)\right) \\
= & \mathrm{E}\left(\boldsymbol{\eta}_2\mathrm{E}(\boldsymbol{\eta}_1^\top|\boldsymbol{\eta}_2)\lim_{n\to\infty}\hat{D}_n^{-1}\right) + \mathrm{E}\left(\boldsymbol{\eta}_2\left(\boldsymbol{\eta}_2^\top - \mathrm{E}(\boldsymbol{\eta}_2)^\top\right)\boldsymbol{\Omega}^{**-1}\boldsymbol{\Omega}_{12}\lim_{n\to\infty}\hat{D}_n^{-1}\right) \\
= & \mathrm{E}\left(\boldsymbol{\eta}_2\boldsymbol{\eta}_2^\top \lim_{n\to\infty}\hat{D}_n^{-1}\right)\boldsymbol{\Omega}^{**-1}\boldsymbol{\Omega}_{12} \\
& -\mathrm{E}\left(\boldsymbol{\eta}_2 \lim_{n\to\infty}\hat{D}_n^{-1}\right)\mathrm{E}\left(\boldsymbol{\eta}_2\right)^\top\boldsymbol{\Omega}^{**-1}\boldsymbol{\Omega}_{12} \\
= & \boldsymbol{\Omega}^{**}\mathrm{E}\left(\chi^{-2}_{k_2+2,\alpha}(\Delta)\right) + \boldsymbol{\zeta}\boldsymbol{\zeta}^\top\mathrm{E}\left(\chi^{-2}_{k_2+4,\alpha}(\Delta)\right) - \boldsymbol{\zeta}\boldsymbol{\zeta}^\top\mathrm{E}\left(\chi^{-2}_{k_2+2,\alpha}(\Delta)\right) \\
= & \boldsymbol{\Omega}^{**}\mathrm{E}\left(\chi^{-2}_{k_2+2}(\Delta)\right) + \boldsymbol{\zeta}\boldsymbol{\zeta}^\top\mathrm{E}\left(\chi^{-2}_{k_2+4}(\Delta)\right) - \boldsymbol{\zeta}\boldsymbol{\zeta}^\top\mathrm{E}\left(\chi^{-2}_{k_2+2}(\Delta)\right).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\boldsymbol{V}(\hat{\boldsymbol{\theta}}_{SE}) &= \boldsymbol{G}^{-1}_{11.2} + (k_2-2)^2\left[\boldsymbol{\Omega}^{**}\mathrm{E}(\chi^{-4}_{k_2+2}(\Delta)) + \boldsymbol{\zeta}\boldsymbol{\zeta}^\top\mathrm{E}(\chi^{-4}_{k_2+4}(\Delta))\right] \\
&- 2(k_2-2)\left[\boldsymbol{\Omega}^{**}\mathrm{E}(\chi^{-2}_{k_2+2}(\Delta)) + \boldsymbol{\zeta}\boldsymbol{\zeta}^\top\mathrm{E}(\chi^{-2}_{k_2+4}(\Delta)) - \boldsymbol{\zeta}\boldsymbol{\zeta}^\top\mathrm{E}(\chi^{-2}_{k_2+2}(\Delta))\right] \\
&= \boldsymbol{G}^{-1}_{11.2} + \left[(k_2-2)^2\mathrm{E}\left(\chi^{-4}_{k_2+2}(\Delta)\right) - 2(k_2-2)\mathrm{E}(\chi^{-2}_{k_2+2}(\Delta))\right]\boldsymbol{\Omega}^{**} \\
&+ \left[(k_2-2)^2\mathrm{E}(\chi^{-4}_{k_2+4}(\Delta)) + 2(k_2-2)\mathrm{E}(\chi^{-2}_{k_2+2}(\Delta))\right. \\
&- \left. 2(k_2-2)\mathrm{E}(\chi^{-2}_{k_2+4})\right]\boldsymbol{\zeta}\boldsymbol{\zeta}^\top.
\end{aligned}
$$

Let $R_{n+l}(\Delta) = \left(1-(k_2-2)\hat{D}_n^{-1}\right)^l I\left(\hat{D}_n < (k_2-2)\right)$, where $l = 1, 2$

$$
\begin{aligned}
\boldsymbol{V}(\hat{\boldsymbol{\theta}}_{PSE}) &= \mathrm{E}\left(\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\theta}}_{PSE}-\boldsymbol{\theta})\sqrt{n}(\hat{\boldsymbol{\theta}}_{PSE}-\boldsymbol{\theta})^{\top}\right), \\
&= \mathrm{E}\left(\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\theta}}_{SE}-\boldsymbol{\theta})\sqrt{n}(\hat{\boldsymbol{\theta}}_{SE}-\boldsymbol{\theta})^{\top}\right) \\
&\quad +\mathrm{E}\left(\lim_{n\to\infty}R_{n+2}(\Delta)\sqrt{n}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})\sqrt{n}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})^{\top}\right) \\
&\quad -2\mathrm{E}\left(\lim_{n\to\infty}R_{n+1}(\Delta)\sqrt{n}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})\sqrt{n}(\hat{\boldsymbol{\theta}}_{SE}-\boldsymbol{\theta})^{\top}\right) \\
&= \boldsymbol{V}(\hat{\boldsymbol{\theta}}_{SE})+\mathrm{E}\left(\lim_{n\to\infty}R_{n+2}(\Delta)\boldsymbol{\eta}_2\boldsymbol{\eta}_2^{\top}\right) \\
&- \quad 2\mathrm{E}\left(\lim_{n\to\infty}R_{n+1}(\Delta)\boldsymbol{\eta}_2\left(\boldsymbol{\eta}_3^{\top}+\left(1-(k_2-2)\hat{D}_n^{-1}\right)\boldsymbol{\eta}_2^{\top}\right)\right), \\
&= \boldsymbol{V}(\hat{\boldsymbol{\theta}}_{SE})-\mathrm{E}\left(\lim_{n\to\infty}R_{n+2}(\Delta)\boldsymbol{\eta}_2\boldsymbol{\eta}_2^{\top}\right)-2\mathrm{E}\left(\lim_{n\to\infty}R_{n+1}(\Delta)\boldsymbol{\eta}_2\boldsymbol{\eta}_3^{\top}\right).
\end{aligned}
$$

Consider the second term:

$$
\begin{aligned}
& -\mathrm{E}\left(\lim_{n\to\infty}R_{n+2}(\Delta)\boldsymbol{\eta}_2\boldsymbol{\eta}_2^{\top}\right) \\
=& \quad -\mathrm{E}\left(\lim_{n\to\infty}\left(1-(k_2-2)\hat{D}_n^{-1}\right)^2 I\left(\hat{D}_n<(k_2-2)\right)\boldsymbol{\eta}_2\boldsymbol{\eta}_2^{\top}\right) \\
=& \quad -\boldsymbol{\Omega}^{**}\mathrm{E}\left(I\left(\chi^2_{k_2+2}(\Delta)<(k_2-2)\right)\left[1-(k_2-2)\chi^{-2}_{k_2+2}(\Delta)\right]^2\right) \\
-& \quad \boldsymbol{\zeta}\boldsymbol{\zeta}^{\top}\mathrm{E}\left(I\left(\chi^2_{k_2+4}(\Delta)<(k_2-2)\right)\left[1-(k_2-2)\chi^{-2}_{k_2+4}(\Delta)\right]^2\right).
\end{aligned}
$$

Consider the third term:

$$
\begin{aligned}
& -2\mathrm{E}\left(\lim_{n\to\infty} R_{n+1}(\Delta)\boldsymbol{\eta}_2\boldsymbol{\eta}_3^\top\right) = -2\mathrm{E}\left(\lim_{n\to\infty} \boldsymbol{\eta}_2\mathrm{E}\left[R_{n+1}(\Delta)\boldsymbol{\eta}_3^\top|\boldsymbol{\eta}_2\right]\right) \\
= & -2\mathrm{E}\left(\lim_{n\to\infty} \boldsymbol{\eta}_2\left[\mathrm{E}\left(\boldsymbol{\eta}_3^\top\right) + \mathrm{Cov}\left(\boldsymbol{\eta}_2,\boldsymbol{\eta}_3\right)\boldsymbol{\Omega}^{**-1}\left(\boldsymbol{\eta}_2 - \mathrm{E}\left(\boldsymbol{\eta}_2\right)\right)\right] R_{n+1}(\Delta)\right) \\
= & -2\mathrm{E}\left(\lim_{n\to\infty} \boldsymbol{\eta}_2\mathrm{E}\left(\boldsymbol{\eta}_3^\top\right) R_{n+1}(\Delta) + \mathbf{0}\right) \\
= & -2\mathrm{E}\left[\lim_{n\to\infty} \boldsymbol{\eta}_2 I\left(\hat{D}_n < (k_2-2)\right) - (k_2-2)\hat{D}_n^{-1}\boldsymbol{\eta}_2 I\left(\hat{D}_n < (k_2-2)\right)\right]\mathrm{E}\left(\boldsymbol{\eta}_3^\top\right) \\
= & \ 2\Psi_{(k_2-2)+4}((k_2-2),\Delta)\boldsymbol{\zeta}\boldsymbol{\zeta}^\top \\
- & \ 2(k_2-2)\mathrm{E}\left(\chi^{-2}_{k_2+2}(\Delta) I\left(\chi^2_{k_2+2}(\Delta) < (k_2-2)\right)\right)\boldsymbol{\zeta}\boldsymbol{\zeta}^\top \\
= & \left[2\Psi_{(k_2-2)+4}((k_2-2),\Delta)\right. \\
- & \left. 2(k_2-2)\mathrm{E}\left(\chi^{-2}_{k_2+2}(\Delta) I\left(\chi^2_{k_2+2}(\Delta) < (k_2-2)\right)\right)\right]\boldsymbol{\zeta}\boldsymbol{\zeta}^\top.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
\boldsymbol{V}\left(\hat{\boldsymbol{\theta}}_{PSE}\right) &= \boldsymbol{V}\left(\hat{\boldsymbol{\theta}}_{SE}\right) - \mathrm{E}\left[\left(1-(k_2-2)\chi^{-2}_{k_2+2}(\Delta)\right)^2 I\left(\chi^2_{k_2+2}(\Delta) < (k_2-2)\right)\right]\boldsymbol{\Omega}^{**} \\
&+ \left[2\Psi_{(k_2-2)+4}((k_2-2),\Delta)\right. \\
&- 2(k_2-2)\mathrm{E}\left(\chi^{-2}_{k_2+2}(\Delta) I\left(\chi^2_{k_2+2}(\Delta) < (k_2-2)\right)\right) \\
&- \left.\mathrm{E}\left[\left(1-(k_2-2)\chi^{-2}_{k_2+4}(\Delta)\right)^2 I\left(\chi^2_{k_2+4}(\Delta) < (k_2-2)\right)\right]\right]\boldsymbol{\zeta}\boldsymbol{\zeta}^\top.
\end{aligned}
$$

The risk expressions in Theorem 3.3 now follow from (3.8) which completes the proof. In order to explain and quantify the properties of the theoretical results, we conduct a simulation study to compare the performance of the suggested estimators.

3.2.1 Simulation Studies

In this section, we carry out a Monte Carlo simulation study to examine the risk (namely MSE) performance of the estimators. This simulation study is based on a zero-inflated Poisson autoregression model with different numbers of covariates. Our sampling experiment consists of different combinations of the length of time series, i.e., $n = 200$, and 300. In this study we simulate time series data by the ZIP autoregression:

$$\log(\lambda_t) = \boldsymbol{x}_{t-1}^\top \boldsymbol{\beta} + \sigma\varepsilon_t$$

and

$$\log\left(\frac{\boldsymbol{\omega}_t}{1-\boldsymbol{\omega}_t}\right) = \boldsymbol{z}_{t-1}^\top \boldsymbol{\gamma},$$

where the covariates x_{t-1} and z_{t-1} are taken to be lagged values of the response y_t hence the ZIP autoregression. Here, ε_t is an unobservable realization from the standard normal distribution, included in λ_t to optionally induce extra overdispersion in the data. We generate y_t using the $rzip$ function in the **ZIM** package in R. In the $rzip$ function, we must supply λ and ω. We calculate λ and ω based on the lag responses and true values of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$. We set the true values of $\boldsymbol{\theta} = (\theta_1^\top, \theta_2^\top)^\top = ((1.1, 1.8, 0.68, -1.7)^\top, \boldsymbol{b}^\top)^\top$ where $\boldsymbol{b}$ is a zero vector with different lengths. We set $\sigma = 0.10$.

For simulation, we consider the particular hypothesis $H_0 : \boldsymbol{\theta_2} = \mathbf{0}$ vs. $H_A : \boldsymbol{\theta_2} \neq \mathbf{0}$, where $\boldsymbol{\theta}_2$ is a $k_2 \times 1$ vector with $k = k_1 + k_2$. The summary of simulation result is provided for $(k_1, k_2) = \{(4,5), (4,9), (4,14), (4,18)\}$, $\alpha = 0.05$ for different sample sizes. Under H_0, the number of simulations was varied initially and it was determined that 1000 of each set of observations were adequate, since a further increase in the number of replications did not significantly change the result. We define the parameter $\Delta = ||\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}||^2$, where

$\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_{\mathbf{1}}^{\top}, \mathbf{0}^{\top})^{\top}$ and $||\cdot||$ is the Euclidian norm. In order to investigate the performance of the estimators for $\Delta > 0$, further responses were generated under local alternative hypotheses (i.e., for different Δ between 0 and 2). All computations were conducted using the **R** statistical system Ihaka and Gentleman (1996). We calculate the simulated mean squared errors (SMSE) by using the empirical formula:

$$SMSE(\boldsymbol{\theta}^*) = \sum_{i=1}^{p+q} (\boldsymbol{\theta}_{\boldsymbol{i}}^* - \boldsymbol{\theta}_{\boldsymbol{i}})^2.$$

The objective here is to investigate the behaviour of the proposed estimators for $\Delta \geq 0$. The criterion for comparing the performance of any estimator $\boldsymbol{\theta}^*$ in this study is the mean squared error (MSE), where $\boldsymbol{\theta}^*$ is any of the estimators $\hat{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}_{PT}$, $\hat{\boldsymbol{\theta}}_{SE}$, and $\hat{\boldsymbol{\theta}}_{PSE}$. The simulated relative mean squared error (RMSE) of $\boldsymbol{\theta}^*$ to $\hat{\boldsymbol{\theta}}$ is defined as

$$\text{RMSE}(\hat{\boldsymbol{\theta}} : \boldsymbol{\theta}^*) = \text{MSE}(\hat{\boldsymbol{\theta}})/\text{MSE}(\boldsymbol{\theta}^*).$$

Observe that an RMSE > 1 indicates the degree of superiority of $\boldsymbol{\theta}^*$ over $\hat{\boldsymbol{\theta}}$.

Our theoretical results were applied to various simulated data sets. Tables 3.1 to 3.8 provide the estimated relative efficiency for various estimators over $\hat{\boldsymbol{\theta}}$ for $n = 200$ and 300. The results can be summarized as follows:

Table 3.1: Simulated relative MSEs of RMLE, PT, SE and PSE with respect to UMLE $\hat{\boldsymbol{\theta}}$ when the hypothesis misspecifies for $n = 200$ and $k_2 = 5$.

Δ	$RMLE$	PT	SE	PSE
0.0	1.507	1.483	1.289	1.292
0.1	1.497	1.467	1.268	1.270
0.3	1.333	1.238	1.174	1.174
0.7	0.997	0.977	1.039	1.039
1.0	0.840	1.000	1.014	1.014
1.2	0.752	1.000	1.005	1.005
1.6	0.679	1.000	1.002	1.002
1.8	0.652	1.000	1.001	1.001
2.0	0.630	1.000	1.001	1.001

Table 3.2: Simulated relative MSEs of RMLE, PT, SE and PSE with respect to UMLE $\hat{\boldsymbol{\theta}}$ when the hypothesis misspecifies for $n = 300$ and $k_2 = 5$.

Δ	$RMLE$	PT	SE	PSE
0	1.470	1.448	1.268	1.269
0.1	1.418	1.398	1.239	1.241
0.3	1.240	1.113	1.115	1.116
0.7	0.932	0.997	1.019	1.019
1.0	0.798	1.000	1.005	1.005
1.2	0.750	1.000	1.002	1.002
1.6	0.689	1.000	1.001	1.001
1.8	0.677	1.000	1.000	1.000
2.0	0.658	1.000	1.000	1.000

Table 3.3: Simulated relative MSEs of RMLE, PT, SE and PSE with respect to UMLE $\hat{\boldsymbol{\theta}}$ when the hypothesis misspecifies for $n = 200$ and $k_2 = 9$.

Δ	$RMLE$	PT	SE	PSE
0.0	1.524	1.444	1.472	1.494
0.1	1.484	1.434	1.424	1.434
0.3	1.429	1.402	1.342	1.348
0.7	1.223	1.039	1.150	1.150
1.0	1.079	1.001	1.075	1.075
1.2	0.970	1.000	1.041	1.041
1.6	0.845	1.000	1.017	1.017
1.8	0.789	1.000	1.011	1.011
2.0	0.755	1.000	1.008	1.008

Table 3.4: Simulated relative MSEs of RMLE, PT, SE and PSE with respect to UMLE $\hat{\boldsymbol{\theta}}$ when the hypothesis misspecifies for $n = 300$ and $k_2 = 9$.

Δ	$RMLE$	PT	SE	PSE
0.0	1.268	1.267	1.237	1.246
0.1	1.270	1.269	1.238	1.244
0.3	1.251	1.232	1.193	1.195
0.7	1.137	1.000	1.078	1.078
1.0	1.016	1.000	1.034	1.034
1.2	0.939	1.000	1.019	1.019
1.6	0.838	1.000	1.007	1.007
1.8	0.792	1.000	1.005	1.005
2.0	0.760	1.000	1.003	1.003

Table 3.5: Simulated relative MSEs of RMLE, PT, SE and PSE with respect to UMLE $\hat{\boldsymbol{\theta}}$ when the hypothesis misspecifies for $n = 200$ and $k_2 = 14$.

Δ	$RMLE$	PT	SE	PSE
0.0	1.964	1.957	1.861	1.909
0.1	1.941	1.941	1.844	1.876
0.3	1.848	1.815	1.740	1.758
0.7	1.500	1.144	1.356	1.356
1.0	1.239	1.004	1.176	1.176
1.2	1.116	1.001	1.106	1.106
1.6	0.903	1.000	1.041	1.041
1.8	0.848	1.000	1.028	1.028
2.0	0.795	1.000	1.018	1.018

Table 3.6: Simulated relative MSEs of RMLE, PT, SE and PSE with respect to UMLE $\hat{\boldsymbol{\theta}}$ when the hypothesis misspecifies for $n = 300$ and $k_2 = 14$.

Δ	$RMLE$	PT	SE	PSE
0.0	1.489	1.489	1.454	1.469
0.1	1.490	1.486	1.446	1.460
0.3	1.477	1.440	1.404	1.408
0.7	1.253	1.020	1.169	1.169
1.0	1.088	1.000	1.074	1.074
1.2	1.005	1.000	1.046	1.046
1.6	0.866	1.000	1.017	1.017
1.8	0.814	1.000	1.010	1.010
2.0	0.786	1.000	1.007	1.007

Table 3.7: Simulated relative MSEs of RMLE, PT, SE and PSE with respect to UMLE $\hat{\boldsymbol{\theta}}$ when the hypothesis misspecifies for $n = 200$ and $k_2 = 18$.

Δ	$RMLE$	PT	SE	PSE
0.0	2.818	2.818	2.632	2.701
0.1	2.712	2.696	2.530	2.614
0.3	2.450	2.450	2.285	2.324
0.7	1.884	1.273	1.638	1.638
1.0	1.515	1.012	1.322	1.322
1.2	1.289	1.000	1.190	1.190
1.6	1.021	1.000	1.076	1.076
1.8	0.931	1.000	1.049	1.049
2.0	0.858	1.000	1.030	1.030

Table 3.8: Simulated relative MSEs of RMLE, PT, SE and PSE with respect to UMLE $\hat{\boldsymbol{\theta}}$ when the hypothesis misspecifies for $n = 300$ and $k_2 = 18$.

Δ	$RMLE$	PT	SE	PSE
0.0	1.730	1.727	1.673	1.704
0.1	1.737	1.733	1.679	1.704
0.3	1.657	1.634	1.588	1.595
0.7	1.382	1.037	1.264	1.264
1.0	1.177	1.000	1.120	1.120
1.2	1.055	1.000	1.068	1.068
1.6	0.912	1.000	1.027	1.027
1.8	0.853	1.000	1.018	1.018
2.0	0.805	1.000	1.012	1.012

We summarize our findings as follows.

(i) From Tables 3.1-3.8 and figures 3.1-3.2 , we observe that the maximum RMSE occurred at and near $\Delta^* = 0$. As evident by Tables 3.1-3.8, the RMLE consistently outperforms the other estimators at and near $\Delta^* = 0$ due to its unbiasedness property, and the RMSE of all estimators is asymptotically converging to 1. Therefore, if the restricted maximum likelihood estimator is nearly correctly specified, then the RMLE is the optimal estimator. On the contrary, as the hypothesis error i.e., Δ^* deviates from zero, the risk of RMLE increases and becomes unbounded while the risk of shrinkage and positive shrinkage estimators remain below the risk of UMLE and merge with it as $\Delta^* \to \infty$. It can be safely concluded that the risk of RMLE explodes as Δ^* increases, but it has less impact on shrinkage and positive shrinkage estimators, which is consistent with the theory.

(ii) For small Δ^*, we find that the PSE is outperforming the SE. For large values of Δ^*, we find that the RMSE's are same for SE and PSE. Therefore, the PSE outperforms the SE at and near $\Delta^* = 0$.

(iii) The PT estimator outperforms the PSE for all Δ^* when the number of insignificant covariates increases. The PT outperforms the PSE only for small Δ^*, and the roles are reversed as Δ^* increases.

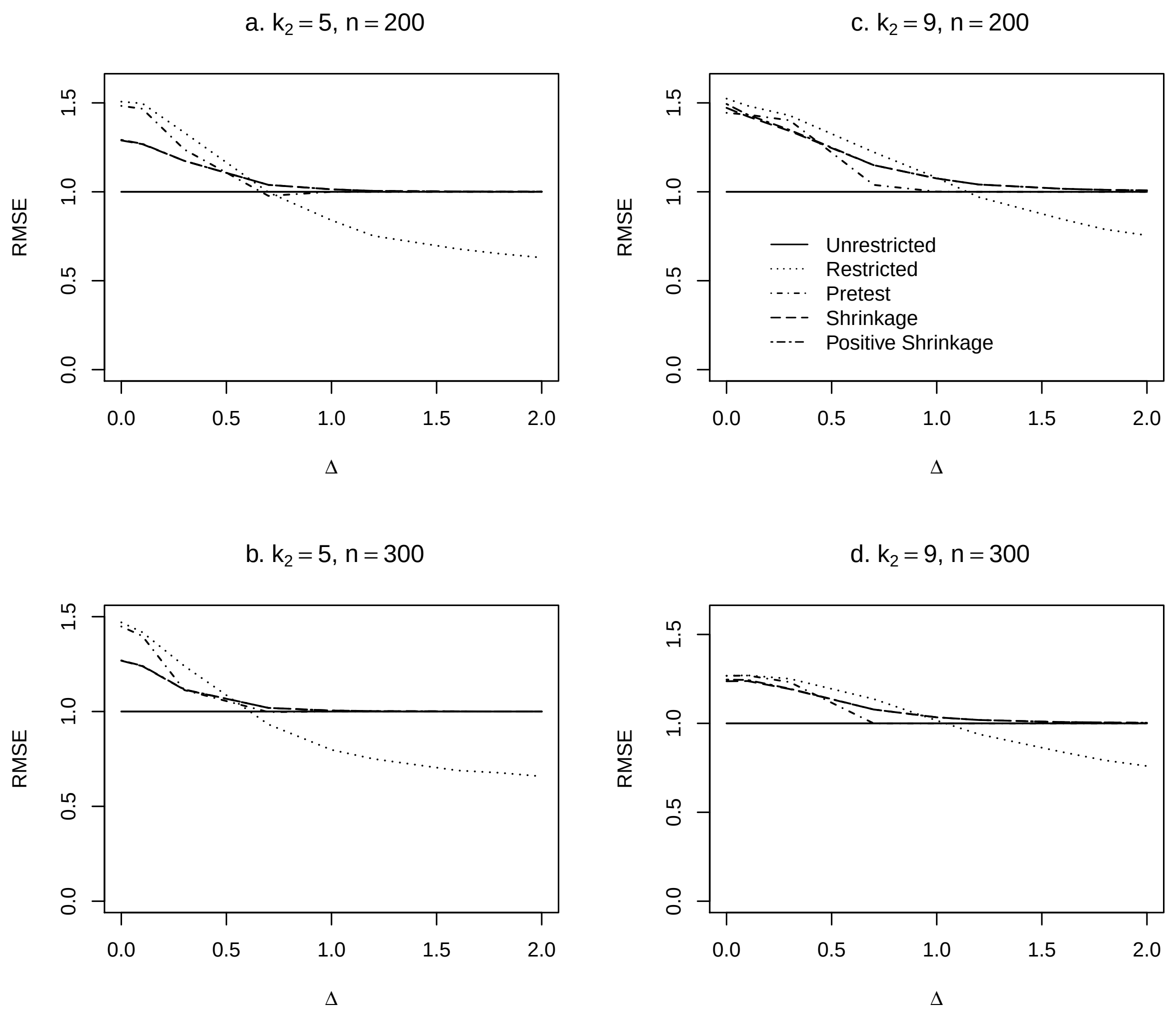

Figure 3.1: Simulated Relative MSE with respect to UMLE, $\hat{\boldsymbol{\theta}}$ of the estimates for $\Delta \geq 0$. Here $k_1 = 4$, $k_2 = 5$; $t = 200, 300$ for the first column and $k_1 = 4$, $k_2 = 9$; $n = 200, 300$ for the second column.

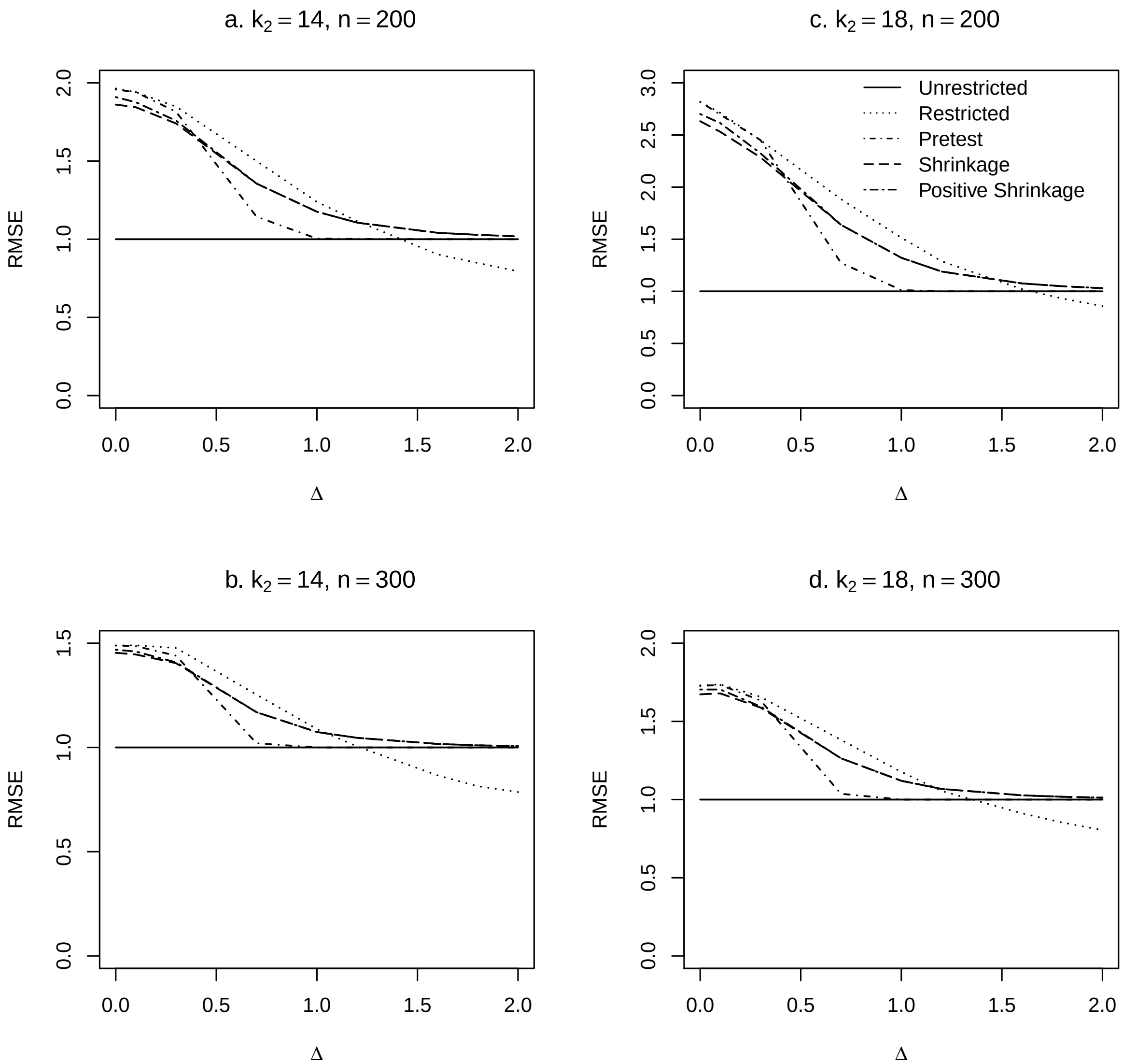

Figure 3.2: Simulated Relative MSE with respect to UMLE, $\hat{\boldsymbol{\theta}}$ of the estimates for $\Delta \geq 0$. Here $k_1 = 4$, $k_2 = 14$; $t = 200, 300$ for the first column and $k_1 = 4$, $k_2 = 18$; $n = 200, 300$ for the second column.

Chapter 4

Conclusions and Future Research

In this thesis, we considered the different estimation methods for zero inflated autoregressive conditional Poisson (ZIACP) models with and without covariates. Application of these methods has been demonstrated in ZIACP models with real data examples. We summarize the findings as follows:

Chapter one summarizes the key concepts from Harvey's book on dynamic models for volatility and heavy tails (Harvey, 2013) that are relevant to modeling time varying parameters. Asymptotic distribution of the estimated parameters that govern the behaviour of dynamic models was verified. The generalized autoregressive score (GAS) model which is a relatively new class of models where the conditional mean or variance is modelled as a function of past values of itself and past values of the scaled score was also briefly discussed.

Chapter two considered the estimation of the parameters of the ZIP distribution using the expectation maximization (EM) algorithm. We summarize the Poisson autoregression results of Fokianos et al. (2009) and these results are extended together with the estimation strategy for ZIP distribution to model the parameters of the ZIACP models using the

EM algorithm approach. A simulation study was conducted to validate and verify the estimation approaches for the ZIACP models. The normality of the estimates was also verified in some cases. Real data examples demonstrated the superiority of the modeling approach to other competing models in the literature.

In Chapter three, we proposed the restricted, the pretest and the shrinkage estimators in the ZIPA model under a restriction, $\boldsymbol{\beta} = 0$. The joint asymptotic distribution of the UMLE and RMLE is provided. Consequently, we derived the asymptotic distributional risks and biases of the proposed estimators. We examined analytically the relative dominance picture of the proposed estimators with respect to the UMLE of $\boldsymbol{\beta}$. We also carried out a Monte Carlo simulation study to compare these estimators in terms of their relative mean squared errors. We concluded that among the proposed estimators, the positive shrinkage estimator performs the best in the sense of giving the smallest mean squared prediction error.

Since forecasting is an important and key concept in time series analysis, the development of robust forecasting techniques for the zero inflated autoregressive conditional Poisson models will be an interesting topic for future research.

The zero inflated autoregressive conditional negative binomial models are useful for analysis of over-dispersed count data with an excess of zeros. There has not been any study investigating the shrinkage methods for this model. Introducing the shrinkage estimation method for this model will be an interesting topic for my future research.

Appendix

Regularity conditions (Kedem and Fokianos, 2005)

A1. The true parameter $\boldsymbol{\theta}$ belongs to an open set $B \subseteq \Re^k$.

A2. The covariate vector $\boldsymbol{\varsigma}_{t-1} = (\boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1})$ almost surely lie in a nonrandom compact subset Γ of $\Re^{p+q}$, such that $P\left(\sum_{t=1}^{n} \boldsymbol{\varsigma}_{t-1}\boldsymbol{\varsigma}_{t-1}^{\top} > 0\right) = 1$. Also $\boldsymbol{\varsigma}_{t-1}^{\top}\boldsymbol{\theta}$ lies almost surely in the domain D of the link inverse link function $h = g^{-1}\ \forall\ \boldsymbol{\varsigma}_{t-1} \in \Gamma$ and $\boldsymbol{\theta} \in B$.

A3. The inverse link function h-defined in (A2) is twice differentiable and $\partial h(\tau)/\partial\tau \neq 0$.

A4. There is a probability measure ν on $\Re^k$ such that $\int_{\Re^k} \boldsymbol{\varsigma}\boldsymbol{\varsigma}^{\top}\nu(d\boldsymbol{\varsigma})$ is positive definite and for Borel sets $A \subset \Re^k$, $1/n\sum_{t=1}^{n} I_{[\boldsymbol{\varsigma}_{t-1}\in A]} \to \nu(A)$ in probability as $n \to \infty$, at the true value of $\boldsymbol{\theta}$.

The assumptions A1 together with A3 guarantee that the second derivative of the log-partial likelihood is continuous with respect to $\boldsymbol{\theta}$. In addition, the condition $h(\tau)/\partial\tau \neq 0$ together with A2, assuming n is large, implies that the conditional information matrix is positive definite with probability 1.

Results of the Estimates of Kurtosis, Skewness and MSE of the ZIP Distribution

Table 4.1: Estimates of the kurtosis, skewness and mean square error (MSE) of the estimated parameters based on 1000 iterations, $n = 50$ and ZIP distribution $\lambda = 2.5$ and $\omega = 0.3$.

Measure	λ_{mme}	ω_{mme}	λ_{mle}	ω_{mle}
Kurtosis	2.9022	3.1227	3.0337	3.0416
Skewness	0.0092	-0.1693	-0.0057	-0.0358
MSE	0.2605	0.0172	0.1953	0.0121

Table 4.2: Estimates of the kurtosis, skewness and mean square error (MSE) of the estimated parameters based on 1000 iterations, $n = 100$ and ZIP distribution $\lambda = 2.5$ and $\omega = 0.3$.

Measure	λ_{mme}	ω_{mme}	λ_{mle}	ω_{mle}
Kurtosis	2.8855	2.9064	2.9618	2.7730
Skewness	0.2083	0.0315	0.0862	0.1049
MSE	0.1204	0.0085	0.0869	0.0060

Table 4.3: Estimates of the kurtosis, skewness and mean square error (MSE) of the estimated parameters based on 1000 iterations, $n = 250$ and ZIP distribution $\lambda = 2.5$ and $\omega = 0.3$.

Measure	λ_{mme}	ω_{mme}	λ_{mle}	ω_{mle}
Kurtosis	2.901	2.8573	2.8681	2.7031
Skewness	0.0986	0.0389	0.0125	0.0816
MSE	0.0501	0.0033	0.0363	0.0022

Results of the ZIACP Modelling Procedure

Table 4.4: Results of simulation for $ZIACP(1)$ model for (ω,γ_0,α_1)=(0.5,2,0.5).

Parameters	Sample size	MLE	MADE	MSE
ω	200	0.4998	0.0331	0.0034
γ_0		2.2200	0.2652	0.1625
α_1		0.5186	0.1622	0.0840
ω	500	0.5002	0.0207	0.0014
γ_0		2.2159	0.2296	0.0968
α_1		0.5320	0.1114	0.0375
ω	1000	0.5003	0.0147	0.0007
γ_0		2.2051	0.2069	0.0660
α_1		0.5477	0.0850	0.0199

Table 4.5: Results of simulation for $ZIACP(1)$ model for (ω,γ_0,α_1)=(0.2,1,0.4).

Parameters	Sample size	MLE	MADE	MSE
ω	200	0.2953	0.0955	0.0127
γ_0		1.3670	0.3678	0.1777
α_1		0.5179	0.1501	0.0555
ω	500	0.2941	0.0941	0.0105
γ_0		1.3548	0.3548	0.1446
α_1		0.5280	0.1347	0.0336
ω	1000	0.2946	0.0946	0.0098
γ_0		1.3532	0.3532	0.1340
α_1		0.5339	0.1347	0.0266

Table 4.6: Results of simulation for $ZIACP(2)$ model for (ω,γ_0,α_1,α_2)=(0.4,2,0.3,0.1).

Parameters	Sample size	MLE	MADE	MSE
ω	200	0.4057	0.0309	0.0031
γ_0		2.2268	0.2953	0.2267
α_1		0.3073	0.1185	0.0445
α_2		0.0854	0.1059	0.0831
ω	500	0.4090	0.0208	0.0013
γ_0		2.2093	0.2345	0.1133
α_1		0.3187	0.0771	0.0184
α_2		0.0956	0.0693	0.0643
ω	1000	0.4083	0.0157	0.0007
γ_0		2.1936	0.2016	0.0721
α_1		0.3209	0.0560	0.0096
α_2		0.1044	0.0488	0.0571

Table 4.7: Results of simulation for $ZIACP(2)$ model for (ω,γ_0,α_1,α_2)=(0.6,3,0.2,0.3).

Parameters	Sample size	MLE	MADE	MSE
ω	200	0.5906	0.0289	0.0024
γ_0		3.1719	0.3710	0.3949
α_1		0.1396	0.1712	0.0846
α_2		0.2753	0.1816	0.1176
ω	500	0.5939	0.0195	0.0012
γ_0		3.1133	0.2247	0.1472
α_1		0.1947	0.1069	0.0333
α_2		0.2833	0.1135	0.0475
ω	1000	0.5960	0.0127	0.0005
γ_0		3.1112	0.1775	0.0831
α_1		0.1953	0.0778	0.0184
α_2		0.2992	0.0744	0.0285

Proof of the Score and Observed Information Matrix of the ZIPA Model

We know that,

$$\frac{\partial \eta_t}{\partial \boldsymbol{\beta}} = \boldsymbol{x}_{t-1},\ \frac{\partial \lambda_t}{\partial \eta_t} = \lambda_t,\ \frac{\partial \omega_t}{\partial \xi_t} = \omega_t(1-\omega_t), \frac{\partial \xi_t}{\partial \boldsymbol{\gamma}} = \boldsymbol{z}_{t-1}.$$

The log partial likelihood of the ZIP model is given by,

$$\begin{aligned}\log \mathcal{L}(\boldsymbol{\theta}) = &\sum_{y_t=0} \log\{\omega_t + (1-\omega_t)exp(-\lambda_t)\} \\ &+ \sum_{y_t>0} \{\log(1-\omega_t) - \lambda_t + y_t \log(\lambda_t) - \log(y_t!)\}.\end{aligned}$$

$$\begin{aligned}\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= \sum_{t=1}^{N} \Bigg\{ \frac{\partial}{\partial \lambda_t} \Big[\mathbf{1}_{\{y_t=0\}} \{\log(\omega_t + (1-\omega_t)exp(-\lambda_t)\} \Big] \\ &+ \frac{\partial}{\partial \lambda_t} \{ \mathbf{1}_{\{y_t>0\}} \Big[\log(1-\omega_t) - \lambda_t + y_t \log(\lambda_t) - \log(y_t!) \Big] \} \Bigg\} \frac{\partial \lambda_t}{\partial \eta_t} \frac{\partial \eta_t}{\partial \boldsymbol{\beta}} \\ &= \sum_{t=1}^{N} \Bigg\{ y_{0,t} \Big(\frac{-(1-\omega_t)exp(-\lambda_t)}{\omega_t + (1-\omega_t)exp(-\lambda_t)} \Big) \lambda_t \boldsymbol{x}_{t-1} + \mathbf{1}_{\{y_t>0\}} (-1 + \frac{y_t}{\lambda_t}) \lambda_t \boldsymbol{x}_{t-1} \Bigg\}.\end{aligned}$$

However, $\mathbf{1}_{\{y_t>0\}} = 1 - \mathbf{1}_{\{y_t=0\}} = 1 - y_{0,t}$. $\implies \mathbf{1}_{\{y_t=0\}} = y_{0,t}$ Hence,

$$\begin{aligned}
\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= \sum_{t=1}^{N} \Bigg\{ -y_{0,t}\Big(\frac{\omega_t + (1-\omega_t)exp(-\lambda_t) - \omega_t}{\omega_t + (1-\omega_t)exp(-\lambda_t)}\Big)\lambda_t \boldsymbol{x}_{t-1} \\
&+ (1-y_{0,t})(y_t - \lambda_t)\boldsymbol{x}_{t-1} \Bigg\} \\
&= \sum_{t=1}^{N} \Bigg\{ -y_{0,t}\Big(1 - \frac{\omega_t}{p_{0,t}}\Big)\lambda_t \boldsymbol{x}_{t-1} + (1-y_{0,t})(y_t - \lambda_t)\boldsymbol{x}_{t-1} \Bigg\},
\end{aligned}$$

where $p_{0,t} = \omega_t + (1-\omega_t)exp(\lambda_t)$ is the probability mass function of $Y_t|\mathcal{F}_{t-1}$ at zero.

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \sum_{t=1}^{N} \Bigg\{ -y_{0,t}\lambda_t \boldsymbol{x}_{t-1} + \frac{y_{0,t}\omega_t\lambda_t\boldsymbol{x}_{t-1}}{p_{0,t}} + (y_t - \lambda_t)\boldsymbol{x}_{t-1} - y_{0,t}y_t + y_{0,t}\lambda_t\boldsymbol{x}_{t-1} \Bigg\}.$$

But the product $y_{0,t}y_t$ is always zero since, when $y_t = 0$ then the product becomes $1 \times 0 = 0$. Also when $y_t > 0$ then the product becomes $0 \times 1 = 0$ Therefore we obtain,

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \sum_{t=1}^{N} \Bigg\{ \frac{y_{0,t}\omega_t\lambda_t}{p_{0,t}} + (y_t - \lambda_t) \Bigg\} \boldsymbol{x}_{t-1}.$$

which simplifies to ,

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \sum_{t=1}^{N} \Bigg\{ y_t - \lambda_t\Big(1 - \frac{\omega_t y_{0,t}}{p_{0,t}}\Big) \Bigg\} \boldsymbol{x}_{t-1}.$$

Also ,

$$
\begin{aligned}
\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}} &= \sum_{t=1}^{N}\Bigg\{\frac{\partial}{\partial \omega_t}[\mathbf{1}_{\{y_t=0\}}\log(\omega_t+(1-\omega_t)exp(-\lambda_t)]\\
&+ \frac{\partial}{\partial \omega_t}\{\mathbf{1}_{\{y_t>0\}}[\log(1-\omega_t)-\lambda_t+y_t\log(\lambda_t)-\log(y_t!)]\}\Bigg\}\frac{\partial \omega_t}{\partial \eta_t}\frac{\partial \eta_t}{\partial \boldsymbol{\gamma}}\\
&= \sum_{t=1}^{N}\Bigg\{y_{0,t}(\frac{(1-exp(-\lambda_t)}{p_{0,t}})+\mathbf{1}_{\{y_t>0\}}(\frac{-1}{1-\omega_t})\Bigg\}\times\omega_t(1-\omega_t)\times\boldsymbol{z}_{t-1}\\
&= \sum_{t=1}^{N}\Bigg\{y_{0,t}\omega_t(1-\omega_t)(\frac{(1-exp(-\lambda_t)}{p_{0,t}})+(1-y_{0,t})(-\omega_t)\Bigg\}\times\boldsymbol{z}_{t-1}\\
&= \sum_{t=1}^{N}\Bigg\{y_{0,t}(1-\omega_t)(\frac{(1-exp(-\lambda_t)}{p_{0,t}})+y_{0,t}-1\Bigg\}\omega_t\boldsymbol{z}_{t-1}\\
&= \sum_{t=1}^{N}\frac{y_{0,t}(1-\omega_t)(1-exp(-\lambda_t))+y_{0,t}(\omega_t+(1-\omega_t)exp(-\lambda_t))}{p_{0,t}}\omega_t\boldsymbol{z}_{t-1}\\
&\quad -\omega_t\boldsymbol{z}_{t-1}\\
&= \sum_{t=1}^{N}\Bigg\{\frac{y_{0,t}(1-\omega_t)+y_{0,t}\omega_t}{p_{0,t}}-1\Bigg\}\omega_t\boldsymbol{z}_{t-1}\\
&= \sum_{t=1}^{N}\Bigg\{\frac{y_{0,t}}{p_{0,t}}-1\Bigg\}\omega_t\boldsymbol{z}_{t-1}.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
S_N(\boldsymbol{\theta}) &= \sum_{t=1}^{N}(\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}+\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}})\\
&= \sum_{t=1}^{N}\Bigg\{(y_t-\lambda_t(1-\frac{\omega_t y_{0,t}}{p_{0,t}}))\boldsymbol{x}_{t-1}+\omega_t(\frac{y_{0,t}}{p_{0,t}}-1)\boldsymbol{z}_{t-1}\Bigg\}. \qquad (4.1)
\end{aligned}
$$

$$d_{11,t}(\boldsymbol{\theta}) = \sum_{t=1}^{N} -\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^2} = \sum_{t=1}^{N} -\frac{\partial(\frac{\partial \log PL(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}})}{\partial \lambda_t}\frac{\partial \lambda_t}{\partial \eta_t}\frac{\partial \eta_t}{\partial \boldsymbol{\beta}}$$

$$= \sum_{t=1}^{N} -\frac{\partial([y_t - \lambda_t(1 - \frac{\omega_t y_{0,t}}{p_{0,t}})]x_{t-1})}{\partial \lambda_t}\frac{\partial \lambda_t}{\partial \eta_t}\frac{\partial \eta_t}{\partial \boldsymbol{\beta}}$$

$$= \sum_{t=1}^{N} \left\{1 - \frac{(\omega_t + (1-\omega_t)exp(-\lambda_t))\omega_t y_{0,t} + \lambda_t \omega_t y_{0,t}(1-\omega_t)exp(-\lambda_t)}{p_{0,t}^2}\right\}$$

$$\times \lambda_t \boldsymbol{x}_{t-1}^{\top}\boldsymbol{x}_{t-1}$$

$$= \sum_{t=1}^{N} \left\{1 - \frac{[\omega_t + (1-\omega_t)exp(-\lambda_t)(1+\lambda_t)]\omega_t y_{0,t}}{p_{0,t}^2}\right\}\lambda_t \boldsymbol{x}_{t-1}^{\top}\boldsymbol{x}_{t-1}.$$

Also ,

$$d_{22,t}(\boldsymbol{\theta}) = \sum_{t=1}^{N} -\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}^2} = \sum_{t=1}^{N} -\frac{\partial([\omega_t(\frac{y_{0,t}}{\omega_t + (1-\omega_t)exp(-\lambda_t)} - 1)]\boldsymbol{z}_{t-1})}{\partial \omega_t}\frac{\partial \omega_t}{\partial \xi_t}\frac{\partial \xi_t}{\partial \boldsymbol{\gamma}}$$

$$= \sum_{t=1}^{N} -\left\{\frac{(\omega_t + (1-\omega_t)exp(-\lambda_t))y_{0,t} - \omega_t y_{0,t}(1 - exp(-\lambda_t))}{p_{0,t}^2} - 1\right\}$$

$$\times \omega_t(1-\omega_t)\boldsymbol{z}_{t-1}^{\top}\boldsymbol{z}_{t-1}$$

$$= \sum_{t=1}^{N} \omega_t(1-\omega_t)\{1 - \frac{y_{0,t}(exp(\lambda_t))}{p_{0,t}^2}\}\boldsymbol{z}_{t-1}^{\top}\boldsymbol{z}_{t-1}.$$

$$d_{12,t}(\boldsymbol{\theta}) = \sum_{t=1}^{N} -\frac{\partial}{\partial \boldsymbol{\gamma}} \left\{ \frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \right\} = \sum_{t=1}^{N} -\frac{\partial([y_t - \lambda_t(1 - \frac{\omega_t y_{0,t}}{p_{0,t}})]x_{t-1})}{\partial \omega_t} \frac{\partial \omega_t}{\partial \xi_t} \frac{\partial \xi_t}{\partial \boldsymbol{\gamma}}$$

$$= \sum_{t=1}^{N} -\omega_t(1-\omega_t) \left\{ \frac{(w_t + (1-\omega_t)exp(-\lambda_t))\lambda_t y_{0,t} - \lambda_t \omega_t y_{0,t}(1 - exp(-\lambda_t))}{p_{0,t}^2} \right\}$$

$$\times \boldsymbol{x}_{t-1}^{\top} \boldsymbol{z}_{t-1}$$

$$= \sum_{t=1}^{N} -\omega_t(1-\omega_t) \{ \frac{(exp(-\lambda_t)\lambda_t y_{0,t}}{p_{0,t}^2} \} \boldsymbol{x}_{t-}^{\top} \boldsymbol{z}_{t-1}.$$

Proof of the Conditional Information Matrix of the ZIPA model

Proof: The conditional information matrix of the ZIPA is given by,

$$G_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} Var\{\boldsymbol{C}_{t-1} v_t(\boldsymbol{\theta}) | \mathcal{F}_t\} = \sum_{t=1}^{n} \boldsymbol{C}_{t-1} \Sigma_t(\boldsymbol{\theta}) \boldsymbol{C}_{t-1}^T,$$

where $\Sigma_t(\boldsymbol{\theta}) = Var\{v_t(\boldsymbol{\theta}) | \mathcal{F}_{t-1}\}$ is a symmetric 2×2 matrix with the elements.

$$\Sigma_t(\boldsymbol{\theta}) = \begin{bmatrix} Var(v_{1,t}(\boldsymbol{\theta})) & Cov(v_{1,t}(\boldsymbol{\theta}), v_{2,t}(\boldsymbol{\theta})) \\ Cov(v_{2,t}(\boldsymbol{\theta}), v_{1,t}(\boldsymbol{\theta})) & Var(v_{1,t}(\boldsymbol{\theta})) \end{bmatrix}.$$

From the score,

$$S_N(\boldsymbol{\theta}) = \sum_{t=1}^{N} (\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} + \frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}})$$

$$= \sum_{t=1}^{N} \left\{ (y_t - \lambda_t(1 - \frac{\omega_t y_{0,t}}{p_{0,t}})) \boldsymbol{x}_{t-1} + \omega_t(\frac{y_{0,t}}{p_{0,t}} - 1) \boldsymbol{z}_{t-1} \right\}.$$

we obtain $v_{1,t}(\boldsymbol{\theta}) = y_t - \lambda_t(1 - \frac{\omega_t y_{0,t}}{p_{0,t}})$ and $v_{2,t}(\boldsymbol{\theta}) = \omega_t(\frac{y_{0,t}}{p_{0,t}} - 1)$

$$\implies Var(v_{1,t}(\boldsymbol{\theta})) = Var(y_t|\mathcal{F}_{t-1}) + \frac{\lambda_t^2\omega_t^2}{p_{0,t}^2}Var(y_{0,t}|\mathcal{F}_{t-1}) + \frac{2\lambda_t\omega_t cov(y_t, y_{0,t})}{p_{0,t}}$$

$$= \lambda_t(1-\omega_t)(1+\lambda_t\omega_t) + \frac{\lambda_t^2\omega_t^2}{p_{0,t}^2}(p_{0,t} - p_{0,t}^2) - 2\lambda_t^2\omega_t(1-\omega_t)$$

$$= \frac{(1-\omega_t)\lambda_t\{exp(-\lambda_t) + \omega_t(1-(1+\lambda_t)exp(-\lambda_t))\}}{p_{0,t}}.$$

Since

$$Var(y_t|\mathcal{F}_{t-1}) = \lambda_t(1-\omega_t)(1+\lambda_t\omega_t)$$

and

$$Var(y_{0,t}|\mathcal{F}_{t-1}) = Var(\mathbf{1}_{\{y_t=0\}}|\mathcal{F}_{t-1}) = E(\mathbf{1}_{\{y_t=0\}})^2 - E^2(\mathbf{1}_{\{y_t=0\}}$$

$$= p_{0,t} - p_{0,t}^2.$$

also

$$Cov(y_t, y_{0,t}) = -\lambda_t(1-\omega_t)p_{0,t}.$$

$$
\begin{aligned}
Var(v_{2,t}(\boldsymbol{\theta})) &= Var(\frac{\omega_t y_{0,t}}{p_{0,t}} - \omega_t | \mathcal{F}_{t-1}) \\
&= \frac{\omega_t^2}{p_{0,t}^2} Var(y_{0,t} | \mathcal{F}_{t-1}) \\
&= \frac{\omega_t^2}{p_{0,t}^2} (p_{0,t} - p_{0,t}^2) \\
&= \frac{\omega_t^2}{p_{0,t}} (1 - p_{0,t}) \\
&= \frac{\omega_t^2}{p_{0,t}} (1 - (\omega_t + (1-\omega_t) exp(-\lambda_t))) \\
&= \frac{\omega_t^2 (1-\omega_t)(1 - exp(-\lambda_t))}{p_{0,t}}.
\end{aligned}
$$

Using the covariance expression below,

$$
Cov(v_{1,t}(\boldsymbol{\theta}), v_{2,t}(\boldsymbol{\theta})) = E(v_{1,t}(\boldsymbol{\theta}) v_{2,t}(\boldsymbol{\theta})) - E(v_{1,t}(\boldsymbol{\theta}))(v_{2,t}(\boldsymbol{\theta})).
$$

But,

$$
\begin{aligned}
v_{1,t}(\boldsymbol{\theta}) v_{2,t}(\boldsymbol{\theta}) &= \frac{y_t \omega_t y_{0,t}}{p_{0,t}} - y_t \omega_t - \frac{\lambda_t \omega_t y_{0,t}}{p_{0,t}} (1 - \frac{\omega_t y_{0,t}}{p_{0,t}}) + \lambda_t \omega_t (1 - \frac{\omega_t y_{0,t}}{p_{0,t}}) \\
&= \frac{y_t \omega_t y_{0,t}}{p_{0,t}} - y_t \omega_t - \frac{\lambda_t \omega_t y_{0,t}}{p_{0,t}} + \frac{\lambda_t \omega_t^2 y_{0,t}^2}{p_{0,t}^2} + \lambda_t \omega_t - \frac{\lambda_t \omega_t^2 y_{0,t}}{p_{0,t}}.
\end{aligned}
$$

Hence

$$
\begin{aligned}
E(v_{1,t}(\boldsymbol{\theta})v_{2,t}(\boldsymbol{\theta})) &= -\omega_t\lambda_t(1-\omega_t) + \frac{\lambda_t\omega_t^2}{p_{0,t}} - \lambda_t\omega_t^2 \\
&= -\omega_t\lambda_t + \frac{\lambda_t\omega_t^2}{p_{0,t}} \\
&= \frac{\lambda_t\omega_t^2 - (\omega_t + (1-\omega_t)exp(-\lambda_t))\omega_t\lambda_t}{p_{0,t}} \\
&= \frac{-\lambda_t\omega_t(1-\omega_t)exp(-\lambda_t)}{p_{0,t}}.
\end{aligned}
$$

since, $E(v_{1,t}(\boldsymbol{\theta})) = 0$, and $E(v_{2,t}(\boldsymbol{\theta})) = 0$. it therefore implies that

$$
Cov(v_{1,t}(\boldsymbol{\theta}), v_{2,t}(\boldsymbol{\theta})) = \frac{-\lambda_t\omega_t(1-\omega_t)exp(-\lambda_t)}{p_{0,t}}.
$$

Bibliography

Ahmed, S., A. Hussein, and P. Sen (2006). Risk comparison of some shrinkage M-estimators in linear models. *Journal of Nonparametric Statistics 18*(4-6), 401–415. (Cited on page 93.)

Ahmed, S. E., S. Hossain, and K. A. Doksum (2012). LASSO and shrinkage estimation in Weibull censored regression models. *Journal of Statistical Planning and Inference 142*(6), 1273—-1284. (Cited on pages 85, 94, 97 and 100.)

Allison, P. D. (2012). *Logistic regression using SAS: Theory and application*. SAS Institute. (Cited on pages 3 and 84.)

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics 31*(3), 307–327. (Cited on pages 1 and 4.)

Cox, D. R., G. Gudmundsson, G. Lindgren, L. Bondesson, E. Harsaae, P. Laake, K. Juselius, and S. L. Lauritzen (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 93–115. (Cited on page 29.)

Creal, D., S. J. Koopman, and A. Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics 28*(5), 777–795. (Cited on page 4.)

Engle, R. F. and J. R. Russell (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 1127–1162. (Cited on page 4.)

Fokianos, K., A. Rahbek, and D. Tjøstheim (2009). Poisson autoregression. *Journal of the American Statistical Association 104*(488), 1430–1439. (Cited on pages i, vi, 1, 5, 28, 29, 30, 32, 35, 36, 55, 63, 76 and 116.)

Ghahramani, M. and A. Thavaneswaran (2009). On some properties of autoregressive conditional poisson (acp) models. *Economics Letters 105*(3), 273–275. (Cited on pages i and 1.)

Haggan, V. and T. Ozaki (1981). Modelling nonlinear random vibrations using an amplitude-dependent autoregressive time series model. *Biometrika 68*(1), 189–196. (Cited on page 31.)

Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics 56*(4), 1030–1039. (Cited on page 87.)

Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*. Number 52. Cambridge University Press. (Cited on pages 2, 5, 8, 9, 10, 12, 13, 14, 16, 17, 23, 25 and 116.)

Hossain, S., K. Doksum, and S. Ahmed (2009). Positive-part shrinkage and absolute penalty estimators in partially linear models. *Linear Algebra and its Applications 430*, 2749–2761. (Cited on pages 85 and 93.)

Ihaka, R. and R. Gentleman (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics 5*(3), 299–314. (Cited on page 108.)

Judge, G. G. and M. E. Bock (1978). The statistical implicatinos of pre-test and stein-rule estimators in econometrics. (Cited on page 93.)

Kedem, B. and K. Fokianos (2005). *Regression models for time series analysis*, Volume 488. John Wiley & Sons. (Cited on pages 12, 27 and 118.)

Kharrati-Kopaei, M. and H. Faghih (2011). Inferences for the inflation parameter in the zip distributions: The method of moments. *Statistical Methodology 8*(4), 377–388. (Cited on pages 37, 46, 48 and 49.)

Kitromilidou, S. and K. Fokianos (2015). Robust estimation methods for a class of log-linear count time series models. *Journal of Statistical Computation and Simulation* (ahead-of-print), 1–16. (Cited on pages i and 78.)

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics 34*(1), 1–14. (Cited on pages 87, 90 and 92.)

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research 16*(3), 285–292. (Cited on page 90.)

McLeish, D. L. and C. G. Small (1988). *The theory and applications of statistical inference functions*. Springer. (Cited on page 38.)

Nanjundan, G. and T. R. Naika (2013). Estimation of parameters in a zero-inflated power series model. (Cited on page 37.)

Nkurunziza, S. and F. Chen (2013). On extension of some identities for the bias and risk functions in elliptically contoured distributions. *Journal of Multivariate Analysis 122*, 190–201. (Cited on page 96.)

Rydberg, T. and N. Shephard (2000). Bin models for trade-by-trade data. *Modelling the number of trades in fixed interval of time. Paper 740*. (Cited on page 28.)

Sprott, D. (1980). Maximum likelihood in small samples: Estimation in the presence of nuisance parameters. *Biometrika 67*(3), 515–523. (Cited on page 38.)

Streett, S. (2000). *Some observation driven models for time series of counts*. Ph. D. thesis, Ph. D. thesis, Colorado State University, Department of Statistics, Fort Collins, Colorado. (Cited on page 28.)

Thomson, T., S. Hossain, and M. Ghahramani (2014). Application of shrinkage estimation in linear regression models with autoregressive errors. *Journal of Statistical Computation and Simulation*, 1–17. (Cited on pages 85 and 96.)

Yang, M. (2012). *Statistical models for count time series with excess zeros*. PhD (Doctor of Philosophy) thesis, University of Iowa, 2012. (Cited on pages i, 55, 71 and 91.)

Zhu, F. (2012). Zero-inflated poisson and negative binomial integer-valued garch models. *Journal of Statistical Planning and Inference 142*(4), 826–839. (Cited on pages i, vi, 55, 64, 65, 66, 67, 74 and 76.)