Generalized Linear Autoregressive Moving Average Models

William T.M. Dunsmuir
Published online on: 21 Dec 2015
3

Generalized Linear Autoregressive Moving Average Models

William T.M. Dunsmuir

CONTENTS
3.1 Introduction.................................................................51
3.2 GLARMA Models for Single Time Series.............................................52
  3.2.1 The GLARMA Likelihood........................................................54
  3.2.2 Parameter Identifiability.........................................................55
  3.2.3 Distribution Theory for Likelihood Estimation...............................56
  3.2.4 Convergence of GLARMA Recursions: Ergodicity and Stationarity.......57
3.3 Application of Univariate GLARMA Models.........................................57
3.4 GLARMA Models for Multiple Independent Time Series..........................61
  3.4.1 Examples.............................................................................61
  3.4.2 General Model for Multiple Independent GLARMA Models with
     Random Effects....................................................................62
3.5 Fixed Effects Multiple GLARMA Model.............................................63
  3.5.1 Maximum Likelihood Estimation................................................63
  3.5.2 Application of Multiple Fixed Effects GLARMA to Road Deaths..........64
3.6 Random Effects Multiple GLARMA Model...........................................67
  3.6.1 Maximum Likelihood Estimation...............................................67
  3.6.2 Laplace Approximation and Adaptive Gaussian Quadrature.................68
  3.6.3 Application to Road Deaths Data.............................................71
  3.6.4 Computational Speed and Accuracy..........................................72
     3.6.4.1 Impact of Increasing the Number of Quadrature Points...............72
     3.6.4.2 Comparison with Numerical Derivatives................................72
3.7 Interpretation of the Serially Dependent Random Effects Models...............73
3.8 Conclusions and Future Directions....................................................73
References..................................................................................74

3.1 Introduction
Generalized linear autoregressive moving average (GLARMA) models are a class of
observation-driven non-Gaussian nonlinear state space models in which the state process
depends linearly on covariates and nonlinearly on past values of the observed process.
Conditional on the state process, the observations are independent and have a distribution
from the exponential family. This could include continuous responses but, in this chapter, we focus entirely on discrete responses such as binary, binomial, Poisson, or negative binomial distributions.

The main advantage of GLARMA models over other observation or parameter-driven models is that they can be fit relatively easily to long time series or to many individual time series. Examples of the latter will be given herein. They provide a natural extension of the GLM modeling framework to include serial dependence terms and provide rapid assessment of the presence and form of serial dependence. This model-based assessment of serial dependence is particularly useful for discrete response data for which standard sample autocorrelation methods can be misleading—see Davis et al. (2000) for example.

Since their genesis in the unpublished paper by Shephard (1995), GLARMA models have found application in many disciplines including financial modeling (Rydberg and Shephard, 2003; Liesenfeld et al., 2006), epidemiological assessments (Davis et al., 2003; Turner et al., 2011), clinical management (Buckley and Bulger, 2012), analysis of crime statistics (Dunsmuir et al., 2008), and primate behavior (Etting and Isbell, 2014).

Benjamin et al. (2003) provide a review of generalized autoregressive models of which the GLARMA models are a subclass. Davis et al. (1999) review various approaches to modeling count time series, whereas Davis et al. (2003, 2005) consider GLARMA models for Poisson response series. Kedem and Fokianos (2002) provide a comprehensive coverage of observation-driven models for discrete response time series with covariates.

In this chapter, we first review the properties of GLARMA models for single time series and describe how they can be estimated using the glarma package (Dunsmuir et al., 2014) for the R language (R Core Team, 2014)—see also Dunsmuir and Scott (2015). Additionally, we describe how univariate GLARMA fitting software can be adapted to model multiple independent series either by a fixed effects analysis or by a random effects analysis. Multiple independent time series share many common features with longitudinal data or panel data collections, but differ in two key aspects. First, in a longitudinal data setting there are typically a large number of independent short temporal observation trajectories, whereas in the applications we discuss, there are a small to moderate number of time series each of which is long. Second, longitudinal data analyses typically restrict the modeling of serial dependence in each trajectory to be the same across the trajectories, whereas for the applications we have in mind, the serial dependence can vary across the series. Indeed, some series may not require serial dependence terms at all, whereas others can have substantial serial dependence. A major focus of this chapter is on describing how to combine GLARMA models for single time series with random effects on regressors between series.

### 3.2 GLARMA Models for Single Time Series

In this section, we describe the GLARMA model class for a single time series. Let $y_1, \ldots, y_n$ be the available observations on the discrete response series. Associated with these are vectors, $x_t$, of $K$ regressors observed for $t = 1, \ldots, n$. Let $\mathcal{F}_t = \{Y_s : s < t, x_s : s \leq t\}$ denote the past information available on the response series and the past and present information on the regressors. The distribution of $Y_t$ conditional on $\mathcal{F}_t$ is assumed to be of exponential family form

$$f(y_t|W_t) = \exp \{y_t W_t - a_t b(W_t) + c_t\}, \quad (3.1)$$
where \{a_t\} and \{c_t\} are sequences of constants, with \(c_t\) often depending on the observations \(y_t\). The response distribution could also be the negative binomial for which an additional parameter is required. The information in \(F_t\) is summarized in the state variable \(W_t\); that is, \(W_t\) is a function of the elements in \(F_t\). We denote the conditional means and variances of the responses as \(\mu_t := E(Y_t|W_t)\) and \(\sigma_t^2 := \text{var}(Y_t|W_t)\). Throughout, we will use the canonical link connecting \(\mu_t\) and \(W_t\), in which case, \(\mu_t = a_t b(W_t)\) and \(\sigma_t^2 = a_t \ddot{b}(W_t)\), where \(b(u)\) and \(\ddot{b}(u)\) are the first and second derivatives, respectively, of \(b(u)\) with respect to \(u\).

While (3.1) is not the fully general form of the exponential family (see McCullagh and Nelder, 1989), it covers several popular and useful distributions. The basic theory and computational methodology presented here can be readily modified to include other response distributions or more general specifications of link functions. An example of such an extension is the use of the negative binomial response distribution. Let \(\mu_t = \exp(W_t)\). The \texttt{glarma} package uses the negative binomial density in the form

\[
 f(y_t|W_t, \alpha) = \frac{\Gamma(\alpha + y_t)}{\Gamma(\alpha) \Gamma(y_t + 1)} \left[ \frac{\alpha}{\alpha + \mu_t} \right]^\alpha \left[ \frac{\mu_t}{\alpha + \mu_t} \right]^{y_t}, \quad y_t = 0, 1, 2, \ldots
\]

Note that \(\sigma_t^2 = \mu_t + \mu_t^2/\alpha\). As \(\alpha \to \infty\), the negative binomial density converges to the Poisson density. Also note that if \(\alpha\) is known, this density belongs to the one-parameter exponential family with appropriate definitions of \(\theta_t, b(\theta_t), a_t, c_t\). If \(\alpha\) is not known, (3.2) is not a member of the one-parameter exponential family.

Specification of \(W_t\) for observation-driven models takes various forms; see Benjamin et al. (2003) for a general discussion or Kedem and Fokianos (2002) for a comprehensive treatment of various models. Following our previous work (Davis et al., 1999, 2003), we consider the case where the state vector in (3.1) is linear in the covariates

\[
 W_t = x_t^T \beta + Z_t. \tag{3.3}
\]

Here the “noise” process \(\{Z_t\}\), which induces serial dependence in the states and hence in the observations, is of the form

\[
 Z_t = \sum_{j=1}^{\infty} \gamma_j(\psi) e_{t-j}, \tag{3.4}
\]

with the parameters \(\psi\) specified separately from the regression coefficient \(\beta\).

The predictive residuals in (3.4) are defined as

\[
 e_t = \frac{Y_t - \mu_t}{\nu_t}, \tag{3.5}
\]

where the scaling sequence \(\{\nu_t\}\) is to be selected. We consider three choices (currently as supported in the \texttt{glarma} package): \(\nu_{p,t} = \sigma_t\), giving classical Pearson residuals; \(\nu_{s,t} = \sigma_t^2\), giving the “score-type” residuals suggested by Creal et al. (2008); and, \(\nu_{i,t} = 1\) (referred to as “identity” scaling), mainly used for binary response GLARMA models. Note that \(\{e_t\}\) are martingale differences, and hence are zero mean and uncorrelated. The Pearson type \(\{e_t\}\) also have unit variance, and hence are weakly stationary white noise. Often, when the identity scaling is used in the Poisson and negative binomial cases, the resulting state equation can be explosive leading to infinite means or zero means, both forms of degeneracy.
This may not be very crucial for the binomial response case but little is currently known about such stability issues. We discuss this topic in more detail later in this chapter.

Following Davis et al. (1999), for a GLARMA model, the general form of \( \{Z_t\} \) in (3.4) is specified via an ARMA-type recursion

\[
Z_t = \sum_{i=1}^{p} \phi_i (Z_{t-i} + e_{t-i}) + \sum_{i=1}^{q} \theta_i e_{t-i}.
\]

The \( \{Z_t\} \) defined in this way can be thought of as the best linear predictor of a stationary invertible ARMA process with driving noise \( \{e_t\} \) of scaled predictive residuals.

The recursions to construct the \( Z_t \) component of the state and its derivatives require initialization. In the \texttt{glarma} package we set \( e_t = 0 \) and \( Z_t = 0 \) for \( t \leq 0 \) ensuring that the conditional and unconditional expected values of \( e_t \) are zero for all \( t \).

### 3.2.1 The GLARMA Likelihood

Given \( n \) successive observations \( y_t \) (\( t = 1, \ldots, n \)) on the response series, and fixed initial conditions (such as using zeros) for the recursions in (3.6), the likelihood is constructed as the product of conditional densities of \( Y_t \) given \( F_t \) (or equivalently the state \( W_t \)) giving the log-likelihood corresponding to the distribution (3.1) as

\[
l(\delta) = \sum_{t=1}^{n} \{y_t W_t(\delta) - a_t b(W_t(\delta)) + c_t\},
\]

where \( \delta = (\beta, \phi, \theta) \). For the negative binomial response distribution, the log-likelihood is more complicated because the shape parameter \( \alpha \) also has to be estimated along with \( \beta, \phi, \) and \( \theta \). For this case, we expand the parameters to \( \delta = (\beta, \phi, \theta, \alpha) \). Because the \( e_t \) in (3.5), the \( Z_t \) in (3.6), and the \( W_t \) in (3.3) are functions of the unknown parameter \( \delta \), they must be recomputed at each iteration to maximize the likelihood. For \( t = 1, \ldots, n \), recursive expressions for calculating \( e_t, Z_t, \) and \( W_t \) as well as their first and second partial derivatives with respect to \( \delta \) are available in Davis et al. (2005) for the Poisson case. Corresponding formulas for the binomial and negative binomial cases are easily derived in a similar way. The essential computational cost lies with the recursions for \( Z_t \) and \( W_t \) (\( t = 1, \ldots, n \)) and their first and second derivatives. These recursions are common to all response distributions. It is only the computation of \( e_t \) and the \( \log f_{Y_t|W_t}(y_t|W_t; \delta) \) that depend on the response distribution being used and choice of scaling \( \gamma_t \).

It may be tempting to consider maximization of (3.7) by using readily available existing GLM (generalized linear modeling) software. Without modification, this is not possible because, for \( t = 1, \ldots, n \), the \( e_t \) and hence \( Z_t \) appearing in the state equation for \( W_t \) are functions of unknown parameters. One suggestion that is sometimes made is, for an initial value of \( \delta \), calculate \( e_t(\delta) \) and hence via (3.6), \( Z_t(\delta) \). Using these, create an augmented set of covariates \( \tilde{x}_t = (x_t^T, Z_{t-1}, \ldots, Z_{t-p}, e_{t-1}, \ldots, e_{t-\text{max}(p,q)}) \) and update the parameters estimates using standard GLM software for the regression term \( \tilde{x}_t^T \delta \) to update the estimate of \( \delta \). Recursion of this method does converge but unfortunately not to the maximum likelihood estimate and hence is biased. In any case, there is not much additional computational cost using the GLARMA likelihood calculations directly.
The log-likelihood is maximized from a suitable starting value \( \delta \) using Newton–Raphson iteration or a Fisher scoring approximation. Define the vector of first and second derivatives of the log-likelihood by \( d(\delta) = \partial l(\delta)/\partial \delta \) and \( D_{\text{NR}}(\delta) = \partial^2 l(\delta)/\partial \delta \partial \delta^T \), where the matrix of second derivatives of the log-likelihood for (3.7) is given by

\[
D_{\text{NR}}(\delta) = \sum_{t=1}^{n} \left[ y_t - a_t \hat{b}(W_t) \right] \frac{\partial^2 W_t}{\partial \delta \partial \delta^T} - \sum_{t=1}^{n} a_t \hat{b}(W_t) \frac{\partial W_t}{\partial \delta} \frac{\partial W_t}{\partial \delta^T}. \tag{3.8}
\]

At the true parameter \( \delta \), \( E[y_t - a_t \hat{b}(W_t)|\mathcal{F}_t] = 0 \). Hence the expected value of the first summation in (3.8) is zero which motivates the Fisher scoring–type approximation based only on first derivatives given by

\[
D_{\text{FS}}(\delta) = -\sum_{t=1}^{n} a_t \hat{b}(W_t) \frac{\partial W_t}{\partial \delta} \frac{\partial W_t}{\partial \delta^T}. \tag{3.9}
\]

Note that \( E[D_{\text{FS}}(\delta)] = E[D_{\text{NR}}(\delta)] \). Also, using the fact that \( d(\delta) \) is a sum of martingale differences, the usual identity \( E[D_{\text{NR}}(\delta)] = -E[d(\delta) d(\delta)^T] \) holds. These expectations cannot be computed in closed form. From an initial value for \( \delta \), Newton–Raphson (using \( D_{\text{NR}} \)) or approximate Fisher scoring (using \( D_{\text{FS}} \)) methods are used to find the maximum likelihood estimate \( \hat{\delta} \).

### 3.2.2 Parameter Identifiability

The GLARMA component \( Z_t \) of the state variable given in (3.6) can be rewritten as

\[
Z_t = \sum_{i=1}^{p} \phi_i Z_{t-i} + \sum_{i=1}^{\tilde{q}} \tilde{\theta}_j e_{t-i}, \tag{3.10}
\]

where \( \tilde{q} = \max(p, q) \) and if \( p \leq q \), \( \tilde{\theta}_j = \theta_j + \phi_j \) for \( j = 1, \ldots, p \) and \( \tilde{\theta}_j = \theta_j \) for \( j = p + 1, \ldots, q \), while if \( p > q \), \( \tilde{\theta}_j = \theta_j + \phi_j \) for \( j = 1, \ldots, q \) and \( \tilde{\theta}_j = \phi_j \) for \( j = q + 1, \ldots, p \). Hence, if \( Z_t = 0 \) for \( t \leq 0 \) and \( e_t = 0 \) for \( t \leq 0 \) and if \( \tilde{\theta}_j = 0 \) for \( j = 1, \ldots, \tilde{q} \), the recursion (3.10) renders \( Z_t = 0 \) for all \( t \), there is no serial dependence in the GLARMA model, and it reduces to a standard GLM model. This is equivalent to \( \phi_j = -\theta_j \) for \( j = 1, \ldots, p \) and \( \theta_j = 0 \) for \( j = p + 1, \ldots, \tilde{q} \). Consequently, the null hypothesis of no serial dependence requires only these constraints on \( \theta \) and \( \phi \) and there can be nuisance parameters that cannot be estimated. This has implications for (1) convergence of the iterations required to optimize the likelihood and (2) on testing that there is no serial dependence in the observations, beyond that induced by the regression component \( x_t^\top \beta \).

In cases where \( p > 0 \) and \( q = 0 \) (equivalent to an ARMA \((p, p)\) specification with constraint \( \theta_j = \phi_j \)) or where \( p = 0 \) and \( q > 0 \) (a pure \( \text{MA}(q) \)) case, identification issues do not arise and there are no nuisance parameters to take into account when using a score, Wald or likelihood ratio test for the GLARMA parameters. However, in other “mixed” model cases where both \( p > 0 \) and \( q > 0 \), some care is needed when fitting models and testing
for serial dependence. To illustrate, consider the case where no serial dependence exists but \( p = q > 0 \) is specified. Then the likelihood iterations are unlikely to converge because the likelihood surface will be “ridge-like” on the manifold where \( \phi_j = -\theta_j \), an issue that is encountered for standard ARMA model fitting. Corresponding to this, the second derivative matrix \( D_{NR}(\delta) \) will be singular or the state variable \( W_t \) can degenerate or diverge. Because of this possibility, it is prudent to start with low orders for \( p \) and \( q \) and avoid specifying them as equal. Once stability of estimation is reached for a lower-order specification, increasing the values of \( p \) or \( q \) could be attempted.

The likelihood ratio test that there is no serial dependence versus the alternative that there is GLARMA-like serial dependence with \( p = q > 0 \) will not have a standard chi-squared distribution because the parameters \( \phi_j \), for \( j = 1, \ldots, p \), are nuisance parameters which cannot be estimated under the null hypothesis. Testing methods such as those proposed by Hansen (1996) or Davies (1987) need to be developed for this situation. Further details on these points can be found in Dunsmuir and Scott (2015).

### 3.2.3 Distribution Theory for Likelihood Estimation

The consistency and asymptotic distribution for the maximum likelihood estimate \( \hat{\delta} \) is rigorously established only in a limited number of special cases. In the stationary Poisson response case in Davis et al. (2003) where \( x_t^T \equiv 1 \) (intercept only) and \( p = 0 \) and \( q = 1 \), these results have been proved rigorously. Similarly, for simple models in the Bernoulli stationary case in Streett (2000) these results hold. Simulation results are also reported in Davis et al. (1999, 2003) for nonstationary Poisson models. Other simulations not reported in the literature support the supposition that \( \hat{\delta} \) has a multivariate normal distribution for large samples for a range of regression designs and for the various response distributions considered here.

For inference in the GLARMA model, it is assumed that the central limit theorem holds so that

\[
\hat{\delta} \overset{d}{\approx} N(\delta, \hat{\Omega}), \tag{3.11}
\]

where the approximate covariance matrix is estimated by \( \hat{\Omega} = -D_{NR}(\hat{\delta})^{-1} \) or \( \hat{\Omega} = -D_{FS}(\hat{\delta})^{-1} \). In the glarma package, this distribution is used to obtain standard errors and to construct Wald tests of the hypotheses that subsets of \( \delta \) are zero. It is also assumed that Wald tests and equivalent likelihood ratio tests will be asymptotically chi-squared with the correct degrees of freedom, results which would follow straightforwardly from (3.11) and its proof when available. Regardless of the technical issues involved in establishing a general central limit theorem, the earlier approximate result seems plausible since, for these models, the log-likelihood is a sum of elements in a triangular array of martingale differences. Conditions under which this result would likely hold include identifiability conditions as discussed earlier, conditions on the regressors similar to those used in Davis et al. (2000, 2003) and Davis and Wu (2009), where the covariates \( x_t \) are assumed to be a realization of a stationary time series or is defined as \( x_t = x_{nt} = f(t/n) \) where \( f(u) \) is a piecewise continuous function from \( u \in [0,1] \) to \( \mathbb{R}^K \). Additional conditions on the coefficients are also needed to ensure that \( Z_t \) and hence \( W_t \) do not degenerate or grow without bound. Indeed, little is known so far about suitable conditions to ensure this.
3.2.4 Convergence of GLARMA Recursions: Ergodicity and Stationarity

To date, the stationarity and ergodicity properties of the GLARMA model are only partially understood. These properties are important to ensure that the process is capable of generating sample paths that do not degenerate to zero or do not explode as time progresses, as well as for establishing the large sample distributional properties of parameter estimates. Davis et al. (2003) provide results for the simplest of all models: Poisson responses specified with \( p = 0, q = 1, \) and \( x_t^T \beta = \beta \). Results for simple examples of the stationary Bernoulli case are given in Streett (2000).

For the Poisson response distribution GLARMA model, failure to scale by the variance or standard deviation will lead to unstable Poisson means (that diverge to infinity or collapse to zero as an absorbing state for instance) and existence of stationary and ergodic solutions to the recursive state equation is not assured—see Davis et al. (1999, 2003, 2005) for details. For the binomial situation, this lack of scaling should not necessarily lead to instability in the success probability as time evolves since the success probabilities, \( p_t \), and observed responses, \( Y_t \), are both bounded between 0 and 1. Thus, degeneracy can only arise if the regressors \( x_t \) become unbounded. As recommended in Davis et al. (1999), temporal trend regressors should be scaled using a factor relating to the sample size \( n \).

Asymptotic results for various types of observation-driven models without covariates are increasingly becoming available. Tjøstheim (2012) (also see Tjøstheim [2015; Chapter 4 in this volume]) has provided a review of the ergodic and stationarity properties for various observation-driven models—primarily in the Poisson response context—as well as presenting large sample theory for likelihood estimation. Wang and Li (2011) discuss the binary (BARMA) model and present some asymptotic results for that case. However, the state equation for the BARMA model differs from that for the binary GLARMA model in that the latter involved scaled residuals while the former uses identity residuals and is also structurally different in its use of past observation of \( \{Y_t\} \). Davis and Liu (2015) present general results for the one-parameter exponential family response distributions and a semiparametric observation-driven specification of the state equation. Woodard et al. (2011) present some general results on stationarity and ergodicity for the GARMA models similar to those available in Benjamin et al. (2003). However, because the state equation recursions involve applying the link function to both the responses and the mean responses, none of these results apply to the specific form of GLARMA models presented here. Also, none of these recent results consider the case of covariates; hence they are not, as yet, applicable to likelihood estimation for regression models for discrete outcome time series.

3.3 Application of Univariate GLARMA Models

We now illustrate the fitting of GLARMA models to binomial and binary time series arising in the study of listener responses to a segment of electroacoustic music. This example also motivates studying multiple independent time series as an ensemble (as will be discussed in the next section). The background to the analysis presented here is in Dean et al. (2014b). Members of a panel comprising three musical expertise groups (8 electroacoustic musicians, 8 musicians, and 16 nonmusicians) provided real-time responses to a segment of electroacoustic music. Aims of these types of experiments are to determine the way in which features of the music (in this case the sound intensity) impact listener response
measured in various ways (we concentrate on the "arousal" response). Dean et al. (2014b) present a variety of standard time series methods for modeling the arousal responsiveness in terms of lags of the musical intensity. Questions such as: “Do the musical expertise groups display differences between them with respect to the impact of intensity on their group average responses?” and “Are there substantial differences between individuals within the panel or within each musical group?” were addressed. For example, in Dean et al. (2014b), the transfer function coefficients for the impact of changes, at lag 1, of musical intensity on the changes in arousal were modeled using 11 lag transfer function models of the form

$$\nabla Y_{jt} = \omega_0 + \sum_{k=1}^{11} \omega_j \nabla X_{t-k} + \alpha_{jt}, \quad (3.12)$$

where $\nabla A_t = A_t - A_{t-1}$ and $\alpha_{jt}$ was modeled by an autoregression of at most order 3. There was evidence that variation between individuals in their responsiveness was substantial and suggested use of a cross-sectional time series analysis as in Dean et al. (2014a). However, analysis of individual responses suggested varying levels of “stasis”; that is, frequent and sometime prolonged periods during which their arousal response did not change. In some cases, this led to a high level of zeros in their differenced responses, which has the potential to impact the validity of model estimates based on traditional Gaussian linear time series analysis. In fact, listeners varied substantially in the amount of time that their responses are in stasis, for example, from 15% to 90%. Response distributions with such large numbers of zeros constitute a challenge for conventional time series analysis.

To examine how robust their findings were based on standard Gaussian linear time series methods, Dean et al. (2014b) also considered an approach similar to that employed in Rydberg and Shephard (2003) for decomposing transaction level stock price data into components of change and size of change. For each listener, binary responses were defined as $D_{jt} = 1$ if the change in their response from time $t - 1$ to $t$ was positive, otherwise $D_{jt} = 0$. This is one of the two components of a potential trinomial model for the change process for which the approach of Liesenfeld et al. (2006) could be used; currently, the glarma package does not handle trinomial responses.

In order to answer “Are there variations between the three musical groups with respect to their group average responses to the same musical excerpt?” the binary time series were aggregated at each time into binomial counts of 8, 8, and 16 respondents at each time in the EA, M, and NM musical groups. The multiple GLARMA fixed effects modeling (described in the next section) was used to examine the differences between group average responses. For this, it is assumed that the aggregated counts obey a binomial distribution. The independence of trials assumption is not in doubt since the individuals responded independently to the musical excerpt. However, the assumption that each individual in the group shares the same probability of response at each time appears in doubt as we now explain.

We consider the nonmusician group of 16 listeners. Let $S_t = \sum_{j=1}^{16} D_{jt}$ count the number of respondents whose arousal change was positive and assume that $S_t \sim \text{Bin}(\pi_t, 16)$, where

$$\text{logit}(\pi_t) = \omega_0 + \sum_{k=1}^{11} \omega_k \nabla X_{t-k} + \alpha_t \quad (3.13)$$
and $\alpha_t$ follows a suitably specified GLARMA process. For this group, we found that $(p, q) = (3, 0)$ was adequate (based on minimum AIC). The fitted transfer function coefficients with individual 95% confidence limits under the assumption of normality are shown in Figure 3.1a. Most transfer functions coefficients are individually significant. The observed binary responses (as probabilities) along with the fixed effects fit and GLARMA model fit is shown in Figure 3.2a. The nonrandomized probability integral transform (PIT) residual plot for this fit is shown in Figure 3.3. Clearly the binomial assumption is not correct as the PIT plot suggests that the binomial distribution is not providing a good prediction of the probability of small or large counts. This is not surprising since the 16 individuals in this group show substantial variability in their “stasis” levels, which of course wil impact the average probability of a positive change in arousal. The PIT analysis suggests that aggregation of individual binary responses in this way is not appropriate.

Consequently, we turn to analysis of two individual binary responses to illustrate the application of GLARMA model for binary data using the individual models

$$\logit(\pi_{jt}) = \omega_{j,0} + \sum_{k=1}^{11} \omega_{j,k} \nabla X_{t-k} + \alpha_{jt},$$  \hspace{1cm} (3.14)

where $\{\alpha_t\}$ is a GLARMA process for the $j$th series. For most of the series, $p = 1$ and $q = 0$ seemed appropriate; hence, we settle on this for all 32 series. We illustrate the results of such fits on two listeners: listener 22, who had 88% stasis, and listener 23, who had 52% stasis. With such a high proportion of zero changes in arousal responses, it is not clear that application of standard Gaussian time series transfer function modeling would be reasonable for these two cases. We modeled the binary response series $\{D_{jt}\}$ using a binary GLARMA model with probabilities specified as in (3.14). Figure 3.1b and c shows the estimated values of the transfer function coefficients $\omega_{j,k}$ along with 95% significance levels. There are clear differences between the two individual responses to the same musical excerpt. The fitted values for these cases are shown in Figure 3.2b and c. There are also substantial differences between overall level as measured by the intercept terms ($\hat{\omega}_{22,0} = -3.24 \pm 0.39$, $\hat{\omega}_{23,0} = -1.21 \pm 0.54$), which is consistent with the relative levels of stasis observed for
FIGURE 3.2
Binomial time series for 16 NM groups, response with GLARMA fits of response to changes in musical intensity and two individual binary time series with fits. (a) Binomial Fit: 16 Nonmusician Listeners (b) Binary Fit: Listener 22 (c) Binary Fit: Listener 23.

Binomial GLARMA model: 16 nonmusician listeners

FIGURE 3.3
PIT residual plot for GLARMA model for binomial counts of positive responses in 16 nonmusician group listeners.
these two individuals as reported earlier. The GLARMA model autoregressive coefficients also substantially differed ($\hat{\phi}_{22} = 0.687 \pm 0.079$ and $\hat{\phi}_{23} = 0.894 \pm 0.024$). This also holds when all 32 individual responses are modeled in this way.

We have also applied GLARMA modeling with the glarma package on binary and binomial time series to study responsiveness to musical features in the complete panel of 32 listeners. A clear conclusion from this extended analysis is that there is a strong need to consider modeling of individual responses in order to address questions such as: “Are there differences between musical expertise groups?” Also clear is the need to accommodate differing amounts of serial dependence for each series in the ensemble, something that even current longitudinal data analysis for mixed models does not readily allow. In the next section we explain two approaches to allowing for variation between individual time series.

3.4 GLARMA Models for Multiple Independent Time Series

3.4.1 Examples

The musicology example of Section 3.3 is an example where ensembles of individual independent responses need to be modeled together. In a public health policy setting, Bernat et al. (2004) assessed, using a pooled cross-sectional random effects analysis, the impact of lowering the legal allowable blood alcohol concentration (BAC) in motor vehicle drivers from 0.10 to 0.08 on monthly counts of single vehicle night time fatalities in 17 U.S. states. In that study, serial dependence was detected in some series but could not be modeled with software and methods available at that time. The purpose of the remainder of this chapter is to present methods that overcome these gaps.

Two approaches are developed here. The first is referred to as the fixed effects plus GLARMA specification. Here, regression effects and serial dependence parameters may be constrained across series to test various hypotheses of interest primarily about regression effects, but also, about variation in serial dependence between series. For long longitudinal data each series can be estimated and modeled individually using the GLARMA models previously discussed and combined likelihoods for constrained parameterizations constructed on which to base inference. Here, the length of the individual time series allows considerable flexibility in serial dependence structures between individual series. In traditional longitudinal data analysis where the number of repeated measures is low, it may not be possible to allow for individuality of this type.

The second approach is based on a random effect specification of the regression component while allowing individual series to have different serial dependence structures and strengths. This is not possible for traditional longitudinal data and the methods proposed here is a substantial extension of existing methodology, which is enabled by the length of the individual series.

Before defining models that combine fixed or random effects with GLARMA serial dependence, we mention some other recent examples along these lines. Xu et al. (2007) present a parameter-driven specification for serial dependence with random effects on covariates. Their approach is limited by the requirement that only autoregressive serial dependence is covered and all series must share the same structure and parameter values. Additionally, the method has not been demonstrated, neither in simulation nor in
application, on series longer than 20 time points. Zhang et al. (2012) specify an exponential
decay autocorrelation on a latent process and use marginal estimation methods for their
combined serial dependence and random effects specification. They also assume that the
structure and strength must be the same for all series. These approaches, which force the
same dependence structure on all series, may be a legacy from the modeling of traditional
longitudinal data in which the individual series are very short and for which the ability to
detect different serial dependence properties between series is limited. It is our view that
for long longitudinal data, this restriction is artificial for two reasons. First, in all exam-
ples we have encountered, there is strong evidence that the serial dependence structure
and strength varies between series; some series will have strong serial dependence, while
others will have none at all. One explanation for this is that the covariates used for the
ensemble for fixed and random effect terms may not be inclusive of covariates that are
unavailable and impact individual series—in such circumstances serial dependence may
be stronger simply because it acts as a proxy for unobserved covariates. The second reason
relates to the fact that the use of random effects on regression variables should logically
extend to their use for serial dependence parameters. The models and methods presented
here provide considerably more flexibility than those proposed by Xu et al. (2007) and
Zhang et al. (2012). However, being based on GLARMA models, which require regular
time spacing of observed outcomes, they cannot handle irregular spacings. For this irreg-
ularity spaces observations, parameter-driven models for the individual series would be
more appropriate.

It is our view that, more frequently, increasingly long longitudinal data will become
available and methods such as those proposed here will be become needed. For exam-
ple, studies with panels of subjects equipped with automatic data loggers measuring their
physical condition and activity levels are now quite feasible.

### 3.4.2 General Model for Multiple Independent GLARMA Models with Random Effects

Let $Y_{jt}$ be the observation at time $t = 1, \ldots, n_j$ on the $j$th series of counts, where $j = 1, \ldots, J$ and let $x_{jt}, r_{jt}$ be the covariates for the $j$th series. We also let $r_{jt}$ denote $d$ random effect covariates that apply to all series through coefficients represented as vectors of normally distributed random effects $U_j \sim \text{i.i.d } N(0, \Sigma(\lambda))$, where $\Sigma$ is a $d \times d$ covariance matrix determined by parameters $\lambda$.

In addition to assuming that the series $\{Y_{jt}\}$ are independent across the $J$ cases, for each $j$, we also assume that, given the state process $W_{j,t}$, $Y_{jt}$ are independent with exponential family distribution (3.1), where

$$W_{j,t} = x_{jt}^T \beta^{(j)} + r_{jt}^T U_j + Z_{jt}$$

is the linear state process for the $j$th case series. Typically, some of the covariates appear in
both $x_{jt}$ and $r_{jt}$ in which case corresponding components of the $\beta^{(j)}$ will not depend on $j$ and some or all of the covariates can be the same across all series.

Serial dependence is modeled with the $Z_{jt}$ assumed to satisfy (3.6) with degrees $(p^{(j)}, q^{(j)})$

$$Z_{jt} = \sum_{l=1}^{p^{(j)}} \phi^{(j)}_l (Z_{j,t-l} + e_{j,t-l}) + \sum_{l=1}^{q^{(j)}} \theta^{(j)}_l e_{j,t-l}.$$ (3.16)
Let $\tau^{(j)} = (\phi^{(j)}_1, \ldots, \phi^{(j)}_{p^{(j)}}, \theta^{(j)}_1, \ldots, \theta^{(j)}_{q^{(j)}})$. We distinguish three special cases for the general state equation (3.15).

1. **Generalized linear mixed model**: In this specification, the serial dependence term $Z_{jt}$ is absent, giving the standard generalized linear mixed model for longitudinal data discussed in Diggle et al. (2002) and Fitzmaurice et al. (2012)

$$W_{j,t} = x_{jt}^T \beta^{(j)} + r_{jt}^T U_j.$$  

The only source of within series correlation is in the random effects component. Various packages (such as SAS and R) are available for model fitting.

2. **Fixed effects multiple GLARMA model**: Random effects are not included but there is a serial dependence in the form of an observation-driven model:

$$W_{j,t} = x_{jt}^T \beta^{(j)} + Z_{jt},$$  

where $Z_{jt}$ is given by (3.16).

3. **Random effects multiple GLARMA model**: Both random effects and serial dependence are present. This is the general specification for $W_{jt}$ given by (3.15), with $Z_{jt}$ specified by the GLARMA model (3.16).

Model Class 1 was used in Bernat et al. (2004). A previous example of Model Class 2 type of modeling was considered in Dunsmuir et al. (2004), who investigated the commonality of regression impacts on three series of daily asthma presentation counts. Extension to datasets with substantially more time series and more complex hypotheses on the fixed effects parameters required considerable development of the previous software. The random effects multiple GLARMA model is developed in detail in Dunsmuir et al. (2014). The BAC dataset will be used to illustrate fitting of both models (3.18) and (3.15).

### 3.5 Fixed Effects Multiple GLARMA Model

#### 3.5.1 Maximum Likelihood Estimation

In this section, we develop maximum likelihood estimation for the model with individual series state equations specified as in (3.18). Let the log-likelihood for the $j$th series be as in (3.7) and denoted by $l_j(\theta^{(j)})$ where $\theta^{(j)T} = (\beta^{(j)T}, \tau^{(j)T})$. The log-likelihood across all series is

$$l(\theta) = \sum_{j=1}^{J} l_j(\theta^{(j)})$$  

where $\beta^{T} = (\beta^{(1)T}, \ldots, \beta^{(J)T})$ and $\tau^{T} = (\tau^{(1)T}, \ldots, \tau^{(J)T})$. Let $\theta = (\beta^{T}, \tau^{T})^T$ denote all parameters.
64 Handbook of Discrete-Valued Time Series

A primary focus is on testing if the parameterization across series can be simplified so that regression coefficients or serial dependence parameters can be constrained to be the same. We consider only linear constraints of the form $\theta = A\psi$ where $A$ has fewer columns than rows and $\psi$ denotes the lower-dimensional vector of parameters in the constrained model. Typically $\psi^T = (\psi^T_\beta, \psi^T_\tau)$ since we will generally not be interested in relating the regression coefficients to the serial dependence parameters. In that case $A$ will be block diagonal appropriately partitioned. Denote the log-likelihood with respect to the constrained parameters as $l(\psi) = l(A\theta)$.

Maximization of (3.19) with respect to the constrained parameters can be over a high-dimensional parameter space. Initial estimates of $\hat{\theta}$ are obtained by maximizing (3.19) without constraints which is the same as maximizing all individual likelihoods separately and combining the resulting $l_j\left(\hat{\theta}^{(j)}_1, \hat{\tau}^{(j)}_1\right)$. These unconstrained estimates are used to initialize the constrained parameters via $\psi(0) = (A^TA)^{-1}A^T\hat{\theta}$. Next, using the appropriate components of $\hat{\theta}(0) = A\hat{\psi}(0)$, each component log-likelihood $l_j\left(\hat{\theta}^{(j)}_1\right)$ and its derivatives are calculated using the standard GLARMA software. Finally, these are combined to get the overall $l\left(\hat{\psi}(0)\right)$. Derivatives with respect to $\psi$ can be obtained using the identities $\partial l(\psi)/\partial \psi = A^T\partial l(A\psi)/\partial \theta$ and $\partial^2 l(\psi)/\partial \psi \partial \psi^T = A^T\partial^2 l(A\psi)/\partial \theta \partial \theta^T A$. This procedure is then iterated to convergence using the Newton–Raphson or Fisher scoring algorithm. Fisher scoring was found to be more stable in the initial stages. Once the derivative $\partial l(\psi)/\partial \psi$ settles down, the iterative search for the optimum can be switched to the Newton–Raphson updates, which typically gives speedier convergence.

Similar to the single series case, the asymptotic properties of the MLEs $\hat{\psi}$ have not been established for the general model previously described. Asymptotic results for longitudinal data typically let the number of cases $J$ tend to infinity and the lengths of individual series $n_j$ are typically held fixed. For this scenario, asymptotic theory is typically straightforward since it relies on large numbers of independent trajectories. For our applications, $J$ is often bounded and we perceive of all $n_j$ as tending to infinity in which case asymptotic results rely on those for individual time series which, as previously noted, are rather underdeveloped. We assume, however, that asymptotic results hold and perform inference in the usual way. For example, the matrices of second derivatives $\tilde{I}(\hat{\psi})$, computed in the course of the Newton–Raphson or Fisher scoring maximization procedures, are used to estimate standard errors for individual parameters. Also, to test the null hypothesis of common regression effects, we use the likelihood ratio statistic $G^2 = -2\left[l(\hat{\psi}_0) - l(\hat{\psi}_1)\right]$, where $\hat{\psi}_0$ is the estimate obtained under the null hypothesis and $\hat{\psi}_1$ the estimate obtained under the alternative. Degrees of freedom, for the chi-squared approximate reference distribution for $G^2$, are calculated in the usual way. These ideas were first illustrated in Dunsmuir et al. (2004) for three series of daily asthma counts which are assessed for common seasonal patterns, day of the week, weather, and pollution effects.

3.5.2 Application of Multiple Fixed Effects GLARMA to Road Deaths

Bernat et al. (2004) assessed the impact of lowering the legal allowable BAC in motor vehicle drivers from 0.10 to 0.08 on monthly counts of single vehicle night time fatalities in 17 U.S. states for which at least 12 months of post intervention data were available. The study design selected 72 consecutive months of data with 36 months prior to the decrease in
allowable BAC and up to 36 months after for each of the 17 states used in the analysis. The mixed effects model presented in Bernat et al. (2004) assumed that the observed counts of single vehicle night time deaths, $Y_{jt}$ in month $t$ for state $j$, had, conditional on the random effects $U_j$, a Poisson distribution with log mean given by

$$W_{jt} = \beta_0 + \beta_1 I_1(t) + \beta_2 I_{2,j}(t) + \beta_3 x_{3,j,t} + \beta_4 x_{4,j,t} + O_{jt} + r_{jt}^T U_j,$$  \hspace{1cm} (3.20)

where $I_1(t)$ is the indicator variable for the change in BAC from 0.1 to 0.08 coded as 0 for $t \leq 36$ and 1 for $t > 36$ for all states in the study, and $I_{2,j}(t)$ the indicator variable taking the value 0 for $t < T_j$ and 1 for $t \geq T_j$, where $T_j$ is the month in which an administrative license revocation law was enacted. This potential confounder was enacted in seven of the states during the period of data used there. The other two regression variables are $x_{3,j,t}$, the number of Friday and Saturday nights, and a control series $x_{4,j,t}$, the log of other motor vehicle deaths (adjusted for population and seasonal factors), in month $t$ for state $j$. An offset term, $O_{jt}$, was used to adjust for unique state population trends and seasonal factors. Finally, $U_j \sim N(0, G)$ is the multivariate normal distribution with covariance matrix $G$ for the random effects on the selected covariates in $r_{jt}$. Further details concerning the rationale and definition of the model components mentioned earlier can be found in Bernat et al. (2004).

Because of the differences between the time periods over which the data were collected from each of the 17 states and because of the control series used for each state, the assumption that the 17 series are independent was considered reasonable. We also confirmed this assumption using cross-correlation analysis of Pearson residuals from individual GLARMA model fits to the 17 series. Bernat et al. (2004) discussed the likely impact that any serial dependence might have on their key conclusion that there is a statistically significant lowering of overall average single vehicle night time fatalities associated with the lowering of the legal BAC level but suitable software was not available at that time to assess this statistically.

We now illustrate the fixed effects plus GLARMA model defined in (3.18) on these data. The first step was to use the glarma package to fit the model (3.18) to each of the individual series using regression and offset terms as specified in (3.20). Examination of the PIT residual plots for each series suggested overfitting relative to the Poisson distribution. We identified that the main contributor to overfitting was the seasonal offset term used in (3.20); this had been determined using standard seasonal adjustment methods available in the PROC X-11 package in SAS for each series separately and based on only $n = 72$ months of data. As an alternative, we used parametric harmonic seasonal terms to all 17 response series with the other regressors and the population offset only. In almost all cases, these seasonal harmonics were not significant suggesting that use of the control series (without seasonal adjustment) and lag 12 autoregressive terms in the GLARMA model are sufficient for modeling seasonality. In view of this, we performed our reanalysis of these data using the regression specification in (3.20), but dropping the seasonal adjustment in both the offset $O_{jt}$ and logOMVD (so that the control series is now log other motor vehicle deaths per 100,000 population without seasonal adjustment), together with a GLARMA process for serial dependence. We compared various combinations of lags for the autoregressive and moving average terms in the GLARMA model to allow for both serial dependence at low orders and at seasonal lags. Based on AIC the overall best specification was an autoregression of order $p = 12$ with zero coefficients at lags 1 through 11. Hence we initially adopted this seasonal model for all series.
TABLE 3.1
Results for testing various fixed effect multiple GLARMA models for the Road Deaths Series in 17 U.S. States

<table>
<thead>
<tr>
<th>Model</th>
<th>$-2\log L$</th>
<th>$S$</th>
<th>$G^2$</th>
<th>d.f.</th>
<th>p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>FE-I: Unrestricted</td>
<td>5345.69</td>
<td>92</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>FE-II: $\phi_i's$ in 6 groups</td>
<td>5347.87</td>
<td>81</td>
<td>$G^2_{II , \text{v I}} = 2.18$</td>
<td>11</td>
<td>0.998</td>
</tr>
<tr>
<td>FE-III: BAC, ALR, FS same</td>
<td>5391.76</td>
<td>43</td>
<td>$G^2_{III , \text{v II}} = 43.89$</td>
<td>38</td>
<td>0.236</td>
</tr>
<tr>
<td>FE-IV: BAC, ALR, FS, lnOMVD same</td>
<td>5436.57</td>
<td>27</td>
<td>$G^2_{IV , \text{v III}} = 44.81$</td>
<td>16</td>
<td>0.00015</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$G^2_{IV , \text{v II}} = 88.61$</td>
<td>54</td>
<td>0.0021</td>
</tr>
</tbody>
</table>

We next check whether the regression coefficients in $W_t$ vary significantly between individual states. We begin with the overall unrestricted fit to all 17 states. We refer to this as Model FE-I, which has $-2\log L = 5345.69$ with $S = 92$ parameters. Examination of the individual estimates $\hat{\phi}_{12}$ suggested that they could be simplified as follows: Group 1 (State 11, $\hat{\phi}_{12} = -0.081 \pm 0.045$), Group 2 (States 1, 6, 7, 9, 10, 12:17, $\hat{\phi}_{12} = 0.005 \pm 0.013$), Group 3 (States 2, 4, $\hat{\phi}_{12} = 0.066 \pm 0.015$), Group 4 (State 5, $\hat{\phi}_{12} = 0.212 \pm 0.087$), Group 5 (State 8, $\hat{\phi}_{12} = 0.401 \pm 0.096$), Group 6 (State 5, $\hat{\phi}_{12} = 0.545 \pm 0.219$) in which, at most, 6 $\phi_{12}$ coefficients are significant.

The model with the $\phi_{12}$ restricted to these groups is referred to as Model FE-II in Table 3.1. Using the likelihood ratio test we obtain $G^2_{II \, \text{v I}} = 2.18$ on 11 d.f.; hence, restriction of the $\phi_{12}$ would not be rejected. From this model, we then examined whether or not some or all of the regression coefficients (other than the intercept which does vary substantially between states) take common values across all 17 states. Model FE-III restricts the coefficients for BAC, ALR, Friday–Saturday to be the same and (see Table 3.1) $G^2_{III \, \text{v II}} = 43.89$ on 38 d.f. and associated $p$-value of 0.24, which is not sufficiently strong evidence to suggest that the impact of these variables differs between individual states in a statistically significant way. Next, in Model FE-IV, log OMVD was allowed to differ between states. Compared with Model FE-III or Model FE-II, this risk control variable is strongly statistically significant between states with $G^2_{IV \, \text{v III}} = 44.81$ on 16 d.f. and associated $p$-value of 0.00015 and $G^2_{IV \, \text{v II}} = 88.61$ on 54 d.f. and associated $p$-value of 0.0021.

Hence, Model FE-III provides a useful summary of the commonality or otherwise of regression variable impacts on single vehicle night time road deaths across the 17 states. The fitted parameters and associated standard errors are reported in Table 3.2. The six groups for $\phi_{12}$ could be reduced to four by removing the nonsignificant cases of Groups 1 and 2. We did not pursue this here, preferring to move onto the use of a random effects analysis. The impact of lowering the legal BAC level is estimated to be $\hat{\beta}_2 = -0.072 \pm 0.022$ confirming the statistical significance of this association found in Bernat et al. (2004).

The fixed effects GLARMA model analysis provides a good starting point for the random effects GLARMA modeling that we turn to in the next section. In particular, it seems plausible from the results of Table 3.1 that random effects will be needed for the intercept term and the log OMVD term, but not for BAC, ALR, or Friday–Saturday effects. The parameter values reported for Model FE-III in Table 3.2 can provide useful starting values for the random effects model fitting. For fixed effects, we use the point estimates of coefficients for predictors that are common to all series, while for predictors that vary between series, we use the mean values of point estimates of the coefficients.
TABLE 3.2
Parameter estimates for the random effects model for the Road Deaths Series in 17 U.S. States

<table>
<thead>
<tr>
<th></th>
<th>RE-IV No GLARMA Random Effects</th>
<th>FE-III Multiple GLARMA Fixed Effects</th>
<th>RE-III Multiple GLARMA Random Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>s.e.</td>
<td>Estimate</td>
</tr>
<tr>
<td>β₀ (intercept)</td>
<td>−1.649</td>
<td>0.116</td>
<td>−1.801</td>
</tr>
<tr>
<td>β₁ (BAC change)</td>
<td>−0.054</td>
<td>0.022</td>
<td>−0.072</td>
</tr>
<tr>
<td>β₂ (ALR term)</td>
<td>−0.063</td>
<td>0.035</td>
<td>−0.011</td>
</tr>
<tr>
<td>β₃ (Frid-Sat)</td>
<td>0.032</td>
<td>0.011</td>
<td>0.037</td>
</tr>
<tr>
<td>β₄ (logOMVD)</td>
<td>0.395</td>
<td>0.063</td>
<td>0.314</td>
</tr>
<tr>
<td>Intercept RE s.d.</td>
<td>0.241</td>
<td>0.053</td>
<td>0.209</td>
</tr>
<tr>
<td>logOMVD RE s.d.</td>
<td>0.160</td>
<td>0.066</td>
<td>0.140</td>
</tr>
<tr>
<td>φ₈Gp1</td>
<td>—</td>
<td>—</td>
<td>−0.066</td>
</tr>
<tr>
<td>φ₈Gp2</td>
<td>—</td>
<td>—</td>
<td>0.010</td>
</tr>
<tr>
<td>φ₈Gp3</td>
<td>—</td>
<td>—</td>
<td>0.067</td>
</tr>
<tr>
<td>φ₈Gp4</td>
<td>—</td>
<td>—</td>
<td>0.216</td>
</tr>
<tr>
<td>φ₈Gp5</td>
<td>—</td>
<td>—</td>
<td>0.404</td>
</tr>
<tr>
<td>φ₈Gp6</td>
<td>—</td>
<td>—</td>
<td>0.531</td>
</tr>
</tbody>
</table>

2loglikelihood 5552.173 5391.1 5505.082

The results labeled RE-IV is that reported in Bernat et al. (2004) using SAS PROC NLMIXED. The results labeled FE-III is the final fixed effects multiple GLARMA model discussed in Section 3.5.2. The results labeled RE-III is the final random effects multiple GLARMA model discussed in Section 3.6.3.

Note: Values reported against the intercept β₀ and the logOMVD term β₄ are averages of the 17 individual values obtained while the values in the rows labeled “Intercept RE” and “logOMVD RE” are the standard deviations of these individual estimates, respectively.

3.6 Random Effects Multiple GLARMA Model

3.6.1 Maximum Likelihood Estimation

Let $W_j$ be defined as in (3.15), where $U_j$ are multivariate normal. Let $\theta = (\beta^{(1)}, \ldots, \beta^{(J)}, \tau^{(1)}, \ldots, \tau^{(J)}, \lambda)$ now be the collection of parameters in the GLARMA models and the random effects parameters. The joint log-likelihood is now

$$l(\theta) = \sum_{j=1}^{J} l_j(\beta^{(j)}, \tau^{(j)}, \lambda), \quad (3.21)$$

where

$$l_j(\beta^{(j)}, \tau^{(j)}, \lambda) = \log \int_{\mathbb{R}^d} \exp(l_j(\beta^{(j)}, \tau^{(j)}|u)g_u(u; \Sigma(\lambda))du, \quad (3.22)$$
and $g_{\text{U}}(\mathbf{u}; \Sigma(\lambda))$ is the multivariate normal density. To proceed further, we parameterize the covariance matrix as $\Sigma = LL^T$ where $L$ is lower triangular and let $U_j = L \xi_j$ where $\xi_j$ are independent $N(0, I_d)$. Let $\lambda = \text{vech}(L)$ be the half-vectorisation. With this parameterization, rewrite $W_{jt}$ in (3.5) linearly in terms of $\lambda$ as

$$W_{jt} = x_{jt}^T \beta^{(j)} + \text{vech} \left( \xi_j r_{jt}^T \right)^T \lambda + Z_{jt}. \quad (3.23)$$

The log-likelihood (3.22) becomes

$$l_j(\beta^{(j)}, \tau^{(j)}, \lambda) = \log \int_{\mathbb{R}^d} \exp \left( l_j \left( \beta^{(j)}, \tau^{(j)}, \lambda \mid \zeta \right) g(\zeta) \right) d\zeta \quad (3.24)$$

where $g(\zeta)$ is the $d$-fold product of the standard normal density and

$$l_j \left( \beta^{(j)}, \tau^{(j)}, \lambda \mid \zeta \right) = \sum_{t=1}^n \left[ y_{jt} W_{jt} - a_{jt} b(W_{jt}) \right] + \sum_{t=1}^n c(y_{jt}).$$

Note that (3.23) is in the same form as (3.3) but the parameters $\lambda$ are treated as regression parameters for any fixed value of the vector $\xi$ and the random effects covariates $r_{jt}^T$.

The representation of the random effects covariance matrix as $\Sigma = LL^T$ allows the parameter $\lambda$ to enter into the conditional log-likelihood linearly and without bounding constraints. Both properties enable existing GLARMA software to calculate the log-likelihood and derivatives with respect to the parameters. When some elements of $\Sigma$, and hence $L$, are specified as zero to reflect zero covariance between some of the random effects, $\lambda$ is the half vectorization of $L$ with the structural zeros removed. Covariance matrices in which certain combinations of random effects are specified to be zero cannot be represented in this form. However, these can often be accommodated by reordering the random effect variables and setting the appropriate elements of $L$ to zero.

### 3.6.2 Laplace Approximation and Adaptive Gaussian Quadrature

For any fixed $\theta$, computation of the log-likelihood $l(\theta)$ requires calculation of the $J$ integrals defined in (3.24). We now outline an approximate method based on the Laplace approximation and adaptive Gaussian quadrature (AGQ). The integral in (3.24) can be rewritten as

$$L_j(\theta) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \exp \left( F_j(\zeta \mid \theta) \right) d\zeta$$

where the exponent is considered as a function of $\zeta$ for fixed parameters $\theta$ and is defined as

$$F_j(\zeta \mid \theta) = \sum_{i=1}^{n_j} \left[ y_{ij} W_{ij}(\zeta; x_{ij}, \theta) - a_{ij} b(W_{ij}(\zeta; x_{ij}, \theta)) + c(y_{ij}) \right] - \frac{\zeta^T \zeta}{2}, \quad (3.25)$$

© 2016 by Taylor & Francis Group, LLC
where

\[ W_{jt}(\zeta; x_{jt}, \theta) = \left( r_{jt}^T L \right) \zeta + x_{jt}^T \beta^{(i)} + Z_{jt} \]  

(3.26)

is treated as a function of \( \zeta \) for \( x_{jt}^T \beta^{(i)} \) fixed. To find the Laplace approximation, we expand the exponent \( F(\zeta) \) around its modal value in a second-order Taylor series, and ignore the remainder. The resulting integral can be obtained in closed form. Note that \( Z_{jt} \) in (3.26) is a function of \( \zeta \). Hence, the contribution to the first and second derivatives from the summation term in (3.25) required for the Taylor series expansion of \( F_j \) need to be calculated using the GLARMA software with \( \zeta \) treated as a regression parameter for covariates \( (r_{jt}^T L) \) and fixing \( x_{jt}^T \beta^{(i)} \) as the offset term. To find the modal value, we need to find \( \zeta_j^* \) which solves

\[ \frac{\partial}{\partial \zeta_j} F_j(\zeta_j^*) = 0. \]

The Newton–Raphson method is used to find \( \zeta_j^* \) and, at convergence, we set

\[ \Sigma_j^* = \left\{ \frac{\partial^2}{\partial \zeta \partial \zeta^T} F_j(\zeta_j^*) \right\}^{-1}. \]

Since \( \frac{\partial^2}{\partial \zeta \partial \zeta^T} F_j(\zeta_j^*) \) is almost surely positive definite for the canonical link exponential family, the Newton–Raphson method will converge to the modal solution from any starting point; we use \( \zeta_j^{(0)} = 0 \) to initiate the recursions.

The Laplace approximation gives the approximate log-likelihood for the \( j \)th state as

\[ \tilde{l}_j^{(1)}(\theta) = \log \det(\Sigma_j^*(\theta))^{1/2} + F_j(\zeta_j^*(\theta)), \]

which can be combined to give the overall approximate log-likelihood as

\[ \tilde{l}^{(1)}(\theta) = \sum_{j=1}^{J} \tilde{l}_j^{(1)}(\theta) \]  

(3.27)

AGQ methods can be used to improve the approximation as has been successfully done for likelihoods in other statistical models such as nonlinear and non-Gaussian mixed effects modeling. This approach is implemented in a number of widely used software systems as the default method—see Pinheiro and Bates (1995) and Pinheiro and Chao (2006) for examples. Our implementation of AGQ follows that of Pinheiro and Chao (2006). It relies on the mode, \( \zeta_j^* \), and \( \Sigma_j^* \) used in the Laplace approximation to center and scale \( Q \) quadrature points in each of \( d \) coordinates resulting in integrands evaluated at \( d^Q \) points. When \( Q = 1 \), the Laplace approximation is obtained.
The AGQ approximation to the $j$th integral is denoted by $\tilde{L}_j(Q)(\theta)$, with corresponding approximation to the overall likelihood as

$$\tilde{l}(Q)(\theta) = \sum_{j=1}^{J} \log \tilde{L}_j(Q)(\theta). \quad (3.28)$$

Since $z_j^*(\theta)$ and $\Sigma_j^*(\theta)$ are functions of the unknown parameters $\theta$, it is necessary to recompute the Laplace approximation at each iterate of $\theta$ to maximize (3.28).

Maximizing (3.28) using the optimizer optim in R proved to be very slow and unreliable for our applications. An alternative was to use Fisher Scoring or Newton–Raphson updates based on numerical derivatives obtained using the R package numDeriv. This also proved to be very slow. Analytical derivatives require implicit differentiation of $\zeta_j^*(\theta)$ and $\Sigma_j^*(\theta)$ which results in complex expressions requiring substantial modification to the current GLARMA software. We next describe an alternative approach that avoids all of these issues.

First derivatives of the log-likelihood (3.19) with respect to unknown parameters are

$$\dot{i}_j(\theta) = \frac{\partial}{\partial \theta} l(\theta) = \sum_{j=1}^{J} \frac{1}{l_j(\theta)} \int_{R^d} \frac{\partial}{\partial \theta} \left[l_j(\theta|\zeta)\right] \exp(l_j(\theta|z)g(\zeta)) d\zeta. \quad (3.29)$$

Second derivatives, $\ddot{i}_j(\theta)$, are also easy to derive and involve more integrals to be approximated. For any fixed $\zeta$, the integrands in these derivative expressions can be calculated recursively using the unpackaged form of single series GLARMA software. If $S$ denotes the number of parameters in $\theta$, then there are $J \times (1+S+2S(S+1)/2) = J(1+S)^2$, $d$-dimensional integrals to calculate in order to implement the Newton–Raphson method. For instance, for the final model for the BAC example (Model RE-III) with two uncorrelated random effects, we have $J = 17$, $S = 10$ requiring calculation of $2057$ $d = 2$ dimensional integrals at each step of the Newton–Raphson iterations. Fisher scoring is not available here because the summation to compute the whole likelihood is over $J$; hence, insufficient outer products of first derivative vectors would result in an ill-conditioned approximation to the second derivative matrix unless $J$ is quite large.

In our experience, for long longitudinal data applications, the Laplace approximation can provide quite accurate single-point approximations to the integrals required for the likelihood itself. However, the first and second derivatives have integrands that are certainly not positive, nor are they unimodal, and so a single-point integral approximation is inadequate. However, AGQ can provide multipoint approximations for the integrals required for derivatives. In our experience, surprisingly few quadrature points are required to get approximations to the likelihood and the first and second derivatives which are sufficiently accurate for convergence to the optimum of the likelihood and which provide accurate standard errors for inferential purposes. We denote the estimates of $\dot{i}(\theta)$ and $\ddot{i}(\theta)$ obtained by applying AGQ with $Q$ nodes by $\tilde{\dot{i}}(Q)(\theta)$ and $\tilde{\ddot{i}}(Q)(\theta)$, respectively. The same quadrature points and weights that are used for $\tilde{l}(Q)(\theta)$ are also used to obtain $\tilde{\dot{i}}(Q)(\theta)$ and $\tilde{\ddot{i}}(Q)(\theta)$ using one pass of the GLARMA software.
Summary of algorithm to approximate likelihood and derivatives:

1. Initialize parameter value $\theta^{(k)}$
2. For each $j = 1, \ldots, J$
   a. Use GLARMA software treating $\zeta$ as a parameter to find the derivatives needed to find the $J$ Laplace approximations $\zeta^*_j, \Sigma^*_j$.
   b. Select $Q$ quadrature points in each of $d$ directions, relocate, and scale these using $\zeta^*_j$ and $\Sigma^*_j$
   c. Apply GLARMA software to calculate the integrands at the $dQ$ integrating points in order to estimate the likelihood and first and second derivatives at $\theta^{(k)}$.
3. Assemble the complete likelihood and derivatives over the $J$ cases.
4. Use Newton–Raphson iteration to update $\theta^{(k)} \rightarrow \theta^{(k+1)}$. Repeat at Step 2 until convergence.

3.6.3 Application to Road Deaths Data

We used Model FE-III (central columns of Table 3.2) as the starting point for the random effects analysis. The correlation between the 17 pairs of intercepts and coefficients of the logOMVD variables was $\text{Corr} \left( \hat{\beta}_{0,j}, \hat{\beta}_{4,j} \right) = 0.531$ which is significantly different from zero. As a result, we used a correlated bivariate random effect for the intercept and this regressor. The results of fitting this model (Model RE-I) are given in Table 3.3. The point estimate of the parameter controlling the correlation between these two terms was not significant ($\hat{L}_{12} = 0.101 \pm 0.107$) and so we removed the correlation between the random effects to arrive at model RE-II. The likelihood ratio test (see Table 3.3) confirmed that these were not correlated. The random effect variance for the log OVMD regressor is marginally significant compared to its standard error. Use of a likelihood ratio test does not reject this simplification, even after adjusting for the fact that this is a test on the boundary using the standard methodology for mixed effects variance testing as described in Fitzmaurice et al. (2012). Similar to the fixed effects analysis, Groups 1 and 2 showed no significant autocorrelation; however, the large autocorrelation for the sixth group was no longer significant. Since there is clearly potential in these models for serially correlated effects to interact or trade-off with regression random effects, our next step was to refit the model after removing autocorrelation terms for Groups 1, 2, and 6. This resulted in Model RE-III and the likelihood ratio test confirms that autoregressive terms are not required for these three Groups. The final states for which significant autocorrelation is required are

<table>
<thead>
<tr>
<th>Model</th>
<th>$\phi_{12}$ Groups</th>
<th>Random Effects</th>
<th>$-2\log L$</th>
<th>$S$</th>
<th>$G^2$</th>
<th>d.f.</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RE-I</td>
<td>6 levels</td>
<td>2 correlated</td>
<td>5500.85</td>
<td>14</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>RE-II</td>
<td>6 levels</td>
<td>2 uncorrelated</td>
<td>5501.87</td>
<td>13</td>
<td>$G^2_{II \text{ vs } I} = 0.98$</td>
<td>1</td>
<td>0.32</td>
</tr>
<tr>
<td>RE-III</td>
<td>3 levels</td>
<td>2 uncorrelated</td>
<td>5505.08</td>
<td>10</td>
<td>$G^2_{III \text{ vs } II} = 3.21$</td>
<td>3</td>
<td>0.36</td>
</tr>
<tr>
<td>RE-IV</td>
<td>No $\phi_{12}$</td>
<td>2 uncorrelated</td>
<td>5552.17</td>
<td>7</td>
<td>$G^2_{IV \text{ vs } III} = 47.09$</td>
<td>3</td>
<td>$3 \times 10^{-10}$</td>
</tr>
</tbody>
</table>
states 2, 4, 5, and 8 (corresponding to California, Florida, Idaho, and Maine). We next compared this model with the purely random effects model (the analogue of what was fit in Bernat et al., 2004), which is labeled Model RE-IV. The likelihood ratio test overwhelmingly rejects the hypothesis that autocorrelation terms can be removed from the model for these four states.

Did our inclusion of autoregressive terms impact the original conclusions of Bernat et al. (2004) concerning the strength and significance of the BAC intervention on single vehicle night time road deaths? They obtained \( \hat{\beta}_1 = -0.052 \pm 0.021 \) (\( p \)-value = 0.013). In the analysis presented here, we have removed the seasonal offset terms from the response distribution and the logOMVD control variable for reasons discussed earlier. The analogous result is Model RE-IV and for that model \( \hat{\beta} = -0.054 \pm 0.022 \) (\( p \)-value = 0.013), a very similar finding to that in Bernat et al. (2004).

We report estimates for all parameters of Model RE-III as the final column pair in Table 3.2. Inclusion of significant serial dependence terms where needed has actually increased the size of the point estimate of the BAC effect relative to Model RE-IV (no serial dependence model) and left the standard error unchanged, resulting in a reduced \( p \)-value of 0.006 for this term and hence suggesting that the original finding of the significance of the BAC association may have been conservative.

3.6.4 Computational Speed and Accuracy

3.6.4.1 Impact of Increasing the Number of Quadrature Points

In several applications of the earlier method, we have found that usually \( Q = 3, 5, \) or 7 quadrature points in each of the \( d \) random effects coordinates are sufficient for use in optimizing the likelihood and for obtaining accurate standard errors and likelihood values and likelihood ratio statistics.

3.6.4.2 Comparison with Numerical Derivatives

For one iteration of the Newton–Raphson procedure, near convergence to the maximum likelihood, using first and second numerical derivatives calculated using the numDeriv package is of the order of 300 times longer than using the AGQ method for calculating derivatives that we propose. Use of \texttt{optim} for convergence and evaluation of the Hessian for standard error calculations were similarly slow. There was no substantial loss of accuracy either for convergence of the maximum likelihood updates or in the standard errors.

In summary, the AGQ method proposed here to calculate derivatives of the log-likelihood is two orders of magnitude faster than using \texttt{optim} without derivative or numerical derivatives based on numDeriv with no substantial loss of accuracy even for \( Q = 3 \). Use of the AGQ method makes it feasible to fit combined random GLARMA random effects models for long longitudinal data. We have experienced similar comparisons of speed and accuracy in more complex settings, such as the analysis of 32 binary time series of length 393 arising in the musicology study discussed earlier in which up to four random effects were included in the models and another study of 49 times series of length 336 of suicide counts for which a Poisson response was appropriate and up to three random effects were included in the model.
3.7 Interpretation of the Serially Dependent Random Effects Models

For single series GLARMA models, means, variances, and serial covariances for the state process \( \{ W_t \} \) can be readily derived using the definition of \( Z_t \) in (3.4). For the Poisson or negative binomial response GLARMA plus random effects model, the marginal interpretation of the fixed effects coefficients is approximately equal to the conditional interpretation since

\[
E(Y_{jt}) \approx \exp \left( x_{0,t}^T \beta^{(0)} + x_{jt}^T \beta^{(j)} \right) \exp \left( \frac{\sigma^2}{2} + \frac{\gamma^2}{2} \right)
\]

by simple extension of the argument in Davis et al. (2003).

For binomial and Bernoulli responses, calculation of means, variances, autocovariances for the response series and interpretation of regression coefficients are not straightforward. This is a typical issue for interpretation of random effects models and transition models in the binomial or Bernoulli case—see Diggle et al. (2002) for example.

3.8 Conclusions and Future Directions

This chapter has reviewed the fitting of GLARMA models for single time series of exponential family response distributions and illustrated this on some binary and binomial series arising in a study of listener responses to music features. Extensions and utilization of single series GLARMA modeling ideas and software to the long longitudinal data setting were explained. Two approaches to providing a combined analysis of all series in a panel of responses were considered. The first approach was based on a constrained fit across the panel using single series GLARMA software and allowed testing of parameter similarity in each series across the panel of series. The second approach modeled between series parameter variation using random effects. Again, single series GLARMA software can be used to compute a modal approximation to the integrals that constitute the likelihood when there are random effects. This modal approximation is based on the Laplace approximation and this can also be calculated using the GLARMA single series software. The use of AGQ to compute the very large number of integrals required to compute the first and second derivatives of the combined likelihood is explained and, somewhat surprisingly, these can be accurately and speedily computed using a small number of quadrature points, hence making the optimization of the likelihood based on Newton–Raphson iteration feasible in practice. The speed of this method compared to those based on numerical derivatives of standard optimizers is several hundred times faster per iteration. We illustrated the multiple independent time series approaches on a set of long longitudinal data arising in the study of the association between lowering the legal blood alcohol level in drivers on road deaths in 17 U.S. States.

The approach taken here allows considerable flexibility on the specification of the serial dependence in the individual time series. The examples presented here clearly require that flexibility. We know of no other current methods that have this flexibility; those that we
have reviewed appear to continue assuming short longitudinal trajectories in which case it is difficult to allow flexibility in serial dependence specifications. For long longitudinal data, this restriction can be avoided as we have demonstrated in the examples presented herein.

We have also extended the random effects approach to parameter-driven models for individual time series serial dependence, again based on the use of a Laplace approximation as in Davis and Rodríguez-Yam (2005) and AGQ for the random effect integration. We aim to extend GLARMA random effects models covered in this chapter and in the parameter-driven version to allow for random effects on serial dependence parameters. This seems to us a natural extension within the modeling perspective of random effects for between series parameter variation.

There is an obvious need for rigorous asymptotic theory to properly justify the statistical inference that we have presented based on these new models. However, this will rely on similar theory being developed for individual observation-driven time series models, something that remains underdeveloped.

Finally, we have not discussed forecasting for GLARMA models in this review nor the somewhat related issue of missing data. GLARMA models, requiring recursive calculation of the state equation, are computationally intensive to forecast beyond one or two time points. This has implications for the computation of the likelihood when there are missing data, as it requires a conditional distribution of the response after gaps.

Acknowledgments
The Music Listener data used in Section 3.3 was provided by Roger Dean of the MARCS Institute at the University of Western Sydney and the U.S. Road Deaths data used in Sections 3.5.2 and 3.6.3 was that used in Bernat et al. (2004) and its source is acknowledged there. I would also like to thank Chris McKendry, my previous Honour’s degree student, for his assistance in developing the adaptive Gaussian quadrature multiple independent random effects approach discussed in this chapter. Finally, helpful comments from a referee improved the clarity of the presentation.

References


for multilevel generalized linear mixed models. Journal of Computational and Graphical Statistics,

R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation
for Statistical Computing, Vienna, Austria.


State University, Department of Statistics, Fort Collins, CO.


In R. A. Davis, S. H. Holan, R. Lund and N. Ravishanker, eds., Handbook of Discrete-Valued Time
Series, pp. 77–100. Chapman & Hall, Boca Raton, FL.


Wang, C. and Li, W. K. (2011). On the autopersistence functions and the autopersistence graphs of


in long sequences: Methods and applications to a driving study. The Annals of Applied Statistics,
6(1):27–54.