Clustering Discrete Valued Time Series

CLUSTERING DISCRETE VALUED TIME SERIES

BY

TYLER ROICK, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2017) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE:	Clustering Discrete Valued Time Series
AUTHOR:	Tyler Roick B.Sc., (Mathematics and Statistics) McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Paul D. McNicholas
NUMBER OF PAGES:	x, 48

To my partner in crime, Haley

Abstract

There is a need for the development of models that are able to account for discreteness in data, along with its time series properties and correlation. A review of the application of thinning operators to adapt the ARMA recursion to the integer-valued case is first discussed. A class of integer-valued ARMA (INARMA) models arises from this application. Our focus falls on INteger-valued AutoRegressive (INAR) type models. The INAR type models can be used in conjunction with existing model-based clustering techniques to cluster discrete valued time series data. This approach is then illustrated with the addition of autocorrelations. With the use of a finite mixture model, several existing techniques such as the selection of the number of clusters, estimation using expectation-maximization and model selection are applicable. The proposed model is then demonstrated on real data to illustrate its clustering applications.

Acknowledgements

First and foremost, I wish to extend my most sincere gratitude to my supervisor Dr. Paul McNicholas. His continued support, guidance, and encouragement have lead me to where I am today, and for that I am grateful. Additionally, I would like to thank Dr. Dimitris Karlis for his contributions, continued support, and guidance in my work. I look forward to continuing my research with them in the future.

Secondly, I would like to express my appreciation to Dr. Traian Pirvu and Dr. Ayesha Khan for taking the time to be members of my examining committee.

Finally, to my family and friends, this work would not have been possible without your endless encouragement, patience, and unwavering support. Thank you all for being part of this journey.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In recent years, new types of research problems have presented themselves in the form of data. The type of data being referred to includes seismic activity counts, monitoring system behaviour to detect abnormalities, epileptic seizure data, alcohol drinking patterns, and daily purchases of consumers. These are just a few examples of data that all involve time series of counts. There have been limited ideas discussed in the literature to date that are able to analyze this type of data.

Models are needed to analyze this type of data because normal approximations tend to fail to adequately model time series data with discrete outcomes. Other problems present in this type of data are low count values creating small means, high number of zeros, no symmetry present in the data, and difficult probabilities to compute and interpret.

In this thesis, a new model-based approach to cluster discrete valued time series via a mixture of INAR models is presented. The approach is developed within the framework of model-based clustering by making use of finite mixture models, allowing several existing model-based clustering techniques to be applicable.

In Chapter 2, a review of the previous work that has appeared in current literature is given. This includes background information on mixture models, model-based cluster, and modeling time series of counts which will all be used to develop our methodology in Chapter 3.

In Chapter 3, the framework of our methodology for clustering discrete valued time series via a mixture of INAR models is presented. The implementation of the EM algorithm for parameter estimation, convergence, initialization, model selection, and performance assessment will be covered.

In Chapter 4, our methodology is applied to both simulated and real data sets, and the results of the application are discussed.

In Chapter 5, a summary of the work presented throughout this thesis is given. Thoughts on the direction of future work are considered.

Chapter 2

Background

2.1 Mixture Models and Model-Based Clustering

Model-based clustering is a technique for estimating group memberships, in which no observations are *a priori* labeled, based on parametric finite mixture models. Finite mixture models are based on the assumption that a population is a convex combination of a finite number of densities. A random vector $\mathbf{X}$ is said to arise from a parametric finite mixture distribution if, for all $\mathbf{x} \subset \mathbf{X}$, its density can be written

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x} \mid \boldsymbol{\theta}_g),$$

where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$, is called the gth mixing proportion, $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the gth component density, and $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_g, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the vector of parameters. The component densities $f_1(\mathbf{x} \mid \boldsymbol{\theta}_1), f_2(\mathbf{x} \mid \boldsymbol{\theta}_2), \ldots, f_G(\mathbf{x} \mid \boldsymbol{\theta}_G)$ are usually taken to be of the same type. A more in depth review of finite mixture models can be found in McNicholas (2016a,b).

Next, existing methods for clustering longitudinal data will be discussed. The Cholesky decomposition (Benoît, 1924) is a method used to decompose a matrix into the product of a lower triangular matrix and its transpose. A modified Cholesky decomposition was applied by Pourahmadi (1999, 2000) to the covariance matrix $\boldsymbol{\Sigma}$ of a random variable to obtain,

$$\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}' = \mathbf{D} \Leftrightarrow \boldsymbol{\Sigma}^{-1} = \mathbf{T}'\mathbf{D}^{-1}\mathbf{T},$$

where $\mathbf{T}$ is a unique unit lower triangular matrix and $\mathbf{D}$ is a unique diagonal matrix with strictly positive diagonal entries. A unit lower triangular matrix refers to a lower triangular matrix in which the diagonal elements are all 1. The values of $\mathbf{T}$ can be interpreted as generalized autoregressive parameters, while the values of $\mathbf{D}$ can be interpreted as innovation variances (Pourahmadi, 1999). Further details on the (modified) Cholesky decomposition can be found in McNicholas (2016a, Ch. 8).

McNicholas and Murphy (2010) used a Gaussian mixture model with a modified Cholesky-decomposed covariance structure for each component to model longitudinal data. The gth component density, written for a p-dimensional random variable $\mathbf{X}$, can be found in McNicholas (2016a, Ch. 8). The option of constraining $\mathbf{T}_g$ and/or $\mathbf{D}_g$ to be equal across components together with the option to impose the isotropic constraint $\mathbf{D}_g = \delta_g \mathbf{I}_g$ has given way to a family of eight Gaussian mixtures models, called the Cholesky-decomposed Gaussian mixture model (CDGMM; McNicholas and Murphy, 2010) family. $\mathbf{T}_g$ is a $p \times p$ unit lower triangular matrix and $\mathbf{D}_g$ is a $p \times p$ diagonal matrix, as illustrated previously, following from the modified Cholesky decomposition of $\boldsymbol{\Sigma}_g$. The models of the CDGMM family can be fit with the expectation-maximization (EM) algorithm (Dempster et al., 1977). McNicholas and

Murphy (2010) also considered the cases where elements below a given sub-diagonal of $\mathbf{T}_g$ are equal to zero, thereby, removing autocorrelation over large time lags. Extensive details on the CDGMM family and model fitting can be found in McNicholas (2016a). Moreover, the methodology of McNicholas and Murphy (2010) has recently been extended by Anderlucci and Viroli (2015) to the case where there are multiple responses for each individual at each time point.

A linear model for the component means was considered by McNicholas and Subedi (2012) in which they also use the covariance structures of the CDGMM family. This is done by applying the aforementioned constraints of $\mathbf{T}_g$ and/or $\mathbf{D}_g$ along with the option to impose the isotropic constraint $\mathbf{D}_g = \delta_g \mathbf{I}_g$. The EM algorithm can once again be used for parameter estimation here. McNicholas and Subedi (2012) also considered a t-analogue of the CDGMM family. They develop mixtures of multivariate t-distributions with component scale matrices decomposed as in the CDGMM family, the option of a linear model for the mean, and the option to constrain degrees of freedom to be equal across groups. As with the CDGMM family, the EM algorithm can again be used for parameter estimation. Further details on the use of t-mixtures can be found in McNicholas and Subedi (2012).

2.1.1 EM Algorithm for Model-Based Clustering

The EM algorithm is an iterative procedure used to find maximum likelihood estimates in the case of missing or incomplete data. Each iteration of the EM algorithm involves two steps, the expectation (E) step and the maximization (M) step. The E-step involves computing the expected value(s) of the complete-data log-likelihood, while the M-step maximizes the expected value of the complete-data log-likelihood

with respect to the model parameters. Complete-data refers to the combination of the observed and unobserved data. The iterations of these two steps are repeated until convergence is reached. It is worth noting that Titterington et al. (1985) cite similar approaches to the EM algorithm that were used by Baum et al. (1970), Orchard and Woodbury (1972), and Sundberg (1974).

In a clustering paradigm, the complete-data is comprised of the observed data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ along with the unknown labels $\mathbf{z}_1, \ldots, \mathbf{z}_n$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})$. Here $\mathbf{z}_i$ denotes the group memberships of observation i, where z_{ig} is an indicator variable used to represent whether observation $\mathbf{x}_i$ belongs to group g. The indicator variable can formally be written as

$$z_{ig} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ belongs to component } g \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, n$ and $g = 1, \ldots, G$. The estimation of z_{ig} is the primary objective in terms of model-based clustering.

A well known approach for determining if the EM algorithm has converged is by the use of Aitken's acceleration (Aitken, 1926). The Aitken acceleration procedure estimates the asymptotic maximum log-likelihood at each iteration of the EM algorithm and makes a decision about whether it has converged or not. At iteration k the Aitken acceleration is given by

$$a^{(k)} = \frac{\ell^{(k+1)} - \ell^{(k)}}{\ell^{(k)} - \ell^{(k-1)}},$$

where $\ell^{(k+1)}$, $\ell^{(k)}$, and $\ell^{(k-1)}$ are the log-likelihood values from iterations $k+1$, k, and

$k-1$, respectively. The asymptotic estimate of the log-likelihood (Böhning et al., 1994) at iteration $k+1$ is given by

$$\ell_{\infty}^{(k+1)} = \ell^{(k)} + \frac{1}{1-a^{(k)}}(\ell^{(k+1)} - \ell^{(k)}),$$

where each value is as previously defined. The stopping criterion proposed by Lindsay (1995) suggests that the EM algorithm had converged when

$$\ell_{\infty}^{(k)} - \ell^{(k)} < \epsilon, \tag{2.1}$$

where ϵ is a small value. An alternative stopping criterion was proposed by McNicholas et al. (2010), which suggests that the algorithm had converged when

$$\ell_{\infty}^{(k+1)} - \ell^{(k)} < \epsilon, \tag{2.2}$$

for a small value of ϵ, provided this difference is positive. The only case in which the difference can achieve a negative value is for $a^{(k)} > 1$ which would not be a reasonable place to stop (McNicholas, 2016a). It was shown by McNicholas et al. (2010) that the criterion in (2.2) is equally as strict as (2.1) since $\ell^{(k+1)} \geq \ell^{(k)}$. It was also shown that the criterion in (2.2) is at least as strict as the lack of progress criterion

$$\ell^{(k+1)} - \ell^{(k)} < \epsilon,$$

given by Fraley and Raftery (1998).

2.2 Modeling Time Series of Counts

Time series of counts can be seen as when a number of events or objects per time period are observed over time. There exists cases where the discrete-value of a time series may be large and easily analyzed using a continuous-valued model. Although, this is not always the case as in some counting processes the values of the time series are small numbers. If the latter is the case, then models for stationary real-valued processes such as the autoregressive moving average (ARMA) models will result in the multiplication of an integer by a real number, which commonly results in a noninteger value. The ARMA models are defined by the recursion

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \epsilon_t + \beta_1 \epsilon_{t-1} + \cdots + \beta_q \epsilon_{t-q}, \tag{2.3}$$

where $\alpha_i,\ i = 1, \ldots, p$, are the coefficients of the autoregressive model of order p, $\beta_j,\ j = 1, \ldots, q$, are the coefficients of the moving average model of order q, and $\{\epsilon_t\}$ is a sequence of independent and identically distributed (i.i.d.) random variables, with zero mean and variance σ^2. If (2.3) results in the multiplication of an integer by a real number, then it is prevented from being applied to the integer-valued case. To fix this problem, a new operation, called a thinning operation (Steutel and van Harn, 1979) is proposed. Thinning operations are probabilistic operations and are used to replace the scalar multiplication of equations. Applying the thinning operation to (2.3) would thereby ensure the right hand side to be integer-valued.

The first and most popular (Weiß, 2008) form of thinning operations is known as binomial thinning (Steutel and van Harn, 1979). Binomial thinning was introduced by Steutel and van Harn (1979) to accommodate the terms of "self-decomposability" and

"stability" for integer-valued time series. This becomes important in regards to the INAR(1) process to be discussed following thinning operators. Let "$\circ$" represent the binomial thinning operator defined in Definition 1. Additionally, let $\mu_X = \mathrm{E}[X]$ and $\sigma_X^2 = \mathrm{Var}[X]$, then some basic properties of binomial thinning with proofs provided by Freeland (1998) and da Silva (2005) are as follows:

- $\mathrm{E}[\alpha \circ X] = \alpha\mu_X$
- $\mathrm{Var}[\alpha \circ X] = \alpha^2\sigma_X^2 + \alpha(1-\alpha)\mu_X$
- $\mathrm{Cov}[\alpha \circ X, X] = \alpha\sigma_X^2$

Definition 1. *Let X be a non-negative integer-valued random variable and let $\alpha \in [0,1]$. Define the random variable*

$$\alpha \circ X := \sum_{i=1}^{X} Y_i,$$

where the Y_i are i.i.d. Bernoulli indicators according to B(1, α), which are also independent of X. It can then be said that $\alpha \circ X$ arises from X by binomial thinning.

Binomial thinning can be applied to model autocorrelated processes with several types of marginal distributions. The concept of binomial thinning has since been adapted for other processes. Notable modifications to binomial thinning have resulted in generalized thinning (Latour, 1998), where Y_i are now i.i.d. random variables that have full range $\{0, 1, \ldots\}$ with mean α and variance β. It is worth noting that binomial thinning is a special case of generalized thinning when $\beta = \alpha(1-\alpha)$. Another modification of binomial thinning, called signed binomial thinning (Kim and Park, 2004), allows for the inclusion of negative integers within the range where Y_i are i.i.d.

coming from $\mathrm{B}(1, |\alpha|)$. There also exists extended thinning (Zhu and Joe, 2003), where in comparison to binomial thinning Y_i are now independent count variables that are also independent of X. Binomial thinning is also a special case of extended thinning. There have also been a number of larger extensions to binomial thinning, some of which even utilize the operation itself. Some of these extensions include random coefficient thinning (e.g., Joe, 1996; Zheng et al., 2007), a prime example in which binomial thinning is utilized. Random coefficient thinning is essentially binomial thinning where α is allowed to be random. Special cases of random coefficient thinning include beta-binomial thinning (McKenzie, 1985, 1986; Joe, 1996) where α follows a beta(α, β) distribution and binomial thinning in the case of one-point distribution. Another extension is iterated thinning (Al-Osh and Aly, 1992), which is interpreted as two nested thinning operators. Iterated thinning also reduces to binomial thinning in the case of one-point distribution. Binomial thinning was also extended with the use of the quasi-binomial distribution (Consul and Mittal, 1975; Shenton, 1986). The three parameter quasi-binomial distribution is defined by

$$\mathrm{P}(X = x) = \binom{n}{x} \frac{1 - p + n\psi}{1 - n\psi} p(p + x\psi)^{x-1}(1 - p - x\psi)^{n-x-1}, \tag{2.4}$$

for $0 < p < 1$, $p + n\psi < 1$, and $x = 0, 1, \ldots, n$. The distribution is similar to that of the binomial distribution, but introduces an extra parameter, ψ, that attempts to describe additional variance. A primary benefit of the quasi-binomial distribution is that it approaches the generalized Poisson (GP) distribution as a limit (Consul and Mittal, 1975). The distribution defined in (2.4) is used to define a generalized thinning operation called quasi-binomial thinning (Alzaid and Al-Osh, 1993). For an elegant and extensive summary of the aforementioned thinning operations see Weiß

(2008).

Consider the application of thinning operators to time series with an infinite range of counts. Several models for count data have been proposed based on the application of a thinning operation. Many of these models are obtained as discrete analogues of the usual linear time series models. In particular, replacing the scalar multiplication in the ARMA recursion (2.3) by a binomial thinning operation leads to a family of integer-valued ARMA (INARMA) models. The first proposed INARMA model was the first-order integer-valued autoregressive, INAR(1), process which was proposed by McKenzie (1985, 1988) and Al-Osh and Alzaid (1987) for modeling and generating sequences of dependent counting processes. The INAR model in general mimics the structure and correlation of the linear AR model.

Definition 2. *A discrete time non-negative integer-valued process* $\{X_t\}_{\mathbb{Z}}$ *is said to be a INAR*(1) *process if it satisfies the following recursion*

$$X_t = \alpha \circ X_{t-1} + \epsilon_t,$$

where $\alpha \in [0, 1]$*, "*$\circ$*" represents the binomial thinning operator and* $\{\epsilon_t\}_{\mathbb{Z}}$ *is a sequence of non-negative i.i.d. integer-valued random variables with mean* μ_ϵ *and variance* σ^2_ϵ*. All thinning operations are performed independently of each other and of* $\{\epsilon_t\}_{\mathbb{Z}}$*, and the thinning operations at each time* t *and* ϵ_t *are independent of* $\{X_s\}_{s<t}$*.*

The marginal distribution of X_t can be written in terms of the innovations, ϵ_t (Al-Osh and Alzaid, 1987). The INAR(1) process $\{X_t\}_{\mathbb{Z}}$ is known to be stationary. Let $p_x(z)$ and $p_\epsilon(z)$ denote the marginal probability generating functions (pgf) of X_t and ϵ_t, respectively. Then, using the proof provided by Alzaid and Al-Osh (1988), it

can be shown that the stationary marginal distribution of X_t can be determined from the following equation,

$$p_x(z) \stackrel{!}{=} p_x(1-\alpha+\alpha z)\cdot p_\epsilon(z) \Leftrightarrow p_\epsilon(z) = \frac{p_x(z)}{p_x(1-\alpha+\alpha z)}. \tag{2.5}$$

It suffices to say that the pgf of the INAR(1) process satisfies the definition of a discrete self-decomposable distribution (Steutel and van Harn, 1979). From (2.5), the marginal distribution of an INAR(1) process can be any distribution belonging to the discrete self-decomposable (DSD) family. The DSD family includes the Poisson, GP, and negative binomial (NB) distribution. If the INAR(1) process is used for a stationary process with Poisson marginals it is referred to as Poisson INAR(1). Properties of the Poisson INAR(1) were provided by Freeland (1998), Freeland and McCabe (2004) and Weiß (2007b). It is known that the INAR(1) process is best suited for the case of Poisson marginals, see Weiß (2008). Several other distributions have been considered for the innovations such as binomial, negative binomial, geometric or generalized Poisson (McKenzie, 1986; Alzaid and Al-Osh, 1993; Brännäs, 1993; Berglund and Brännäs, 1999). These models were considered because the innovation distribution drives the properties of the model, such as the allowance of under- and/or overdispersion. Expressions for third-order moments can be found in da Silva and Oliveira (2004, 2005). Extensive details of the stationary INAR(1) process can be found in Al-Osh and Alzaid (1987) and Alzaid and Al-Osh (1988). Some of the more basic properties and proofs provided by Al-Osh and Alzaid (1987) and Alzaid and Al-Osh (1988) are as follows:

- $\mu_X = \dfrac{\mu_\epsilon}{1-\alpha}$

- $\sigma_X^2 = \dfrac{(\alpha\mu_\epsilon + \sigma_\epsilon^2)}{(1-\alpha^2)}$
- $\rho_X(k) = \alpha^k$
- $\mathrm{P}(X_t = k|X_{t-1} = l) = \sum_{j=0}^{min(k,l)} \binom{l}{j} \alpha^j (1-\alpha)^{l-j} P(\epsilon_t = k-j)$
- $\mathrm{E}[X_t|X_{t-1}] = \alpha X_{t-1} + \mu_X \epsilon$

There exists three main approaches to model estimation for a time series where an INAR(1) process is deemed appropriate. Model parameters can be estimated using method of moments or conditional least squares which has been used for the case of the Poisson INAR(1) process. The asymptotic distributions of the previous two estimators for the Poisson case can be found in Freeland and McCabe (2005). The final type of estimation used are maximum likelihood estimates, which can be used because the likelihood function is easily derivable. Extensive details on these approaches can be found in Al-Osh and Alzaid (1987) and Jung et al. (2005) who also provided a comparison of the aforementioned approaches.

The INAR(1) process using the binomial thinning operation was later extended to the p order, known as the INAR(p) process, by Du and Li (1991). The INAR(p) process was also considered using generalized thinning by Gauthier and Latour (1994) and Latour (1998). Du and Li (1991) show the stationarity condition of this INAR(p) process as well as prove that the process is ergodic. The previously mentioned INAR(p) process mimics the second order structure of the well known AR(p) process as mentioned earlier. This INAR(p) process is not the same as the INAR(p) process considered by Alzaid and Al-Osh (1990) in which the second-order structure resembles that of an ARMA($p, p-1$) process. It is worth noting that a separate portrayal of

the INAR(p) process as a p-dimensional INAR(1) process was obtained using vector thinning by Franke and Subba Rao (1995).

Definition 3. *A discrete time non-negative integer-valued process* $\{X_t\}_{\mathbb{Z}}$ *is said to be a INAR*(p) *process if it satisfies the following recursion*

$$X_t = \sum_{i=1}^{p} \alpha_i \bullet X_{t-i} + \epsilon_t, \text{ where } \alpha_i \geq 0 \text{ for } i = 1, \ldots, p-1 \text{ and } \alpha_p > 0,$$

where "$\bullet$*" represents the generalized thinning operator and* $\{\epsilon_t\}_{\mathbb{Z}}$ *is a non-negative sequence of i.i.d. integer-valued random variables with mean* μ_ϵ *and variance* σ_ϵ^2*. The count series,* $\{Y_{j,i}\}$*, of thinning operations* $\alpha_i \bullet X_{t-i} = \sum_{j=0}^{X_{t-i}} Y_{j,i}, i = 1, \ldots, p$*, are mutually independent, and independent of* $\{\epsilon_t\}_{\mathbb{Z}}$*.*

The INAR model has since been extended, and in more recent years, it has been generalized. Generalizations of the INAR model include the generalized INAR(1) process proposed by Zheng et al. (2007) using random coefficient thinning and the corresponding generalization of the INAR(p) process, called the RCINAR(p) process, considered by Zheng et al. (2006). There have also been two proposed cases for the multivariate INAR(p) process. The first case proposed by Franke and Seligmann (1993) for $p = 1$ and the second by Latour (1997) for $p \geq 1$. Extensions of the INAR model will be discussed in the following. It was stated earlier that the INAR(1) process is best suited for the case of Poisson marginals. However, other members of the DSD family were also mentioned, being the cases of NB and GP marginals. While a complex expression for the distribution of the innovations is possible in the case of NB marginals (see Weiß, 2008), it is not possible to obtain in the case of GP marginals. Two cases have been proposed that can be used in the modeling of processes with

NB marginals. The first being the case where α follows a beta distribution, the RCINAR(1) process is easily used to model processes with NB marginals (McKenzie, 1986). The second being the case where Al-Osh and Aly (1992) used iterated thinning in combination with processes having NB marginals. From this, the idea for the iterated thinning INAR(1), known as IINAR(1), process arose. For properties of the IINAR(1) process see Al-Osh and Aly (1992). There also exists a case which is well suited for count variables with a GP distribution. Alzaid and Al-Osh (1993) used quasi-binomial thinning to define a stationary AR(1)-like process with GP marginals. This is where the quasi-binomial INAR(1), known as QINAR(1), process arises from. For properties of the QINAR(1) process see Al-Osh and Aly (1992). Further details on all three of these cases can be found in Weiß (2008).

Consider the application of thinning operators to time series with a finite range of counts. All first-order models that have previously been mentioned can not be applied in this case. To model a process of binomial counts, McKenzie (1985) proposed a modification to the INAR(1) recursion that would still utilize binomial thinning, called the Binomial AR(1) model. A complete explanation along with the interpretation of the binomial AR(1) model can be found in Weiß (2008). Properties of the model can be found in McKenzie (1985) and Weiß (2007a). A separate approach utilizing a new form of thinning, called hypergeometric thinning (Al-Osh and Alzaid, 1991), was also proposed. Al-Osh and Alzaid (1991) applied hypergeometric thinning to defined autocorrelated processes with binomial marginals, leading to the BARMA processes. The original BARMA model was the first-order binomial autoregressive (BAR(1)) process. Properties of the BAR(1) process can be found in Al-Osh and Alzaid (1991) and Weiß (2008).

The ideas discussed here cover only a limited area of the field. Models such as the replicated INAR(p) process, known as the RINAR(p) process (see da Silva, 2005), have also been proposed. Higher order members of the INARMA family also exist, see Jung and Tremayne (2006) for a recent review. Other approaches for modeling time series of count data can be found in Jung et al. (2006).

Chapter 3

Methodology

3.1 The Model

It was previously mentioned that the likelihood of the standard INAR(1) function is easily derivable. This is due to the standard INAR(1) process being a stationary Markov chain. The conditional likelihood of such a model can be written as

$$\mathcal{L}(\boldsymbol{\Theta}) = \prod_{t=2}^{T} P(X_t | x_{t-1}, \boldsymbol{\Theta}), \tag{3.1}$$

where $\boldsymbol{\Theta} = (\alpha, \boldsymbol{\theta})$ refers to the vector of parameters. Here, α refers to the probability of success for binomial thinning and $\boldsymbol{\theta} = (\lambda, \phi)$ are the parameters associated with the distribution of the innovation terms. The parameters λ and ϕ refer to the mean and dispersion of the innovations, respectively. Note that $t = 1$ is excluded from the conditional likelihood as it refers to the distribution of the innovations. Considering the previously given definition of binomial thinning, the conditional distribution of the model can be seen to be a convolution between the binomial distribution and

that of the distribution of the innovation terms. The conditional likelihoods for INAR processes where the observations are related at higher-order lags are similar to that of equation (3.1). The general conditional likelihood where the observations are related at higher-order lag times, assuming the same structure as the INAR(1) process, can be written as

$$\mathcal{L}(\boldsymbol{\Theta}) = \prod_{t=1}^{s} P(X_t) \prod_{t=s+1}^{T} P(X_t | x_{t-s}, \boldsymbol{\Theta}), \tag{3.2}$$

where the first product of (3.2) corresponds to the distribution of the innovation terms only.

To make the likelihoods of the INAR processes comparable, a finite mixture of them are taken. Although the observations are assumed to have come from an INAR process, they may come from any finite mixture of INAR processes with equal or different orders. The case where each observation may come from a different process is not considered, as this would become computationally cumbersome. The observations are said to have come from a mixture of INAR processes included in the model with a specific probability. That is to say that each individual belongs to a specific INAR process which does not change over time, but the process may have different orders. The finite mixture of likelihoods for the INAR model can be written as

$$\mathcal{L}_i(\boldsymbol{\Theta}) = \sum_{g=1}^{G} \pi_g \mathcal{L}_{ig}(\boldsymbol{\Theta}), \tag{3.3}$$

where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$, are the mixing proportions. In the model, $\mathcal{L}_i(\boldsymbol{\Theta})$ refers to the likelihood of the ith individual and $\mathcal{L}_{ig}(\boldsymbol{\Theta})$ refers to the likelihood of the ith individual coming from the gth process. The likelihood for each individual

is found over time from $t = 1$ to T_i. It is assumed that each INAR process is allowed to differ in terms of order and parameter values. The number of components G is considered to be unknown and will be estimated using the observations. The finite mixture of likelihoods in (3.3) can then be seen to follow a similar structure to the standard definition of a mixture model given previously.

3.2 Model Fitting

Considering that the model follows a similar structure to that of the definition of a finite mixture model, estimation via the EM algorithm is considered. As the focus of this method is for model-based clustering purposes, the scenario in which there are n observations, none of which have known group memberships, is also considered.

At each E-step, until convergence, the component indicator variables are updated using their conditional expected values

$$\hat{z}_{ig} = \frac{\pi_g \mathcal{L}_{ig}(\boldsymbol{\Theta})}{\sum_{g=1}^{G} \pi_g \mathcal{L}_{ig}(\boldsymbol{\Theta})} = \frac{\pi_g \mathcal{L}_{ig}(\boldsymbol{\Theta})}{\mathcal{L}_i(\boldsymbol{\Theta})}. \tag{3.4}$$

In the succeeding M-step, the expected complete-data log-likelihood is maximized with respect to the model parameters. The mixing proportions are first updated

$$\hat{\pi}_g = \frac{n_g}{n},$$

for $g = 1, \ldots, G$, where $n_g = \sum_{i=1}^{n} \hat{z}_{ig}$. The M-step here is not a closed form expression meaning that the model specific parameters can not be calculated in a finite number

of operations. To obtain the model specific parameters, the weighted likelihood

$$\mathcal{L}^g(\boldsymbol{\Theta}) = \sum_{i=1}^{n} z_{ig}\mathcal{L}_{ig}(\boldsymbol{\Theta}),$$

can be maximized via the `optim` function in R. At each successive iteration of the above steps, the likelihood is increased until a set convergence condition is met. To determined if the EM algorithm has converged, Aitken's acceleration is used with the stopping criterion proposed by McNicholas et al. (2010).

3.3 Initialization

For each number of components, G, there must be G initial values given for the parameters of $\boldsymbol{\Theta}$. The objective is to obtain the true values of the model parameters in order to optimize $\hat{z}_{ig}$. The ability to accurately predict starting values for the parameters proves to be heavily dependent on the distribution of the innovations. In the case of equal-dispersion, herein referred to as equidispersion, the innovations are assumed to follow a Poisson distribution. Equidispersion is the result of the parameter ϕ from

$$E[X_i] = \phi Var[X_i],$$

being found to equal 1. Estimation proves to be much faster and more accurate in the case of Poisson distributed innovations as there is one less parameter to consider. In the case of overdispersion, where $\phi > 1$, the innovations are thought to follow a negative binomial distribution. Overdispersion is the result of the variance being larger than the expectation (see Figure 4.6). In negative binomial regression, the distribution tends to be given in terms of its mean, allowing the variance to be

written as

$$Var[X_i] = E[X_i] + \frac{(E[X_i])^2}{\phi},$$

where ϕ denotes the dispersion parameter. It is clear from this that the variance must be larger than the expectation. Note that it is also appropriate to use a gamma-Poisson mixture in which the mean of the Poisson distribution can be thought of as a gamma distributed random variable. Both cases are appropriate as they both introduce an additional free parameter, the dispersion parameter. It is worth noting that the weighted likelihood, $\mathcal{L}^g(\mathbf{\Theta})$, frequently fails to be optimized if dispersion is not accounted for and Poisson innovations are used. Although very rare, the case of under-dispersion is handled similarly.

In all cases, starting values are obtained with the use of k-means clustering. The initial values of the means, λ_g, are thought to be similar to the first group of centers found by k-means. The mixing proportions, π_g, come from the respective cluster sizes which are turned into proportions. For $\phi = 1$, the probability of success, α_g, for the binomial distribution is estimated by minimizing the average of the absolute difference of sums between simulated data and that of the observed data for the clusters found by k-means. This is done using the previously estimated values of λ_g and π_g, respectively. The simulated data that the observed data is compared to is created using the most influential lag time. A similar approach is used in the case of $\phi \neq 1$, although both ϕ_g and α_g must be estimated here. Minimizing the absolute mean of the difference between the observed data and simulated data provides moderately accurate starting values for both. The model proves to be more accurate when used as an iterative approach, meaning that initialization must only be done for the smallest number of components fitted. Subsequent number of components, G, use the maximized

parameter values found when $G-1$ components were fitted and add a new component centered at the mean with a small probability. Agglomerative hierarchical clustering may also be used in a similar fashion for initialization, but has shown to be sensitive depending on the data.

3.4 Model Selection and Performance Assessment

The models for this method are considered to be the possible mixtures of INAR processes. The INAR processes to be included in the mixtures are decided by their respective autocorrelations. For example, in Figure 4.1 the two most influential autocorrelations are of order five and order ten. If these were the only two desired autocorrelations to be included in the model, then any mixture of these two autocorrelations may be used. This means that the possible models are mixtures of the form $G-H$ INAR(5) and H INAR(10), where G is the number of components and $H \leq G$. It is obvious that H is restricted by G as a negative number of INAR processes can not be fitted, but as G increases so does the total possible number of mixtures.

With the use of mixture models, an objective criterion is needed to select the 'best' model. Bayes factors are known to have desirable properties for model selection, but are not evaluated with ease. Instead, the Bayesian information criterion (BIC; Schwarz, 1978) is a crude approximation for the Bayes factor and will be used to select the best model. When comparing two models, the difference in the BIC gives a rough approximation to the logarithm of the Bayes factor assuming equal priors (Kass and Raftery, 1995). Given a model with parameters $\boldsymbol{\Theta}$, the Bayesian information criterion is given by

$$\text{BIC} = 2\ell(\hat{\boldsymbol{\Theta}}) - \rho \log n,$$

where $\ell(\hat{\boldsymbol{\Theta}})$ is the maximized log-likelihood, $\hat{\boldsymbol{\Theta}}$ is the maximum likelihood estimate of $\boldsymbol{\Theta}$, ρ is the number of free parameters, and n is the number of observations. The use of the BIC for model selection is a well known idea in model-based clustering. Justifications for its use can be found in Leroux (1992), Kass and Wasserman (1995), Kass and Raftery (1995), and Keribin (2000).

Although in a real clustering scenario the true group memberships are not known, the effectiveness of the model will still be evaluated through simulated data and data with known group memberships. The model is evaluated using a cross tabulation of the maximum *a posteriori* (MAP) classification of the predicted group memberships and that of the true group memberships. Using the results of the cross tabulation, the performance can be quantified numerically though the use of the adjusted Rand index (ARI; Hubert and Arabie, 1985). The Rand index (Rand, 1971) is based on pairwise agreement, written as

$$\frac{\text{number of pairwise agreements}}{\text{number of pairs}},$$

where a value on $[0, 1]$ is obtained, 1 being perfect class agreement. The Rand index alone does not account for agreement by chance, meaning when predicted group memberships are obtained there is a chance they would be classified correctly by chance. The ARI is used as it corrects the Rand index for agreement by chance and has an expected value of 0 under random classification while still having a value of 1 for perfect classification.

Chapter 4

Illustrations

4.1 Overview

In this chapter, the model developed in Chapter 3 will be applied to both real and simulated data sets. Two simulated data analyses and one real data analysis will be carried out. The two simulated data reflect the different aspects covered throughout Chapter 3 in regards to equidispersion and overdispersion. For simplicity in the analyses, only the two most influential INAR processes will be considered in the models. The INAR processes to be included in the model will be decided by the most influential autocorrelations at a multitude of different lag times. We will also only consider three possible models in each analysis. Due to two INAR processes and three models being considered, $G = 1$ components will not be fitted. This is done for consistency purposes while following the iterative approach mentioned previously.

The simulated analyses will be carried out with multiple trials of increasing difficulty. To increase the difficulty in clustering, the parameters of the simulated data will converge together in order to bring the clusters closer and create more overlay.

Both simulated data analyses will be done in a clustering fashion such that the true group memberships of the data will be taken as unknown. This allows us to assess the performance and classification accuracy using the ARI.

4.2 Simulated Data Analyses

4.2.1 Poisson Innovation Simulated Data

INAR data with Poisson distributed innovations are simulated with increasing difficulty. The difficulty is increased in each of three simulations by allowing the parameters to converge and create more overlay between clusters. The true parameters along with the mixing proportions of the three components in each simulation can be found in Table 4.1. In this case, $15,000$ three-component observations are simulated. The dimensions of the simulated data are for 300 individuals over times $t = 1, \ldots, 50$.

Exploring the simulated data, it can be seen from Figure 4.1 that the autocorrelations of all three simulated data are very similar. From the box plots of the autocorrelations only INAR processes of order five and order ten will be considered in the model. Because the data have been simulated for Poisson distributed innovations, it can be seen from Figure 4.2 that the dispersion of the data follow along the Poisson line, where $\phi = 1$. Figure 4.3 shows the simulated data as it would be known in a true clustering scenario along with the true group memberships of the respective clustering difficulty to provide a comparison for Figure 4.4.

For each of three cluster difficulties, $G = 2, \ldots, 5$ components are fit using k-means starting values. The results of each trial can be seen in Table 4.1 along with corresponding MAP classifications. The BIC correctly selects $G = 3$ components using

a mixture of three INAR(5) and zero INAR(10) as the best model for all clustering difficulties. Figure 4.4 shows the estimated group memberships of each clustering scenario and the cluster profiles of the estimated group memberships. The estimated parameters appear to be very close to the true parameters with all clustering difficulties (Table 4.1). In the most difficult clustering scenario an ARI of 0.882 is achieved with a misclassification rate of 5.00% which are both extremely good values for such a difficult problem.

Table 4.1: Clustering results for the easy, moderate, and difficult simulated INAR data with Poisson distributed innovations.

Clustering Difficulty	True Parameters	Estimated Parameters	ARI	Classification Table
Easy	$(\alpha_1, \pi_1, \lambda_1, \phi_1) = (0.40, 0.333, 7.00, 1)$ $(\alpha_2, \pi_2, \lambda_2, \phi_2) = (0.50, 0.250, 4.00, 1)$ $(\alpha_3, \pi_3, \lambda_3, \phi_3) = (0.70, 0.417, 0.50, 1)$	$(\hat{\alpha}_1, \hat{\pi}_1, \hat{\lambda}_1, \hat{\phi}_1) = (0.4, 0.336, 6.98, 1)$ $(\hat{\alpha}_2, \hat{\pi}_2, \hat{\lambda}_2, \hat{\phi}_2) = (0.54, 0.247, 3.78, 1)$ $(\hat{\alpha}_3, \hat{\pi}_3, \hat{\lambda}_3, \hat{\phi}_3) = (0.68, 0.417, 0.58, 1)$	0.991	1: 100, 0, 0 2: 1, 74, 0 3: 0, 0, 125
Moderate	$(\alpha_1, \pi_1, \lambda_1, \phi_1) = (0.40, 0.333, 6.00, 1)$ $(\alpha_2, \pi_2, \lambda_2, \phi_2) = (0.50, 0.250, 4.00, 1)$ $(\alpha_3, \pi_3, \lambda_3, \phi_3) = (0.60, 0.417, 2.00, 1)$	$(\hat{\alpha}_1, \hat{\pi}_1, \hat{\lambda}_1, \hat{\phi}_1) = (0.40, 0.336, 5.99, 1)$ $(\hat{\alpha}_2, \hat{\pi}_2, \hat{\lambda}_2, \hat{\phi}_2) = (0.50, 0.245, 4.01, 1)$ $(\hat{\alpha}_3, \hat{\pi}_3, \hat{\lambda}_3, \hat{\phi}_3) = (0.60, 0.418, 2.00, 1)$	0.949	1: 98, 2, 0 2: 3, 71, 1 3: 0, 0, 125
Difficult	$(\alpha_1, \pi_1, \lambda_1, \phi_1) = (0.40, 0.333, 5.50, 1)$ $(\alpha_2, \pi_2, \lambda_2, \phi_2) = (0.50, 0.250, 4.00, 1)$ $(\alpha_3, \pi_3, \lambda_3, \phi_3) = (0.60, 0.417, 2.00, 1)$	$(\hat{\alpha}_1, \hat{\pi}_1, \hat{\lambda}_1, \hat{\phi}_1) = (0.40, 0.336, 5.51, 1)$ $(\hat{\alpha}_2, \hat{\pi}_2, \hat{\lambda}_2, \hat{\phi}_2) = (0.50, 0.250, 4.00, 1)$ $(\hat{\alpha}_3, \hat{\pi}_3, \hat{\lambda}_3, \hat{\phi}_3) = (0.6, 0.417, 1.98, 1)$	0.882	1: 94, 6, 0 2: 9, 66, 0 3: 0, 0, 125

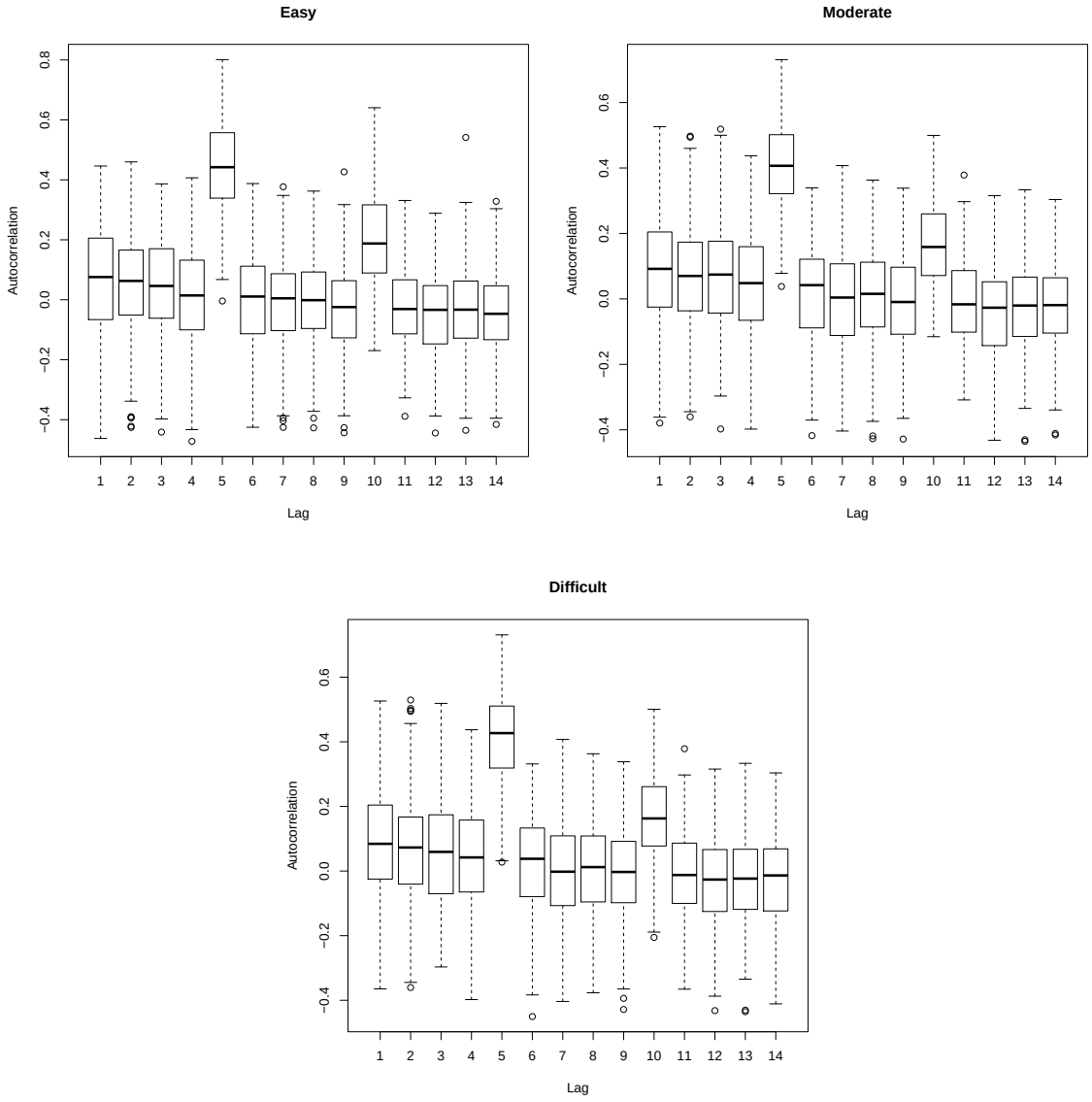

Figure 4.1: Box plots of the autocorrelation at multiple lag times for the easy, moderate, and difficult simulated INAR data with Poisson distributed innovations, respectively.

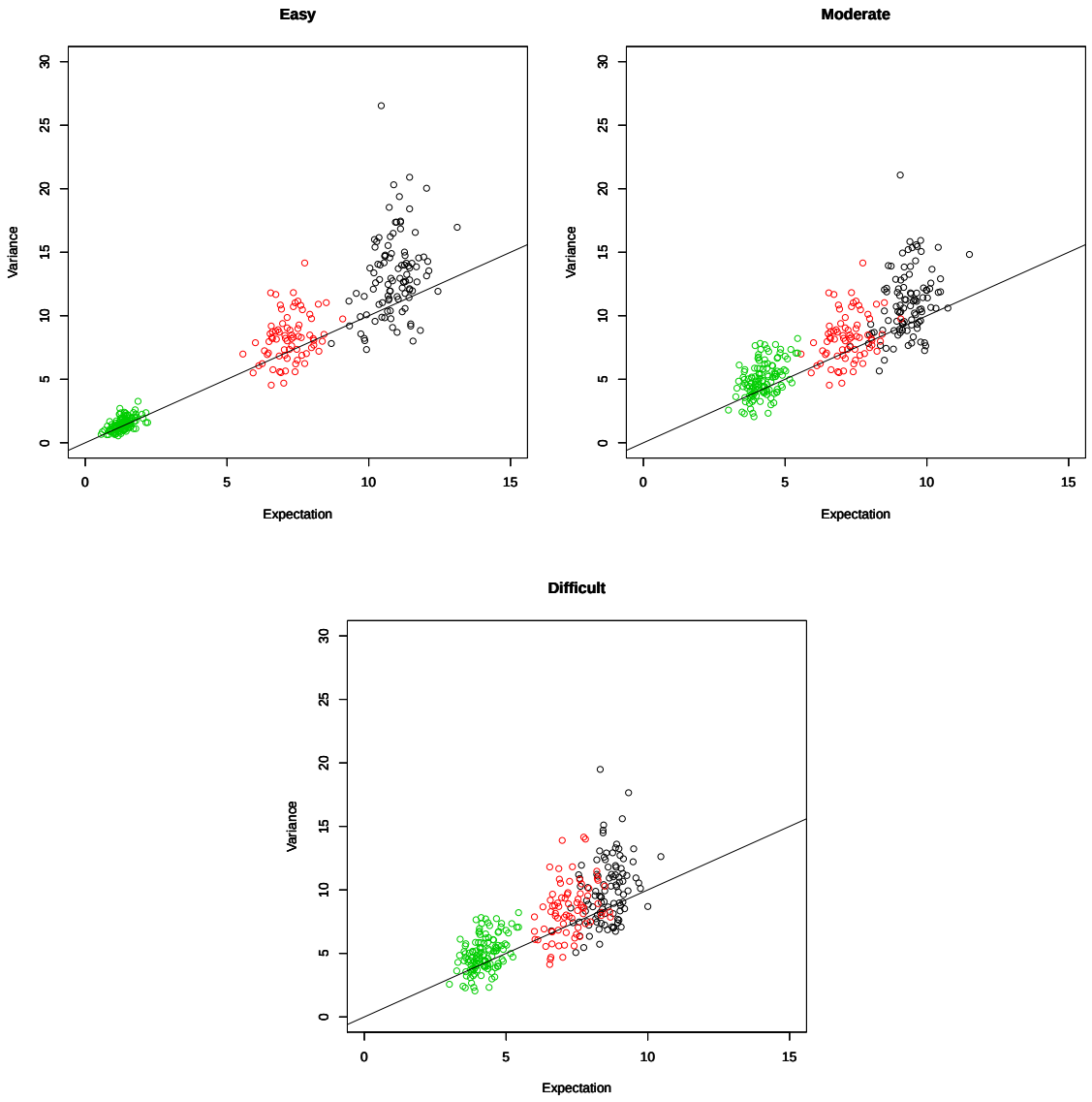

Figure 4.2: Dispersion of the easy, moderate, and difficult simulated INAR data with Poisson distributed innovations, respectively.

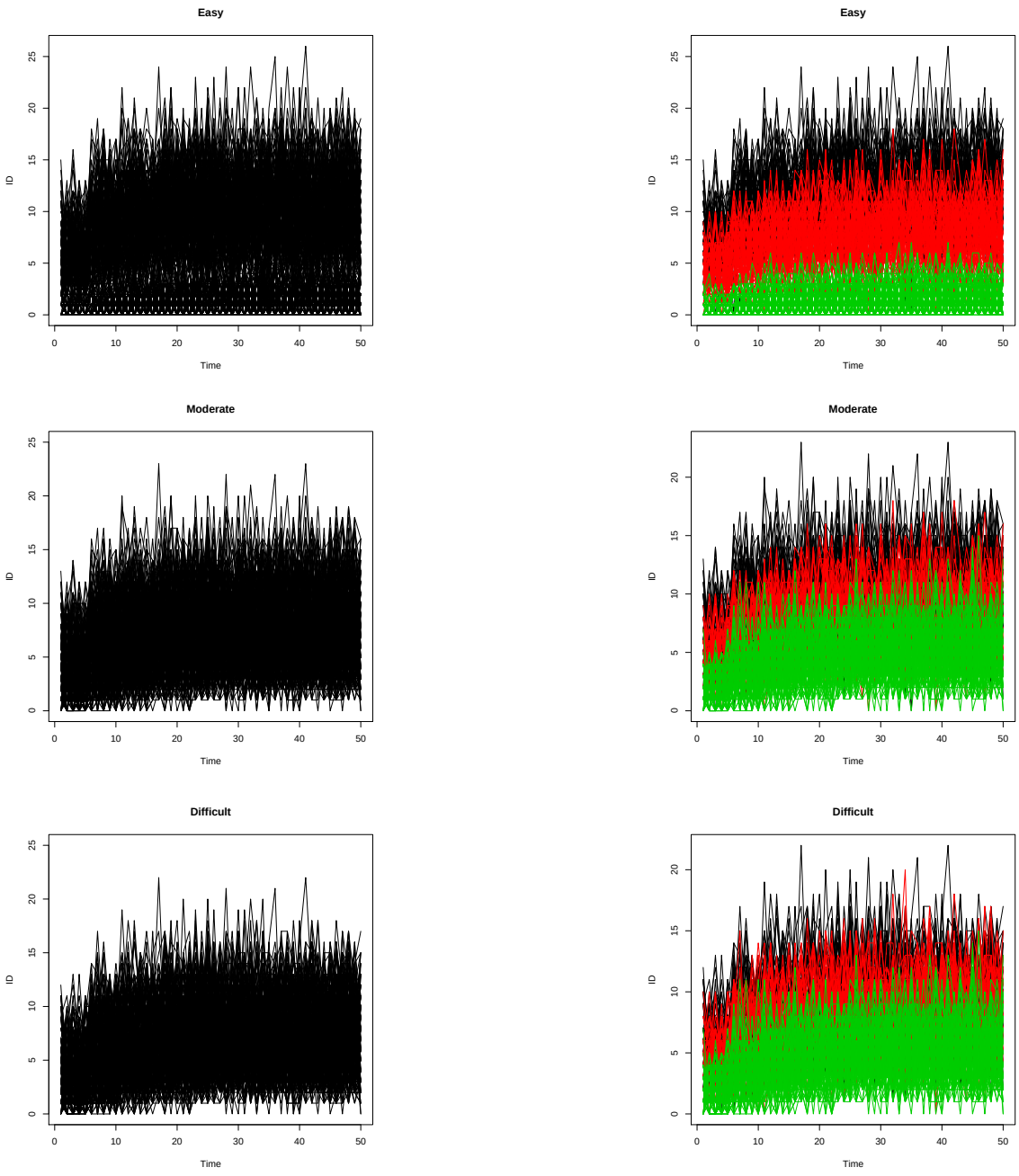

Figure 4.3: Plots of the data with unknown group memberships and the true group memberships for the easy, moderate, and difficult simulated INAR data with Poisson distributed innovations, respectively.

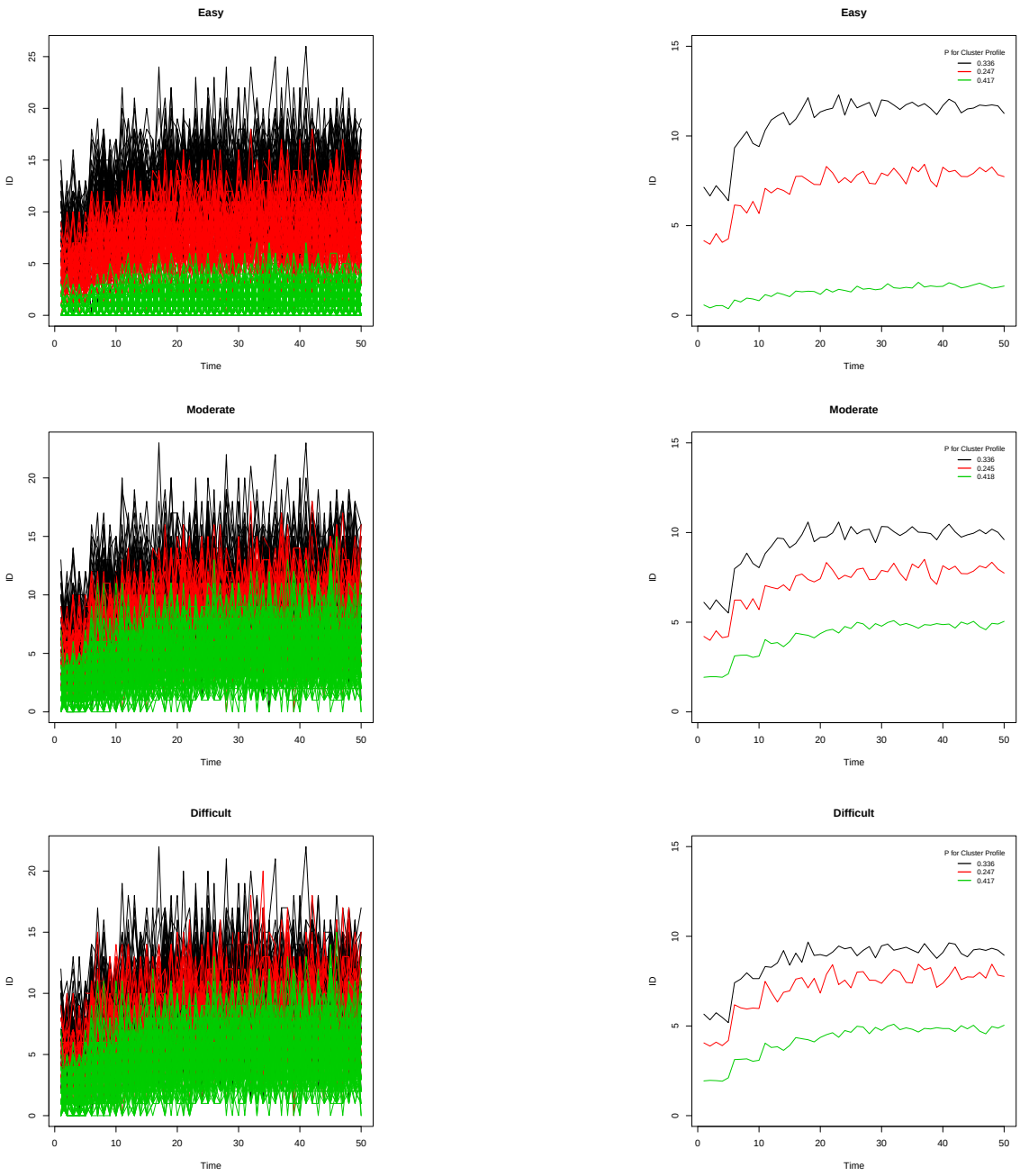

Figure 4.4: Plots of the estimated group memberships and cluster profiles for the easy, moderate, and difficult simulated INAR data with Poisson distributed innovations, respectively.

4.2.2 Negative Binomial Innovation Simulated Data

Following in a similar fashion to the previous section, INAR data with negative binomial distributed innovations are simulated with increasing difficulty. The difficulty is increased in each of three simulations by allowing the parameters to converge and create more overlay between clusters. The true parameters along with the mixing proportions of the three components in each simulation can be found in Table 4.2. In this case, $12,000$ two-component observations are simulated. The dimensions of the simulated data are for 400 individuals over times $t = 1, \ldots, 30$.

Exploring the simulated data, it can be seen from Figure 4.5 that the autocorrelations of all three simulated data are very similar. From the box plots of the autocorrelations only INAR processes of order two and order four will be considered in the model. Because the data have been simulated for negative binomial distributed innovations, it can be seen from Figure 4.6 that the dispersion of the data mainly lies above the Poisson line, thus simulating overdispersion. Figure 4.7 shows the simulated data as it would be known in a true clustering scenario along with the true group memberships of the respective clustering difficulty to provide a comparison for Figure 4.8.

For each of three clustering difficulties, $G = 2, \ldots, 4$ components are fit using k-means starting values. The results of each trial can be seen in Table 4.2 along with corresponding MAP classifications. The BIC correctly selects $G = 2$ components using a mixture of two INAR(2) and zero INAR(4) as the best model for all clustering difficulties. Figure 4.8 shows the estimated group memberships of each clustering scenario and the cluster profiles of the estimated group memberships. The estimated parameters appear to be very close to the values of the true parameters with all

clustering difficulties (Table 4.2). In the most difficult clustering scenario an ARI of 0.730 is achieved with a misclassification rate of 7.25% which are both very reasonable values for such a difficult problem.

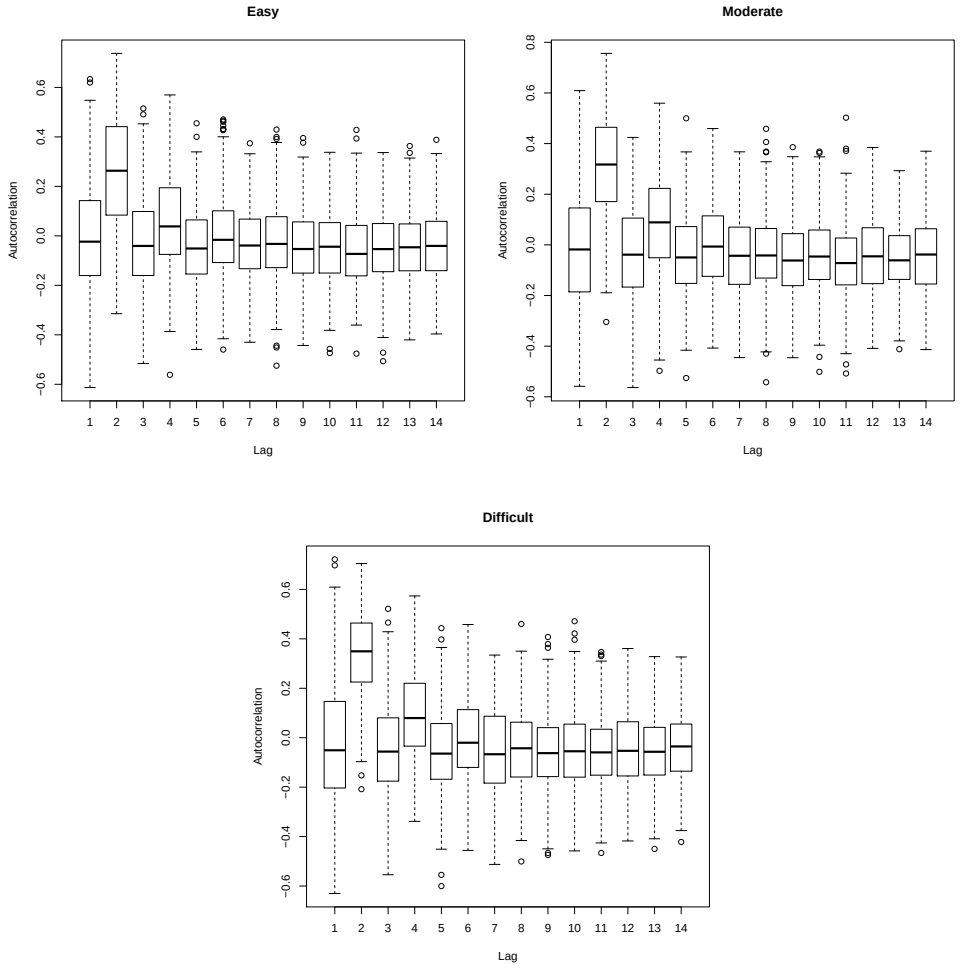

Figure 4.5: Box plots of the autocorrelation at multiple lag times for the easy, moderate, and difficult simulated INAR data with negative binomial distributed innovations, respectively.

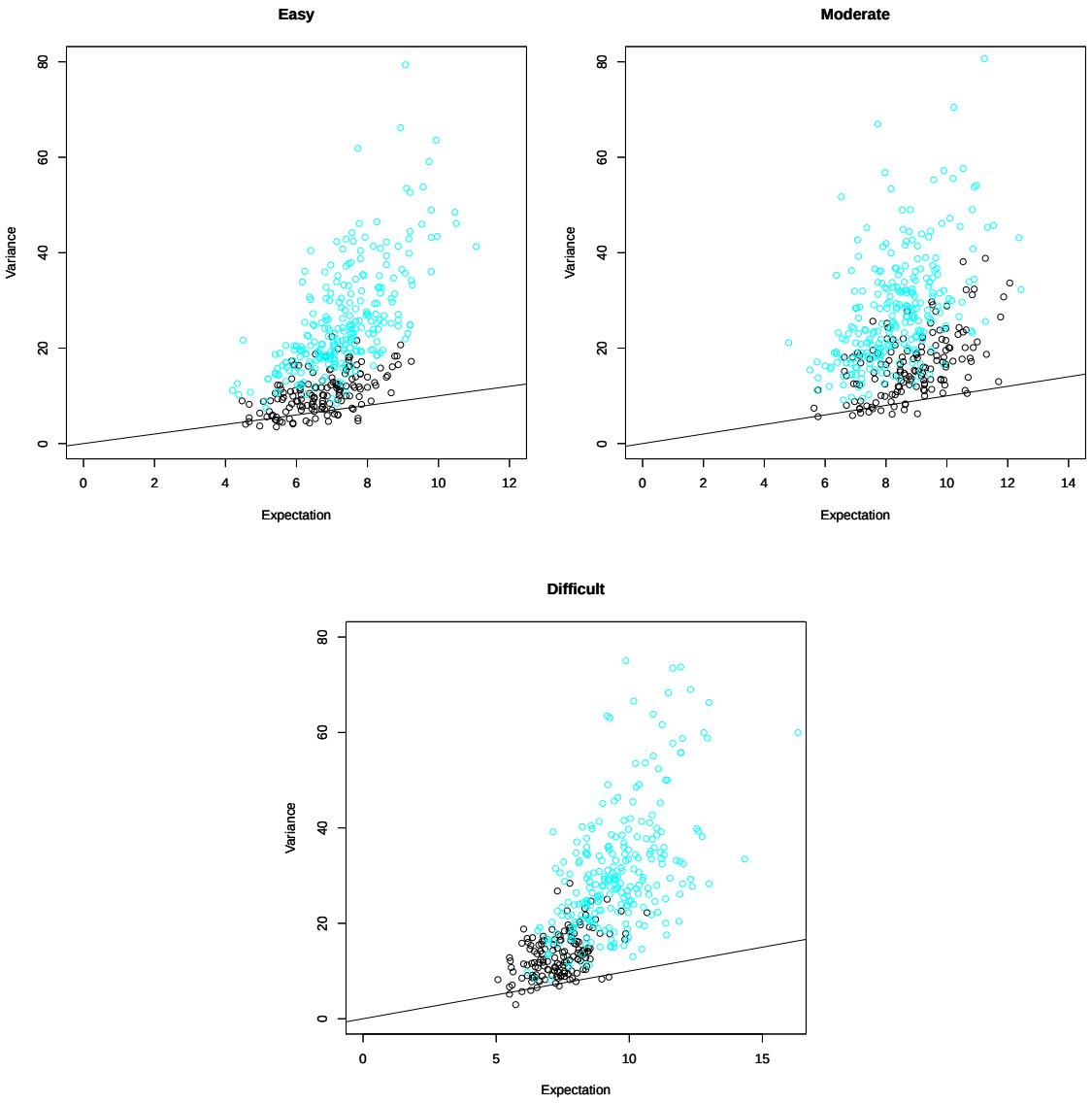

Figure 4.6: Dispersion of the easy, moderate, and difficult simulated INAR data with negative binomial distributed innovations, respectively.

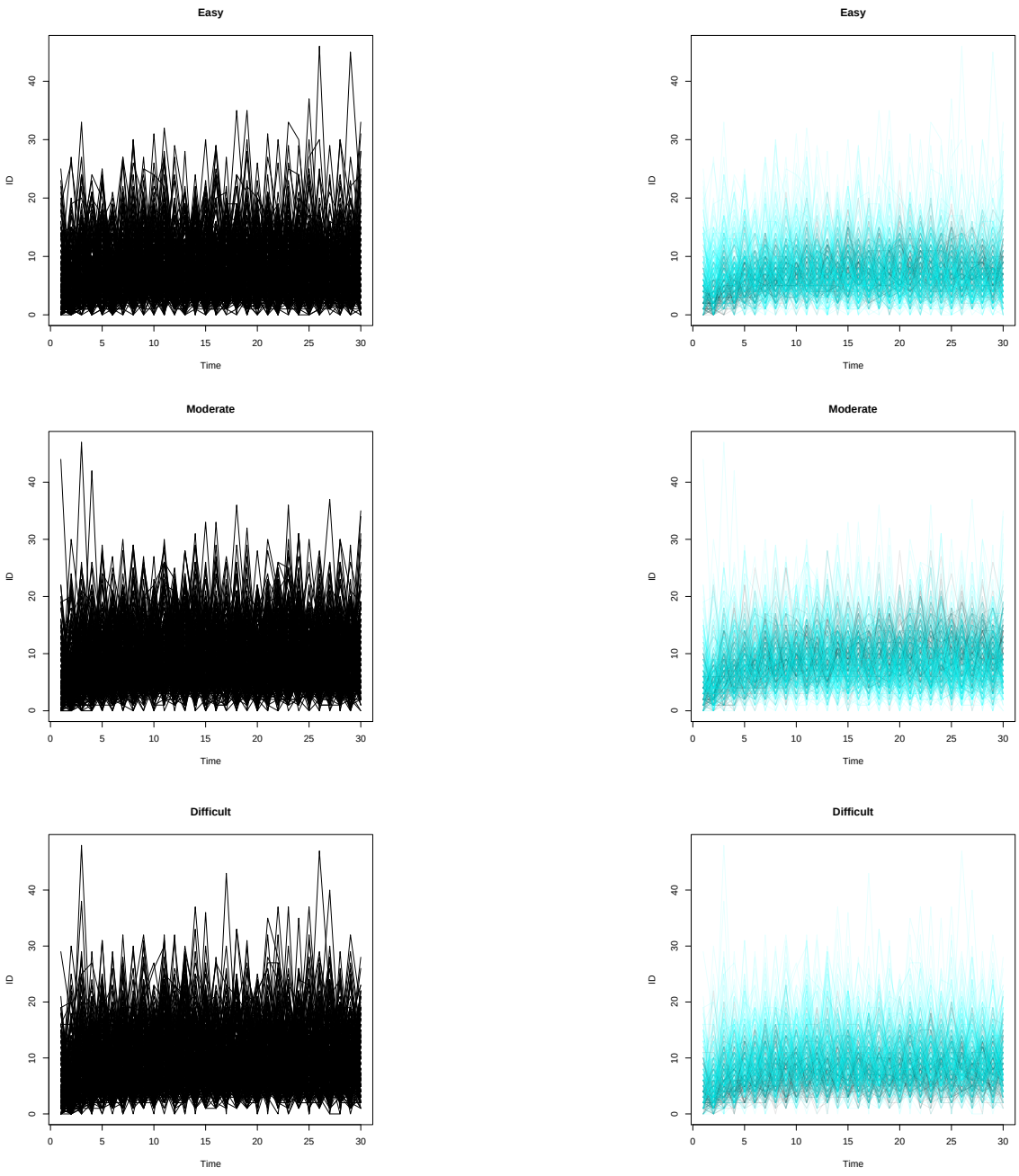

Figure 4.7: Plots of the data with unknown group memberships and the true group memberships for the easy, moderate, and difficult simulated INAR data with negative binomial distributed innovations, respectively.

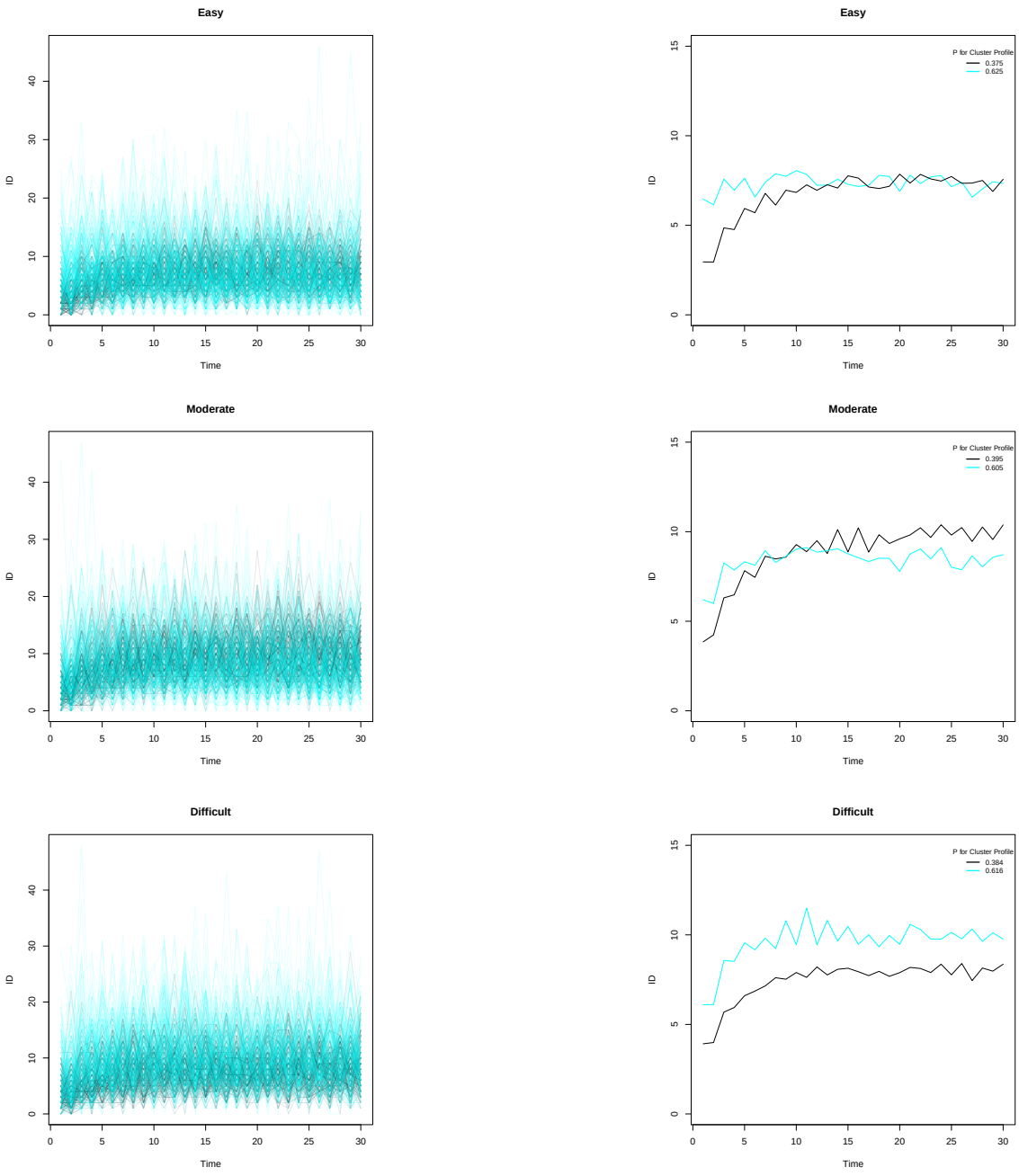

Figure 4.8: Plots of the estimated group memberships and cluster profiles for the easy, moderate, and difficult simulated INAR data with negative binomial distributed innovations, respectively.

Table 4.2: Clustering results for the easy, moderate, and difficult simulated INAR data with negative binomial distributed innovations.

Clustering Difficulty	True Parameters	Estimated Parameters	ARI	Classification Table		
Easy	$(\alpha_1, \pi_1, \lambda_1, \phi_1) = (0.60, 0.375, 3.00, 4)$ $(\alpha_2, \pi_2, \lambda_2, \phi_2) = (0.20, 0.625, 6.00, 2)$	$(\hat{\alpha}_1, \hat{\pi}_1, \hat{\lambda}_1, \hat{\phi}_1) = (0.59, 0.375, 3.03, 3.96)$ $(\hat{\alpha}_2, \hat{\pi}_2, \hat{\lambda}_2, \hat{\phi}_2) = (0.21, 0.625, 5.87, 1.92)$	0.970		1	2
				1	149	1
				2	2	248
Moderate	$(\alpha_1, \pi_1, \lambda_1, \phi_1) = (0.60, 0.375, 4.00, 4)$ $(\alpha_2, \pi_2, \lambda_2, \phi_2) = (0.30, 0.625, 6.00, 2)$	$(\hat{\alpha}_1, \hat{\pi}_1, \hat{\lambda}_1, \hat{\phi}_1) = (0.59, 0.395, 4.00, 3.73)$ $(\hat{\alpha}_2, \hat{\pi}_2, \hat{\lambda}_2, \hat{\phi}_2) = (0.31, 0.605, 5.99, 1.89)$	0.883		1	2
				1	149	1
				2	11	239
Difficult	$(\alpha_1, \pi_1, \lambda_1, \phi_1) = (0.50, 0.375, 4.00, 4)$ $(\alpha_2, \pi_2, \lambda_2, \phi_2) = (0.40, 0.625, 6.00, 2)$	$(\hat{\alpha}_1, \hat{\pi}_1, \hat{\lambda}_1, \hat{\phi}_1) = (0.51, 0.385, 3.96, 3.64)$ $(\hat{\alpha}_2, \hat{\pi}_2, \hat{\lambda}_2, \hat{\phi}_2) = (0.41, 0.615, 5.95, 1.88)$	0.730		1	2
				1	138	12
				2	17	233

4.3 Real Data Analyses

4.3.1 Alcohol Timeline Followback Data

The timeline followback (TLFB; Sobell et al., 1986) method is a tool used to assess subjects' daily alcohol consumption. The alcohol TLFB data being considered was presented in Atkins et al. (2013) and comes from a larger study aimed at event specific prevention. The event specific prevention here refers to intensive daily drinking habits around a number of people's twenty-first birthdays. This data also includes extreme drinking events relative to a random sample of students' drinking (Neighbors et al., 2010). Estimates of daily drinking were evaluated for clinical and nonclinical populations; e.g., adolescents, adults, college students, alcoholics of different severity, and normal male and female drinkers in the general population. Using a calendar, subjects provided retrospective estimates of their daily drinking over a specified time

period. The original focus of the assessment was to study the gender, greek status being that the subject is in a fraternity/sorority or neither, and period of the week in which the drinking occurred. Our focus will fall sheerly on the number of drinks and what can be inferred about the clusters found.

The data is composed of 980 individuals who listed their respective number of drinks over a 30 day period. There were 269 individuals who did not finish the study, due to this reason we will only consider the 711 individuals for which the data was fully recorded. Taking a closer look at the data, Figure 4.9a shows box plots of the autocorrelations. From these box plots, only INAR processes of order one and order seven will be considered in the model. It can be seen from Figure 4.9b that overdispersion is present in the TLFB data. Figure 4.9c shows the simulated data as it would be known in a true clustering scenario.

For the alcohol TLFB data, $G = 2, \ldots, 8$ components are fit using k-means starting values. The BIC selects $G = 6$ components using a mixture of four INAR(1) and two INAR(7). Figure 4.9d shows the estimated group memberships of the TLFB data and Figure 4.9e shows the respective cluster profiles of the estimated group memberships. From the six cluster profiles present in Figure 4.9e, there seems to be individuals on very extreme ends of the spectrum. The red profile appears to be individuals who drank at a specific event and returned to not drinking throughout the remainder of the study. The light blue profile, although very similar to the red profile, appears to be individuals who continued drinking lightly after the specified event. The black, blue, magenta, and green profiles appear to be individuals with heavier drinking habits, but at a variety of different quantities. This could perhaps have to do with the individuals alcohol tolerance level or other social gatherings.

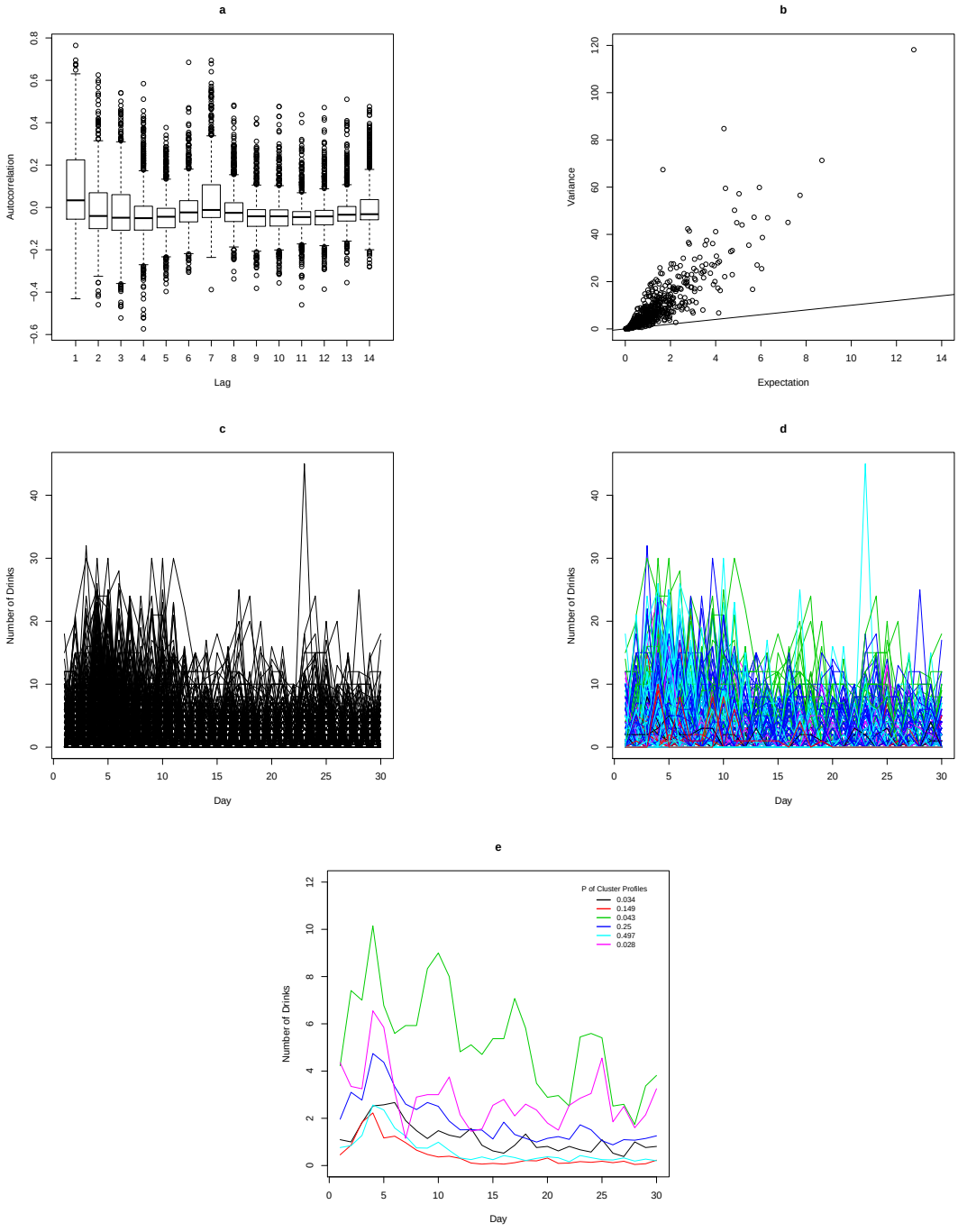

Figure 4.9: Plots of the: a) autocorrelation at multiple lag times, b) dispersion in the data, c) unknown group memberships, d) estimated group memberships, and e) cluster profiles of the estimated group memberships for the alcohol TLFB data.

Chapter 5

Summary and Future Work

5.1 Summary

The current literature on mixture models, model-based clustering, and modeling time series of counts was reviewed. A new model-based approach for clustering discrete valued time series has been introduced. The parameters of the model were estimated using the EM algorithm and a stopping criterion based on Aitken acceleration was used to determine if the model had converged. Model selection was done using the BIC and a performance assessment was carried out using the ARI and misclassification rate in the case of simulated data. The new model-based technique was applied to both simulated and real data to illustrate its clustering capabilities. In the application to simulated data, the technique performed well for a variety of difficulties with both equidispersion and overdispersion present in the data. In the application to real data, a true clustering scenario in which no group memberships were known was analyzed. The technique performed appropriately and reasonable clusters were found for the obscure relationships in the data.

5.2 Future Work

The newly discovered model-based approach for clustering discrete valued time series presents many different directions that could be taken in future work. Some of the more relevant directions to be taken include extending the INAR model to include multivariate time series of counts. Other directions include expanding the model-based approach to include other integer-valued models, e.g., a mixture of INARCH models, and the improvement of computational aspects, e.g., the EM algorithms time consuming maximization step.

Bibliography

Aitken, A. C. (1926). A series formula for the roots of algebraic and transcendental equations. *Proceedings of the Royal Society of Edinburgh 45*, 14–22.

Al-Osh, M. A. and E.-E. A. A. Aly (1992). First order autoregressive time series with negative binomial and geometric marginals. *Communications in Statistics: Theory and Methods 21*(9), 2483–2492.

Al-Osh, M. A. and A. A. Alzaid (1987). First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis 8*(3), 261–275.

Al-Osh, M. A. and A. A. Alzaid (1991). Binomial autoregressive moving average models. *Communications in Statistics: Stochastic Models 7*(2), 261–282.

Alzaid, A. A. and M. A. Al-Osh (1988). First-order integer-valued autoregressive process: distributional and regression properties. *Statistica Neerlandica 42*, 53–61.

Alzaid, A. A. and M. A. Al-Osh (1990). An integer-valued pth-order autoregressive structure (INAR(p)) process. *Journal of Applied Probability 27*, 314–324.

Alzaid, A. A. and M. A. Al-Osh (1993). Generalized Poisson ARMA processes. *Annals of the Institute of Statistical Mathematics 45*(2), 223–232.

Anderlucci, L. and C. Viroli (2015). Covariance pattern mixture models for multivariate longitudinal data. *The Annals of Applied Statistics 9*(2), 777–800.

Atkins, D. C., S. A. Baldwin, C. Zheng, R. J. Gallop, and C. Neighbors (2013). A tutorial on count regression and zero–altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors 27*(1), 166–177.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics 41*, 164–171.

Benoît (1924). Note sur une méthode de résolution des équations normales provenant de l'application de la méthode des moindres carrés à un système d'équations linéires en nombre inférieur celui des inconnues (Procédé du Commandant Cholesky). *Bulletin Géodésique 2*, 67–77.

Berglund, E. and K. Brännäs (1999). Plants' entry and exit in Swedish municipalities. Umeå Economic Studies 497. Umeå University, Sweden.

Böhning, D., E. Dietz, R. Schaub, P. Schlattmann, and B. Lindsay (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics 46*, 373–388.

Brännäs, K. (1993). Estimation and testing in integer-valued AR(1) models. Umeå Economic Studies 335. Umeå University, Sweden.

Consul, P. C. and S. P. Mittal (1975). A new urn model with predetermined strategy. *Biometrical Journal 17*(2), 67–75.

da Silva, I. M. M. (2005). Contributions to the analysis of discrete-valued time series. PhD thesis, University of Porto.

da Silva, M. E. and V. L. Oliveira (2004). Difference equations for the higher-order moments and cumulants of the INAR(1) model. *Journal of Time Series Analysis 25*(3), 317–333.

da Silva, M. E. and V. L. Oliveira (2005). Difference equations for the higher order moments and cumulants of the INAR(p) model. *Journal of Time Series Analysis 26*(1), 17–36.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B 39*(1), 1–38.

Du, J.-G. and Y. Li (1991). The integer-valued autoregressive (INAR(p)) model. *Journal of Time Series Analysis 12*(2), 129–142.

Fraley, C. and A. E. Raftery (1998). How many clusters? Which clustering methods? Answers via model-based cluster analysis. *The Computer Journal 41*(8), 578–588.

Franke, J. and T. Seligmann (1993). Conditional maximum likelihood estimates for INAR(1) processes and their application to modeling epileptic seizure counts. *Developments in Time Series Analysis: in honour of Maurice B. Priestley*, 310–330.

Franke, J. and T. Subba Rao (1995). Multivariate first order integer valued autoregressions. Technical report, Math. Dep., UMIST, England.

Freeland, R. K. (1998). Statistical analysis of discrete time series with applications to the analysis of workers compensation claims data. PhD thesis, University of British Columbia, Canada.

Freeland, R. K. and B. P. M. McCabe (2004). Analysis of low count time series data by Poisson autoregression. *Journal of Time Series Analysis 25*(5), 701–722.

Freeland, R. K. and B. P. M. McCabe (2005). Asymptotic properties of CLS estimators in the Poisson AR(1) model. *Statistics and Probability Letters 73*(2), 147–153.

Gauthier, G. and A. Latour (1994). Convergence forte des estimateurs des paramètres dt'un processus genar(p). *Annales des Sciences Mathématiques du Québec 18*, 49–71.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*(1), 193–218.

Joe, H. (1996). Time series models with univariate margins in the convolution-closed infinitely divisible class. *Journal of Applied Probability 33*, 664–677.

Jung, R. C., M. Kukuk, and R. Liesenfeld (2006). Time series of count data: Modeling, estimation and diagnostics. *Computational Statistics and Data Analysis 51*, 2350–2364.

Jung, R. C., G. Ronning, and A. R. Tremayne (2005). Estimation in conditional first order autoregression with discrete support. *Statistical Papers 46*(2), 195–224.

Jung, R. C. and A. R. Tremayne (2006). Binomial thinning models for integer time series. *Statistical Modelling: An International Journal 6*(2), 81–96.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Kass, R. E. and L. Wasserman (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association 90*(431), 928–934.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā. The Indian Journal of Statistics. Series A 62*(1), 49–66.

Kim, H. Y. and Y. S. Park (2004). A non-stationary integer-valued autoregressive model. In: Proc. of the Spring Conference. *Korean Statistical Society*, 193–199.

Latour, A. (1997). The multivariate GINAR(p) process. *Advances in Applied Probability 29*, 228–248.

Latour, A. (1998). Existence and stochastic structure of a non-negative integer-valued autoregressive process. *Journal of Time Series Analysis 19*(4), 439–455.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics 20*(3), 1350–1360.

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, Volume 5. California: Institute of Mathematical Statistics: Hayward.

McKenzie, E. (1985). Some simple models for discrete variate time series. *Water Resources Bulletin 21*(4), 645–650.

McKenzie, E. (1986). Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Advances in Applied Probability 18*, 679–705.

McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts. *Advances in Applied Probability 20*, 822–835.

McNicholas, P. D. (2016a). *Mixture model-based classification*. CRC Press, Taylor & Francis Group.

McNicholas, P. D. (2016b). Model-based clustering. *Journal of Classification 33*(3), 331–373.

McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of longitudinal data. *The Canadian Journal of Statistics 18*(3), 285–296.

McNicholas, P. D., T. B. Murphy, A. F. McDaid, and D. Frost (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis 54*(3), 711–723.

McNicholas, P. D. and S. Subedi (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference 142*(5), 1114–1127.

Neighbors, C., M. A. Lewis, D. C. Atkins, M. M. Jensen, T. Walter, N. Fossos, C. M. Lee, and M. E. Larimer (2010). Efficacy of web–based personalized normative feedback: A two-year randomized controlled trial. *Journal of Consulting and Clinical Psychology 78*(6), 898–911.

Orchard, T. and M. A. Woodbury (1972). A missing information principle: Theory and applications. In L. M. Le Cam, J. Neyman, and E. L. Scott (Eds.), *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*, pp. 697–715. Berkeley: University of California Press.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika 86*(3), 677–690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika 87*(2), 425–435.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*(336), 846–850.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Shenton, L. R. (1986). Quasibinomial distributions. *In: Kotz/Johnson: Encyclopedia of Statistical Sciences 7*, 458–460.

Sobell, M. B., L. C. Sobell, F. Klajner, D. Pavan, and E. Basian (1986). The reliability of a timeline method for assessing normal drinker college students' recent drinking history: Utility for alcohol research. *Addictive Behaviors 11*(2), 149–161.

Steutel, F. W. and K. van Harn (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability 7*, 893–899.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics 1*(2), 49–58.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley & Sons.

Weiß, C. H. (2007a). Controlling correlated processes with binomial marginals. Preprint 277, Mathematische Institute der Julius-Maximilians-Universität Würzburg.

Weiß, C. H. (2007b). Serial dependence and regression of Poisson INARMA models. *Journal of Statistical Planning and Inference*.

Weiß, C. H. (2008). Thinning operations for modeling time series of counts — A survey. *ASTA-Advances in Statistical Analysis 92*(2), 319–341.

Zheng, H., I. V. Basawa, and S. Datta (2006). Inference for pth-order random coefficient integer-valued autoregressive processes. *Journal of Time Series Analysis 27*(3), 411–440.

Zheng, H., I. V. Basawa, and S. Datta (2007). First-order random coefficient integer-valued autoregressive processes. *Journal of Statistical Planning and Inference 137*(1), 212–229.

Zhu, R. and H. Joe (2003). A new type of discrete self-decomposability and its application to continuous-time Markov processes for modeling count data time series. *Stochastic Models 19*(2), 235–254.