Bayesian Analysis (2007) **2**, Number 4, pp. 817–842

Bayesian Model Diagnostics Based on Artificial Autoregressive Errors

Mario Peruggia*

Abstract. Hierarchical Bayes models provide a natural way of incorporating covariate information into the inferential process through the elaboration of regression equations for one or more of the model parameters, with errors that are often assumed to be i.i.d. Gaussian. Unfortunately, building adequate regression models is a complicated art form that requires the practitioner to make numerous decisions along the way. Assessing the validity of the modeling decisions is often difficult.

In this article I develop a simple and effective device for ascertaining the quality of the modeling choices and detecting lack-of-fit. I specify an artificial autoregressive structure (AAR) in the probability model for the errors that incorporates the i.i.d. model as a special case. Lack-of-fit can be detected by examining the posterior distribution of AAR parameters. In general, posterior distributions that assign considerable mass to a region of the AAR parameter space away from zero provide evidence that apparent dependencies in the errors are compensating for misspecifications of some other aspects (typically conditional means) of the model. I illustrate the methodology through several examples including its application to the analysis of data on brain and body weights of mammalian species and response time data.

Keywords: Allometry, Asymptotic normality, Autocorrelation, Hierarchical models, Response times

1 Introduction

It is common statistical practice to evaluate the fit of a model by examining the behavior of the realized residuals. For example, in multiple linear regression analyses (Weisberg 1985), the residuals are often plotted against one of the covariates or against the predicted values of the corresponding cases. The plots are then scanned for the presence of non-random patterns that would call into question the assumption of an i.i.d. error structure. Realized residuals incompatible with i.i.d. errors may indicate inadequacies in the model and suggest avenues to improve the fit (e.g., transformations of the response and/or some of the predictors, introduction of new predictors, removal of old predictors, etc.). The use of residual plots and related diagnostics (added variables plots, q-q plots, etc.) is well established in the frequentist arena and virtually all statistical software packages provide easy access to computational and graphical tools to perform residual-based model checks.

*Department of Statistics, The Ohio State University, Columbus, OH, mailto:peruggia@stat.osu.edu

Hierarchical Bayes models (Carlin and Louis 2000; Gelman, Carlin, Stern, and Rubin 2004) provide a natural way of incorporating covariate information into the inferential process through the elaboration of regression equations for one or more of the model parameters, with errors that are often assumed to be i.i.d. Gaussian. The approach is conceptually simple and the development of efficient Markov Chain Monte Carlo (MCMC) computational techniques has made its implementation feasible even in rather complicated settings, especially since flexible and user-friendly computational environments such as BUGS (Spiegelhalter, Thomas, Best, and Gilks 1996a) and WinBUGS (Spiegelhalter, Thomas, Best, and Lunn 2003) have become available. Unfortunately, building adequate regression models is a complicated art form that requires the practitioner to make numerous hard decisions along the way. By comparison to its frequentist counterpart, the contents of the Bayesian model-building toolbox look quite scanty. All the more so when the regression model to be constructed is for a parameter at a higher stage in the hierarchy—a quantity for which, typically, the intuition is not as well developed as the intuition for quantities that are directly observable.

In this article I demonstrate that residual analysis is also a powerful diagnostic tool for Bayesian model building. Rather than examining the realized residuals directly (a difficult task, especially when the regression models are for parameters that appear at higher levels of the model hierarchy), I introduce an artificial autoregressive (AAR) structure in the probability model for the errors that incorporates the i.i.d. model as a special case. Lack-of-fit can be detected by examining the posterior distribution of the AAR parameters. In general, posterior distributions that assign considerable mass to a region of the AAR parameter space away from zero provide evidence that apparent dependencies in the errors are compensating for misspecifications of some other aspects (typically conditional means) of the model.

Econometricians have long recognized that, given a data vector $\boldsymbol{Y}$ whose elements are collected sequentially over time, the misspecification of the design matrix $\boldsymbol{X}$ in a linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{e}$ can induce autocorrelation in the elements of $\boldsymbol{e}$ (Judge, Griffiths, Hill, and Lee 1980). Within this context, the Durbin and Watson test is the earliest and most popular frequentist technique for detecting autocorrelation in the residuals (Durbin and Watson 1950, 1951). More recently, a Bayesian comparison of regression models with i.i.d. and AR(1) errors was presented in Carota (1998). This work was further extended to encompass comparisons with higher order autoregressive error structures (as well as moving average and nonparametric error structures) in Carota (2005). In econometric applications the misspecification of the design matrix is typically difficult to remedy, due in part to the fact that some of the important regressors may not be known or measurable. Thus, econometricians end up retaining the misspecified $\boldsymbol{X}$ and are principally concerned with the impact that the presence of autocorrelation has on the estimation of $\boldsymbol{\theta}$ and how to optimally estimate $\boldsymbol{\theta}$ in its presence. Bayesian estimation in this context is considered, for example, by Zellner and Tiao (1964) using locally uniform priors on the model parameters and in Judge et al. (1980, chap. 5) using Jeffreys' prior.

The basic premise that model misspecification leads to autocorrelated errors also motivates the development of the diagnostic tools presented in this article. However, the perspective here is essentially opposite than in econometrics: the aim of the diagnostics is to assess lack-of-fit, uncover specific shortcomings of the model, and suggest possible directions for further elaboration and improvement that would make a model with conditionally i.i.d. errors appropriate. The applicability of the AAR error structure to uncover lack-of-fit in the context of elaborate Bayesian hierarchical models is very flexible and can be readily implemented using popular MCMC simulation software. While the AAR structure implies that the observations are ordered, I need not require that they be collected over time. As the examples presented throughout the article will make clear, in certain cases the ordering will be indeed determined by the time when the observations arose, in others it will be suggested naturally by the specific application, and in others yet it will be entirely arbitrary.

An asymptotically normal approximation to the posterior distribution of the AAR parameter based on an infill scheme provides an interpretative tool that outlines how bias and error variance play off against each other in the assessment of lack-of-fit. The practical strength of the proposed AAR diagnostic lies in its computational simplicity and in the fact that the lack-of-fit of a given regression structure is numerically quantifiable based on the posterior distribution of a *one-dimensional* model parameter. Beside examining graphically the posterior distribution of the AAR parameter, I suggest to compute numerical diagnostics defined in terms of tail probabilities. These are very quick to derive based on the MCMC output and provide easily interpretable summaries of lack-of-fit conditional on the evidence given by the data. Because of their computational simplicity and low dimensionality these summaries enable exploratory analyses for complex hierarchical models (e.g., analyses concerning joint aspects of lack-of-fit) that would be much harder to accomplish if based on a direct examination of the residuals. In practice, for hierarchical models involving many regression equations, a routine, preliminary screening based on the AAR device can be followed by a direct residual analysis for the fits that are identified as problematic, so as to uncover specific reasons for lack-of-fit. In addition to the appeal of computational simplicity, I show that there is useful information contained in the posterior distribution of the AAR parameters that cannot be recovered by a simple examination of the residuals.

The remainder of the article is organized as follows. Section 2 builds some intuition for the technique by considering an elementary example and deriving some asymptotic properties. Sections 3 and 5 evaluate the performance of the technique in a variety of settings using simulated and real data respectively. Section 4 illustrates how to quantify numerically the lack-of-fit on the basis of the posterior distribution of the AAR parameters. Section 6 contains a brief discussion and some concluding remarks.

2 Intuition and Asymptotic Considerations

In this section, I consider a simple model for which direct closed-form calculation can be performed and present some asymptotic results to develop intuition and motivate the development of model diagnostics based on the specification of AAR error structures.

Example 2.1. *I am planning to collect n i.i.d. observations, y_i, from a $N(\mu_y, \sigma_y^2)$*

distribution with known variance σ_y^2 and to make inference about the unknown mean μ_y. My friend Pythia claims she has the ability to make inferences without seeing the data and tells me that the value of the mean is equal to some fixed real number a. Pythia has a spotty track-record of reporting the right mean and I wish to verify her assertion by actually collecting the data and analyzing them. To do so, always assuming conditional independence unless stated otherwise, I specify the following model with AAR errors for the y_i. For $i = 1, \dots, n$, let

$$y_i = a + \eta_i,$$

where a is a fixed constant, and $\eta_1 \sim N(0, \sigma_1^2)$. Furthermore, for $i = 2, \dots, n$, let

$$\eta_i = \phi\eta_{i-1} + \epsilon_i,$$

with $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and $\phi \sim N(0, \sigma_\phi^2)$, where σ_1^2, σ_ϵ^2, and σ_ϕ^2 are fixed and known. The model has conjugate structure, and the posterior distribution of ϕ is easily seen to be normal with mean and variance

$$E(\phi|\boldsymbol{y}) = \frac{\sum_{i=2}^n (y_i - a)(y_{i-1} - a)}{(\sigma_\epsilon^2/\sigma_\phi^2) + \sum_{i=2}^n (y_{i-1} - a)^2} \quad \text{and} \quad V(\phi|\boldsymbol{y}) = \left[\frac{1}{\sigma_\phi^2} + \frac{1}{\sigma_\epsilon^2}\sum_{i=2}^n (y_{i-1} - a)^2\right]^{-1}.$$

Suppose the true mean of the y_i is $\mu_y = 0$. Then, denoting by ρ_y the lag-1 autocorrelation of the y_i (which is zero because of the independence assumption),

$$\begin{aligned} &lim_{n\to\infty}\frac{1}{(n-1)}\sum_{i=2}^n (y_i - a)(y_{i-1} - a) = \\ &= lim_{n\to\infty}\sum_{i=2}^n \frac{y_i y_{i-1}}{n-1} - a\sum_{i=2}^n \frac{y_i}{n-1} - a\sum_{i=2}^n \frac{y_{i-1}}{n-1} + a^2\,\frac{n-1}{n-1} \\ &= \rho_y - a\mu_y - a\mu_y + a^2 = a^2. \end{aligned}$$

Similar calculations show that $lim_{n\to\infty}(n-1)^{-1}\sum_{i=2}^n (y_{i-1} - a)^2 = \sigma_y^2 - 2a\mu_y + a^2 = \sigma_y^2 + a^2$. Thus, $lim_{n\to\infty}E(\phi|\boldsymbol{y}) = a^2/(a^2+\sigma_y^2)$ and $lim_{n\to\infty}(n-1)V(\phi|\boldsymbol{y}) = \sigma_\epsilon^2/(a^2+\sigma_y^2)$. The asymptotic posterior mean of ϕ is zero if Pythia is right (i.e., when $a = 0$). The further she is from the truth (i.e., the larger a^2 is), the closer the asymptotic posterior mean of ϕ will be to one. The asymptotic variance, in turn, is determined by the relative sizes of σ_ϵ^2 and $a^2 + \sigma_y^2$. These considerations suggest that draws from the posterior distribution of ϕ can be used to diagnose Pythia's accuracy. A histogram of posterior ϕ values centered somewhere in the neighborhood of zero would give no indication that her guess should be questioned.

From a more general perspective, let f and g be smooth functions on a finite closed interval, that, without loss of generality, can be taken to be $[0, 1]$. I will consider regular infill asymptotics, assuming that the true data generating mechanism is given by $y_i = f(i/n) + u_i$, where, for $i = 0, \dots, n$, the u_i are independent normal errors with mean zero and variance σ_u^2. The fitted model with AAR errors is given by $y_i = g(i/n) + \eta_i$, with $\eta_i = \phi\eta_{i-1} + \epsilon_i$, where the i.i.d. zero-mean normal noise process $\{\epsilon_i\}$ has variance σ_ϵ^2.

Let $r(t) = f(t) - g(t)$ denote the bias at the point t induced by fitting the wrong regression function g. Simple algebraic manipulations yield:

$$[r(i/n) + u_i] = \phi\,[r((i-1)/n) + u_{i-1}] + \epsilon_i.$$

Treating the process $z_i = [r(i/n) + u_i]$ as zero-mean stationary, ϕ is estimated consistently by

$$\widehat{\phi}_n = \frac{(n+1)^{-1}\sum_{i=1}^{n} z_i z_{i-1}}{(n+1)^{-1}\sum_{i=0}^{n} (z_i)^2}. \tag{1}$$

As n goes to infinity, the numerator of Equation (1) converges to the limit of

$$\sum_{i=1}^{n} \frac{r(i/n)r((i-1)/n)}{n+1} + \sum_{i=1}^{n} \frac{r(i/n)u_{i-1}}{n+1} + \sum_{i=1}^{n} \frac{r((i-1)/n)u_i}{n+1} + \sum_{i=1}^{n} \frac{u_i u_{i-1}}{n+1}.$$

By the definition of a Riemann sum, the first term in the summation converges to $\int_0^1 r^2(t)\,dt$. Also, because of the smoothness assumptions on f and g and of the various distributional assumptions, the SLLN implies that the last three terms converge to zero. Hence the numerator converges to $\int_0^1 r^2(t)\,dt$. Similarly, the denominator converges to $\int_0^1 r^2(t)\,dt + \sigma_u^2$. Thus, $\widehat{\phi}_n$ converges with probability one to

$$\frac{\int_0^1 r^2(t)\,dt}{\int_0^1 r^2(t)\,dt + \sigma_u^2}. \tag{2}$$

The expression in Equation (2) can be interpreted as a *signal to noise ratio* for detecting unmodeled trend: the integrated squared bias in the numerator quantifies the amount of unmodeled trend and the denominator quantifies the amount of unmodel trend plus the amount of noise in the data.

A full Bayesian model would typically specify a parametric family $g(t|\boldsymbol{\theta})$ to which $g(t)$ belongs as well as a prior distribution for the model parameters $\boldsymbol{\theta}$, ϕ, and σ_ϵ^2. For all fixed n, the model will define a likelihood $p_n(\boldsymbol{y}_n|(\boldsymbol{\theta}, \phi, \sigma_\epsilon^2))$ for the data $\boldsymbol{y}_n = (y_0, \ldots, y_n)$ collected under the regular infill scheme. Let $(\boldsymbol{\theta}^*, \phi^*, (\sigma_\epsilon^*)^2)_n$ be the value of $(\boldsymbol{\theta}, \phi, \sigma_\epsilon^2))$ that minimizes the Kullback-Leibler divergence of $p_n(\boldsymbol{y}_n|(\boldsymbol{\theta}, \phi, \sigma_\epsilon^2))$ from the true distribution of the data, $p_n(\boldsymbol{y}_n)$, let $(\boldsymbol{\theta}^*, \phi^*, (\sigma_\epsilon^*)^2) = \lim_{n\to\infty}(\boldsymbol{\theta}^*, \phi^*, (\sigma_\epsilon^*)^2)_n$, and let $V(\boldsymbol{\theta}, \phi, \sigma_\epsilon^2) = \lim_{n\to\infty}(n+1)[I_n(\boldsymbol{\theta}, \phi, \sigma_\epsilon^2)]^{-1}$, where $I_n(\boldsymbol{\theta}, \phi, \sigma_\epsilon^2)$ denotes the Fisher information matrix based on a sample of size $n+1$. Under suitable regularity conditions (see Gelman et al. (2004) for the case of i.i.d. observations), the joint posterior distribution of $\sqrt{n+1}\,[(\boldsymbol{\theta}, \phi, \sigma_\epsilon^2) - (\boldsymbol{\theta}^*, \phi^*, (\sigma_\epsilon^*)^2)]$ will converge to a normal distribution with mean $\mathbf{0}$ and covariance matrix $V[(\boldsymbol{\theta}^*, \phi^*, (\sigma_\epsilon^*)^2)]$. Marginally, the posterior of $\sqrt{n+1}\,(\phi - \phi^*)$ will converge to a univariate normal distribution with mean 0 and variance $(\sigma_\epsilon^*)^2/(\int_0^1 [r^*(t)]^2\,dt + \sigma_u^2)$, where

$$\phi^* = \frac{\int_0^1 [r^*(t)]^2\,dt}{\int_0^1 [r^*(t)]^2\,dt + \sigma_u^2} \tag{3}$$

The expression for the asymptotic mean ϕ^* is the same as expression (2) (and can be interpreted similarly), except that $r(t)$ is replaced by $r^*(t) = f(t) - g(t|\boldsymbol{\theta}^*)$, where $g(t|\boldsymbol{\theta}^*)$ is the member of the family $g(t|\boldsymbol{\theta})$ that best approximates $f(t)$ in the sense of minimizing $\int_0^1 (f(t) - g(t|\boldsymbol{\theta}))^2 \, dt$. An illustration of this asymptotic approximation is presented in Example 3.1.

3 Examples Using Simulated Data

In this section, I examine the finite sample and asymptotic performance of the AAR diagnostic device by presenting some simulated data examples. To explain the intuition behind the methodology, the examples in this section deal with simple models for which more direct diagnostics can be readily developed. In several cases the posterior distributions of the AAR parameters could be derived using low dimensional numerical integration, but I chose to estimate them by MCMC simulation to reflect the way in which the diagnostic device would typically be used for general hierarchical Bayes models.

Example 3.1. *This example illustrates the methodology in the context of a linear regression model. The data set is comprised of* $n = 101$ *pairs of observations* (x_i, y_i) *simulated from the model* $y_i = 1.0 + (0.1)x_i + u_i$, *where, for* $i = 0, \ldots, n$, $x_i = i$ *and the* u_i *are independent standard normal errors.*

Using BUGS, I fit several regression models with an AAR error structure and a possibly misspecified slope. For $i = 0, \ldots, n$, *I assume that* $y_i = \alpha + bx_i + \eta_i$, *where* b *is a fixed constant and* $\alpha \sim N(0, 10^3)$. *The errors* η_i *have the same basic AAR structure as in Example 2.1, but now the variances* σ_1^2 *and* σ_ϵ^2 *are given inverse gamma prior distributions. Specifically,* $\eta_1 \sim N(0, \sigma_1^2)$, *and, for* $i = 1, \ldots, n$, $\eta_i = \phi\eta_{i-1} + \epsilon_i$, *with* $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, $1/\sigma_1^2 \sim Gamma(2.5, 10.0)$, $1/\sigma_\epsilon^2 \sim Gamma(0.1, 0.1)$, *and* $\phi \sim N(0, 1.0)$.

Figure 1 displays the data and five fitted regression lines corresponding to five specifications of the slope b *in the regression equation. The five values of* b *under consideration are* $b_1 = 0.10$, $b_2 = 0.09$, $b_3 = 0.07$, $b_4 = 0.05$, *and* $b_5 = 0.00$. *For* $1 \leq j \leq 5$, *the equation of the fitted regression line is given by* $y = \widehat{\alpha}_j + b_j x$, *where* $\widehat{\alpha}_j$ *is an estimate of the posterior mean of* α *obtained by averaging* $1,000$ *posterior draws* $\alpha_j^{(k)}$, $k = 1, \ldots, 1,000$, *generated by BUGS based on the model with slope* b_j. *In each of the five cases, I discarded the first* $10,000$ *draws and subsampled every tenth of the subsequent* $10,000$ *ones. Denoting by* $\bar{y}$ *the sample mean of the observed* y *values, each estimated intercept* $\widehat{\alpha}_j$ *arises from the shrinkage of the value* $\bar{y} - b_j * 50$ *towards the prior mean of zero.*

As evidenced by the figure, the horizontal line (corresponding to slope b_5*) provides the worst fit. As* x *increases, the residuals from the fitted line progressively shift from having large negative values to having large positive values (see the bottom left panel of Figure 2). This increasing trend is due to the erroneous specification of the conditional mean for the observations. Failure to remove the trend induces strong positive autocorrelations in the residuals that are captured by the AAR model. As a consequence, the posterior distribution of the AAR parameter* ϕ, *illustrated by the histogram in the bot-*

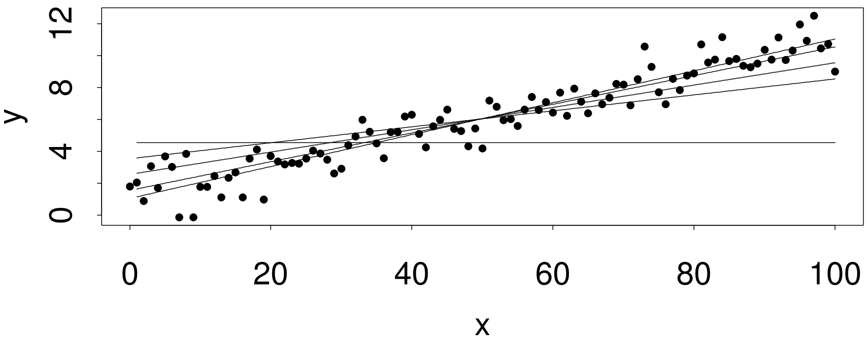


Figure 1: Simulated Data and Five Fitted Lines for the Linear Regression Model of Example 3.1.

tom right panel of Figure 2, is concentrated near one. Based on the 1,000 draws from the posterior distribution of ϕ used to construct the histogram, the posterior mean is estimated at 0.93 and the 95% equal-tailed posterior probability interval is $[0.83, 1.01]$*).*

The best fitting line appears to be the one rising most rapidly (corresponding to slope $b_1 = 0.1$*). This, of course, is not surprising because the data were simulated from a regression model with slope equal to 0.1. In this case, a model of independence for the AAR error structure (*$\phi = 0$*) would be appropriate as evidenced by the absence of trends in the residual plot in the top left panel of Figure 2. Accordingly, the posterior distribution of the AAR parameter ϕ, illustrated by the histogram in the top right panel of Figure 2, is centered in the vicinity of zero. Based on the 1,000 draws from the posterior distribution of ϕ used to construct the histogram, the posterior mean is estimated at 0.02 and the 95% equal-tailed posterior probability interval is* $[-0.19, 0.24]$*. The intermediate cases corresponding to slopes b_2, b_3, and b_4, are characterized by correspondingly worsening fits (see the left intermediate panels of Figure 2) and posterior distributions of the AAR parameter ϕ that become progressively more concentrated about a central location that shifts from zero towards one (see the right intermediate panels of Figure 2).*

The asymptotic posterior mean ϕ^* of ϕ under regular infill can be computed using Equation (3), adjusting for the fact that the unit interval in the original derivation of Section 2 is here replaced by an interval of length 100. Note that the initialization of the AAR process with the specification of the variance of η_1 is not considered explicitly in the derivations of Section 2 because it does not impact on the asymptotic arguments. Observing that $\int_0^{100}(1 - 0.10t - (\alpha + bt))^2\,dt$ *is minimized at* $\alpha^* = 6 - 50b$*, yields* $100^{-1}\int_0^{100}[r^*(t)]^2\,dt = (10b-1)^2(25/3)$ *and*

$$\phi^* = \frac{(25/3)(10b-1)^2}{(25/3)(10b-1)^2 + 1}, \tag{4}$$

with $(25/3)(10b-1)^2$ *capturing the amount of unmodeled trend and* $\sigma_u^2 = 1$ *capturing the amount of noise in the data. Figure 3 illustrates the convergence of the posterior*

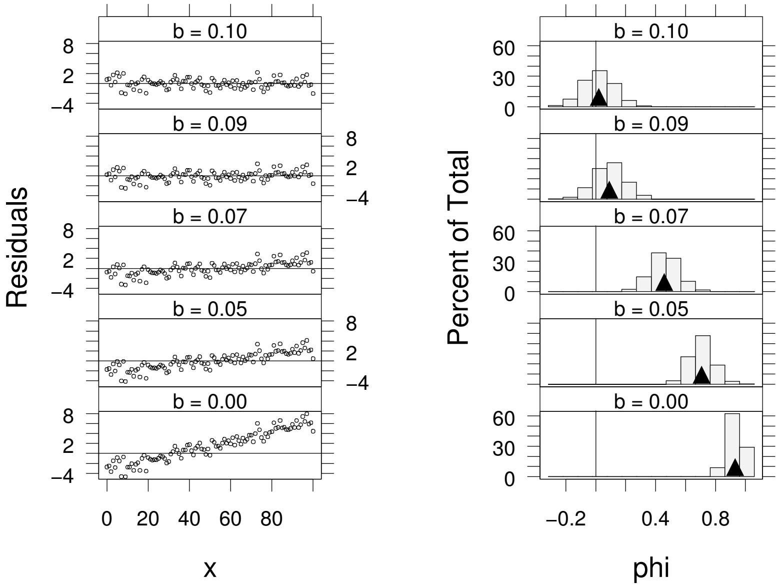


Figure 2: Residuals from the Five Fitted Lines and Corresponding Histograms of the AAR Parameters for the Linear Regression Model of Example 3.1. The triangular symbols represent the estimated posterior means of the AAR parameters.

distribution of ϕ to normality for the case of $b_4 = 0.05$, The three panels display, from bottom to top, histograms of posterior draws of ϕ based on samples of sizes 101, 1,001, and 10,001, respectively. As the sample size increases, the histograms become more symmetric and concentrate more closely around the asymptotic mean value of 0.676 calculated according to Equation (4). In each panel, the value of the asymptotic mean is represented by the vertical segment and the sample mean of the posterior draws is represented by a triangular symbol.

Example 3.2. *In this example, building on Examples 3.1 and on additional evidence (not reported to conserve space) that the methodology can successfully detect if the true conditional mean of the observations is a polynomial of a higher degree than the one specified by the model, I assess the performance of the AAR diagnostic in the context of a linear repeated measures model. I simulated repeated measurements y_{ij} according to the model $y_{ij} = a_i + b_i x_j + c_i x_j^2 + u_{i,j}$, where i, $1 \leq i \leq 10$, can be thought of as indexing an experimental subject and j, $1 \leq j \leq 20$, indexes a measurement taken at $x_j = j$. For the sake of discussion, I can think of the x_j as representing successive time points at which the measurements were taken. For $1 \leq i \leq 9$, I sampled independently a_i from a $N(3, (0.5)^2)$ distribution and b_i from a $N(0.2, (0.05)^2)$ distribution, and set $c_i = 0$. Furthermore, I set $a_{10} = 2.26787$, $b_{10} = 0.44321$, and $c_{10} = -0.01108$. Thus, the regression curves for the first nine subjects are linear with expected measurements of 3.2 at $x_1 = 1$ and 7.0 at $x_{20} = 20$. The regression curve for the last subject is quadratic,*

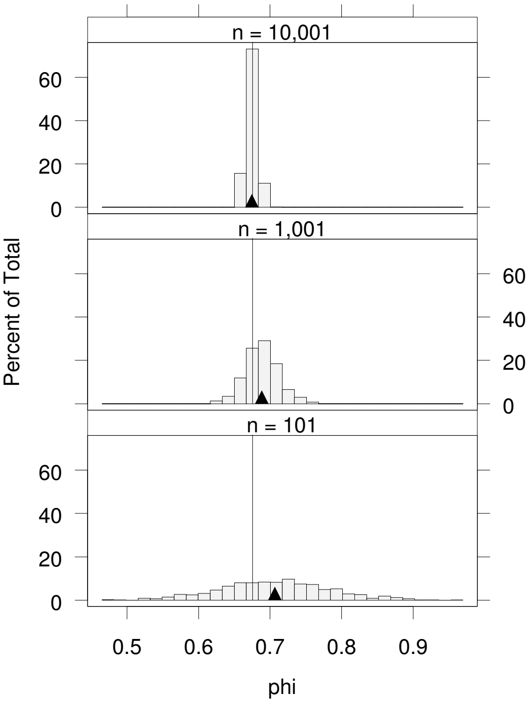


Figure 3: Convergence to Normality Under Regular Infill of the Posterior Distributions of the AAR Parameters in the Linear Regression Model of Example 3.1 for the Case $b_4 = 0.05$. The vertical segments represent the value of the asymptotic mean and the triangular symbols represent the estimated posterior means of the AAR parameters.

takes on values 2.7 at $x_1 = 1$ *and 6.7 at* $x_{20} = 20$, *and has a horizontal tangent at* $x_{20} = 20$. *I generated the measurement errors* u_{ij} *from a* $N(0, (0.2)^2)$ *distribution. The longitudinal regression curves and corresponding simulated data are depicted in the two panels of Figure 4, where the thicker lines represent the curves corresponding to the last subject.*

The correct assessment of the functional form of the regression curves for the various subjects is of paramount importance in the analysis of repeated measures data. Often, the regression curves are modeled as low degree polynomials. Naturally the question arises of whether a straight line is appropriate or a higher degree polynomial is needed to improve the fit. With this question in mind, I fit the model $y_{ij} = \alpha_i + \beta_i x_j + \eta_{i,j}$, *where the subject specific regression parameters are independently distributed as* $\alpha_i \sim N(\alpha_0, \sigma_\alpha^2)$ $\beta_i \sim N(\beta_0, \sigma_\beta^2)$, $1 \leq i \leq 10$, *with* $\alpha_0, \beta_0 \sim N(0, 10^3)$ *and* $1/\sigma_\alpha^2, 1/\sigma_\beta^2 \sim Gamma(0.1, 0.1)$, *all independently.*

For each subject, the error terms $\eta_{i,j}$ *follow the same AAR structure introduced in Example 3.1. Thus, there is an AAR parameter* ϕ_i *corresponding to each of the ten*

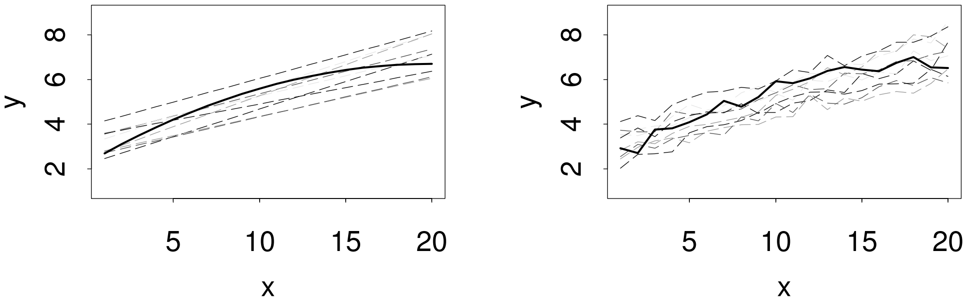


Figure 4: Simulated Longitudinal Regression Curves (left panel) and Data (right panel) for Example 3.2. The thicker lines correspond to the last subject who has a quadratic regression.

subjects. The posterior distributions of the ten AAR parameters are summarized by the histograms of Figure 5, based on 1,000 MCMC draws generated by BUGS with a burn-in of 10,000 and a thinning rate of 1 in 10. While the posterior distributions of the first nine AAR parameters do not appear to indicate any substantial lack-of-fit, the posterior mass of ϕ_{10} is concentrated well to the right of 0, with a posterior mean estimated at 0.75. The AAR diagnostic tool is thus very sensitive to the inadequacy in the specification of the regression curve for subject 10.

4 Quantifying the Lack-of-Fit

The examples and asymptotic results presented in the previous sections motivate the use of the posterior distribution of the AAR parameter to probe a regression model for lack-of-fit. In particular, the large sample derivation based on an in-fill argument provides very helpful interpretative insight. The mean of the limiting posterior distribution of the AAR parameter can be regarded as a signal-to-noise ratio for lack-of-fit constructed by dividing the squared bias in the numerator by the squared bias plus the variance of the true noise process in the denominator (cf. Equation (3)).

By incorporating the AAR diagnostic parameter into the Bayesian framework, the uncertainty about its importance can be assessed as a direct byproduct of the model fitting procedure. Such an assessment is *conditional on the data* and, in general, cannot be retrieved from a direct examination of the residuals. As an illustration, reconsider Example 3.1. For each of the 1,000 values of $\alpha_j^{(k)}$, $k = 1, \ldots, 1,000$, generated by the MCMC algorithm, a set of residuals could be computed according to $y_i - (\alpha_j^{(k)} + b_j x_i)$, $i = 1, \ldots, 101$. The autocorrelation of these residuals could be viewed as a proxy for the AAR parameter. However, any two sets of residuals corresponding to two different values of $\alpha_j^{(k)}$ only differ by a shift and, therefore, yield the same autocorrelation. For instance, for $b_j = b_4 = 0.05$ this common autocorrelation is equal to 0.692. This is, of course, a large value, but, by itself, it fails to convey the overall information about the

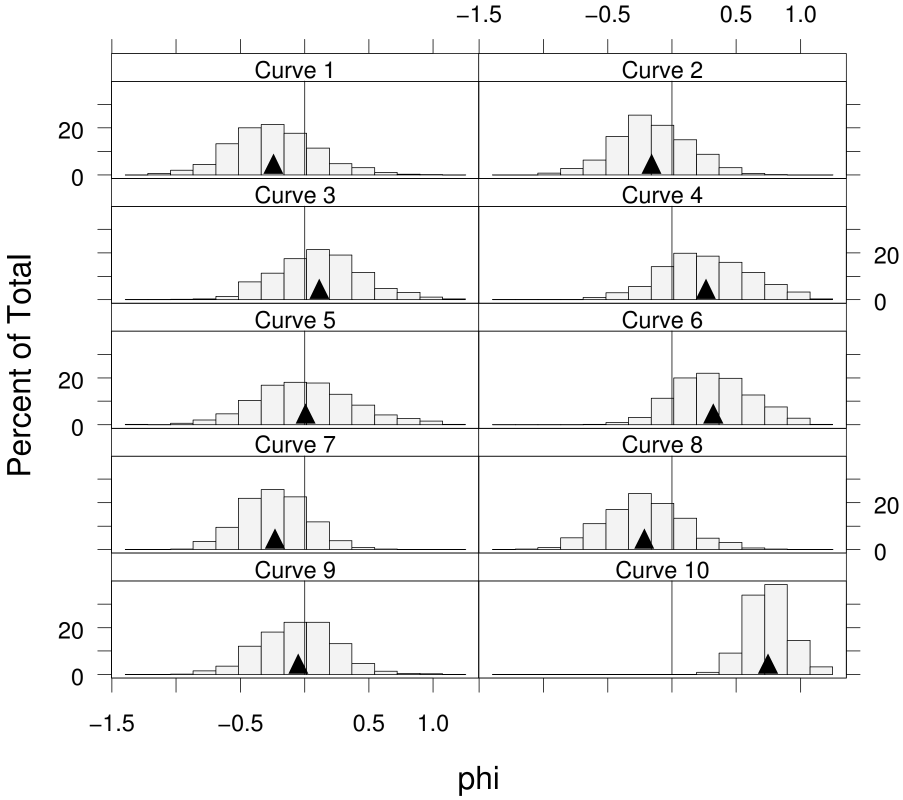


Figure 5: Histograms of the AAR Parameters for Example 3.2. The triangular symbols represent the estimated posterior means of the AAR parameters.

strength of evidence for lack-of-fit that is contained in the posterior distribution of the AAR parameter given the data (cf. Figure 2). For more general models, the residuals corresponding to different MCMC draws will not have constant autocorrelation, but the empirical distribution of the realized autocorrelations does not relate in any simple form to the distribution of the AAR diagnostic parameter. Thus, the posterior distribution of the AAR parameter conveys information about the lack-of-fit of the model beyond what can be uncovered by examining the residuals directly.

In light of the considerations above, I suggest that features of the posterior distribution of the AAR parameter, estimated directly from the MCMC fit, constitute the most direct and effective means of assessing lack-of-fit. Beside examining the histograms of the AAR parameters, one can compute posterior numerical summaries. One numerical summary of special interest is the posterior tail-probability that the AAR parameter be larger than a given threshold t, i.e.,

$$p_t = P(\phi \geq t \,|\, \text{data}). \tag{5}$$

	Slope b_j				
	0.10	0.09	0.07	0.05	0.00
Threshold t					
0.3	0.002	0.027	0.951	1.000	1.000
0.5	0.000	0.000	0.325	0.997	1.000
0.7	0.000	0.000	0.001	0.534	1.000

Table 1: Posterior Probabilities that $\phi \geq t$ Estimated as the Frequencies of the 1,000 MCMC Posterior Draws of ϕ that exceed t.

The value of the threshold can be adjusted by the modeler to reflect the degree of evidence required before the goodness-of-fit of a model is questioned. For example, based on the values reported in Table 1 concerning Example 3.1, setting a threshold of $t = 0.7$ and a requirement that p_t exceed 0.95, only the model corresponding to slope 0.00 would be questioned. A stricter threshold of $t = 0.5$ would lead one to question, in addition, the model with slope 0.05 and the even stricter threshold of $t = 0.3$ would lead one to include also the model with slope 0.07 in the set of questionable models.

5 Examples Using Real Data

Example 5.1. *This example demonstrates that the AAR method can also be employed when the independent variable x does not play the role of time and does not take on equally spaced values. For illustration, I consider the 100 complete pairs of brain and body weights for placental mammalian species published in Sacher and Staffeldt (1974). In addition to the weights, the data set also records the order and sub-order to which each species belongs. This data set was used in MacEachern and Peruggia (2002) to illustrate the performance of some numerical and graphical diagnostic tools in detecting the shortcomings of a simple linear regression (SLR) model compared to a variance component (VC) model. For mammal i, I denote by y_i the natural logarithm of the brain weight and by x_i the natural logarithm of the body weight recentered about the sample mean of the log body weights. A scatterplot of the data shown in Figure 6 suggests that a SLR model should provide an excellent fit to the data.*

The Bayesian specification of the SLR model with AAR errors says that, for $i = 1, \ldots, 100$, $y_i = \alpha + \beta x_i + \eta_i$, where, independently, $\alpha \sim N(0, 10^4)$ and $\beta \sim N(2/3, (2/9)^2)$ (see MacEachern and Peruggia 2002 for a detailed justification of these prior specifications). The prior distribution of the AAR error structure follows the same specifications given in Example 3.1.

The Bayesian VC model with AAR errors incorporates random effect terms γ_ℓ and δ_m for the 13 orders and 19 sub-orders in the taxonomy by saying that, for $i = 1, \ldots, 100$, $y_i = \boldsymbol{X}_i' \boldsymbol{\theta} + \eta_i$, where $\boldsymbol{\theta} = (\alpha, \beta, \gamma_1, \ldots, \gamma_{13}, \delta_1, \ldots, \delta_{19})'$ and $\boldsymbol{X}$ is the $100 \times (2 + 13 + 19 = 34)$ design matrix constructed by adjoining a column of ones, the column x of recentered log body weights, 13 columns of 0-1 order indicators, and 19 columns of 0-1 sub-order indicators. The priors for α and β are specified as in the SLR model and, independently,

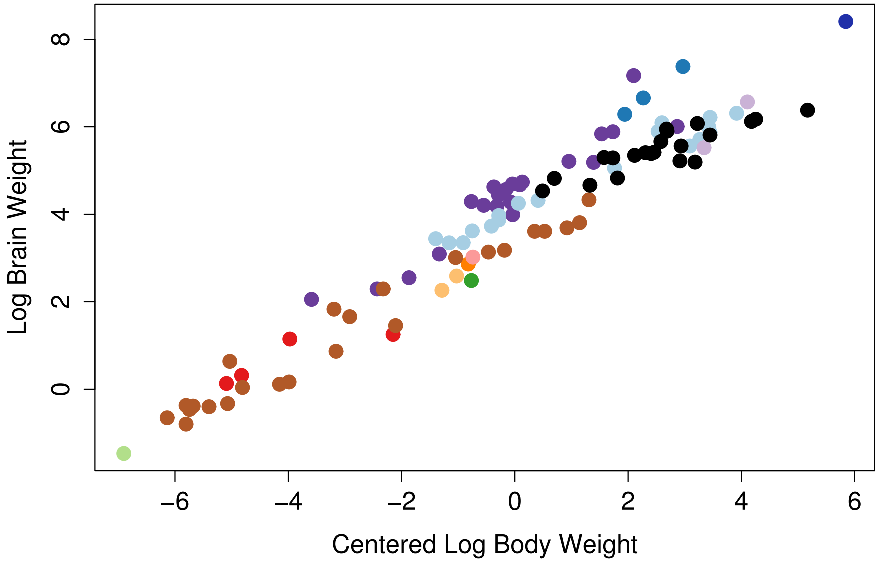


Figure 6: Statterplot of Log Brain Weight vs. Centered Log Body Weight for the 100 Mammalian Species of Example 5.1. The plotting symbols are color-coded according to the taxonomic order of the species.

$\gamma_\ell \sim N(0, \sigma_\gamma^2)$, $\ell = 1, \ldots, 13$, $\delta_m \sim N(0, \sigma_\delta^2)$, $m = 1, \ldots, 19$, $1/\sigma_\gamma^2 \sim Gamma(0.1, 0.1)$, *and* $1/\sigma_\delta^2 \sim Gamma(0.1, 0.1)$. *Once again, the prior distribution of the AAR error structure is as in Example 3.1.*

The histograms of Figure 7, based on 1,000 draws generated using BUGS with a burn-in of 10,000 and a thinning rate of 1 in 10, summarize the posterior distributions of the AAR parameter ϕ *for the SLR and VC models. For the SLR model, the posterior expectation of* ϕ *is estimated at 0.61 with a 95% equal-tailed posterior probability interval given by* $[0.45, 0.78]$. *For the VC model, the posterior expectation of* ϕ *is estimated at 0.28 with a 95% equal-tailed posterior probability interval given by* $[-0.07, 0.59]$.

There is a clear indication that the SLR model is not adequately capturing the variability in the data despite the strong linear trend evidenced by the scatterplot. The reasons for this are carefully outlined in MacEachern and Peruggia (2002). Essentially, the SLR model neglects to account for dependencies of species within orders and sub-orders. The VC model ameliorates the situation by inducing positive correlations within these groups. In this example, the AAR diagnostic device exploits the existence of a trade-off between the residual autocorrelation structure and the presence of random effects in the model (cf. Pinheiro and Bates 2000, p. 398). Specifically, when the correlations between groups are not taken into consideration by the SLR model, the residuals within each order and sub-order tend to behave similarly and the AAR error structure is sensitive to this grouping because the observation are arranged in the data file according to orders

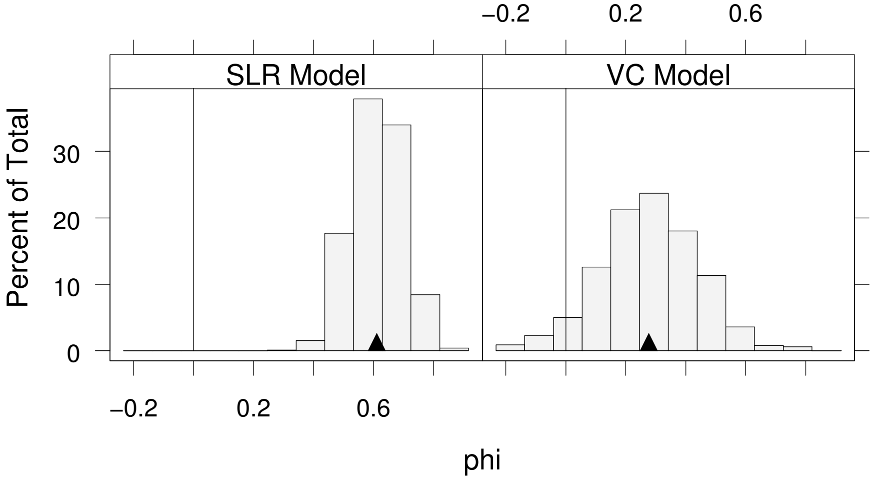
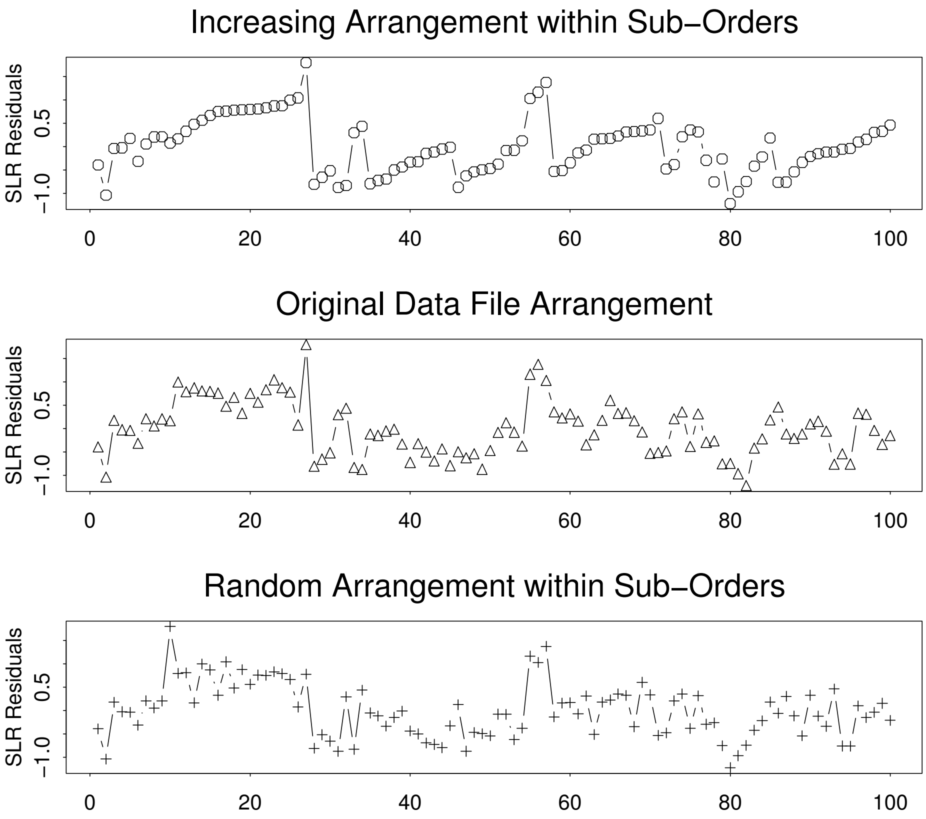


Figure 7: Histograms of the AAR Parameters for the Two Models for the Mammals Data Fit in Example 5.1. The triangular symbols represent the estimated posterior means of the AAR parameters.

and sub-orders.

One might wonder what role the arrangement of the species within each order and sub-order plays in determining the strength of the results. Figure 8 helps to address this question. All three panels display the residuals from the least squares fit of the SLR model. What differs is the order in which the residuals are arranged along the horizontal axis. In all three panels the species sub-orders share the same arrangement (Marsupialia, Insectivora, Chiroptera, Prosimii, Anthropoidea, Edentata, Lagomorpha, Sciuromorpha, Myomorpha, Hystricomorpha, Cetacea, Fissipeda, Pinnipedia, Proboscidea, Hyracoidea, Perissodactyla, Suiformes, Tylopoda, Ruminantia). In addition, the middle panel maintains the species arrangement of the original data file. In the top panel the species are rearranged in such a way that, within sub-order, the residuals come along in increasing order, so as to maximize autocorrelation. In the bottom panel, the arrangement of the species within each sub-order follows a random permutation. Visual inspection confirms that the autocorrelation is highest in the top panel. The autocorrelation also appears to be larger in the middle panel than in the bottom one. In fact, the values of the lag one sample autocorrelations for the arrangements in the three panel are 0.6574, 0.5863, and 0.4839 from top to bottom and one would expect the posterior estimates of the AAR parameter ϕ in the Bayesian SLR model to follow a similar pattern.

As it turns out, with the observations arranged as in the top panel, the posterior distribution of ϕ was estimated on the basis of an MCMC run of size 1,000 as having a mean of 0.68 with a 95% equal-tailed posterior probability interval given by $[0.51, 0.83]$*. The arrangement depicted in the bottom panel yielded a posterior mean for ϕ of 0.54 and a 95% equal-tailed posterior probability interval of* $[0.35, 0.72]$*, also estimated on*


Increasing Arrangement within Sub−Orders

Original Data File Arrangement

Random Arrangement within Sub−Orders

Figure 8: Three Different Orderings of the Residuals from the Least Squares Fit of the SLR Model to the Mammals Data of Example 5.1.

the basis of an MCMC run of size 1,000. These results confirm the intuition developed in the preceding paragraph that the posterior distribution of ϕ for the arrangement in the top panel is stochastically larger than the posterior corresponding to the original arrangement depicted in the middle panel (posterior mean of 0.61 and 95% posterior probability interval of $[0.45, 0.78]$*, as seen earlier), which in turn is stochastically larger than the posterior corresponding to the random arrangement depicted in the bottom panel.*

There are $1! \times 4! \times 1! \times 3! \times 18! \times 1! \times 2! \times 4! \times 11! \times 9! \times 3! \times 14! \times 4! \times 1! \times 1! \times 2! \times 4! \times 2! \times 15!$ *or approximately* 10^{60} *permutations of the species within sub-order that preserve a fixed arrangement of the sub-orders. In addition to the permutation depicted in the bottom panel of Figure 8, I generated another 99 random permutations and fit the Bayesian SLR model with AAR error structure. The 100 estimates of the posterior expectation of* ϕ *ranged from* 0.47 *to* 0.70 *with a mean value of* 0.62 *which is very close to the estimate of*

0.61 *obtained from the original data file arrangement. All 100 estimates are well above the value of 0.28 obtained from the fit of the VC model. This indicates that, regardless of the arrangement of the various species within sub-orders, the AAR diagnostic device is capable of signaling the inadequacies of the SLR model.*

I also considered arrangements that would only keep the taxonomy orders fixed, allowing for permutation of the species to run across sub-orders. Such arrangements typically yield lower estimated values of the posterior mean of the AAR parameter. For 100 randomly generated arrangements of this type the smallest estimate of the posterior expectation of ϕ *was 0.36 and the largest was 0.61 with a mean of 0.51. Here, all 100 estimated posterior expectations were smaller than the estimate of* 0.61 *obtained from the original data file arrangement but, again, they were all larger than the estimate of 0.28 obtained from the VC model. Thus, grouping by taxonomic order alone would still induce dependencies that are strong enough to enable the AAR diagnostic device to detect the limitations of the SLR model compared to the VC model.*

Example 5.2. *In this example I look at a data set of Williams (1959). This data set was analyzed by Carlin and Chib (1995) and Spiegelhalter, Thomas, Best, and Gilks (1996b) to illustrate the use of Bayes factors to compare non-nested regression models. The response variable* y *is the maximum compressive strength parallel to the grain measured on 42 specimens of radiata pine. There are two possible predictors:* x*, denoting the density of the specimens, and* z*, denoting the density adjusted for resin content. Corresponding to the two predictors, there are two competing SLR models with AAR errors that are specified, for* $i = 1 \ldots 42$*, as*

$$\textit{Model A: } y_i = \alpha + \beta x_i + \eta_i \quad \text{and} \quad \textit{Model B: } y_i = \gamma + \delta z_i + \eta_i.$$

Carlin and Chib (1995) and Spiegelhalter et al. (1996b) use different transformations of the variables in their analyses with independent errors. Carlin and Chib (1995) recenter the x_i *and the* z_i *by subtracting off their means while Spiegelhalter et al. (1996b) standardize them by subtracting off their means and dividing by their standard deviations and they also similarly standardize the response values* y_i*. Slightly different priors are also used. Yet, both articles report Bayes factors that demonstrate unequivocally how Model B provides a better fit than Model A.*

I will now show how the AAR error diagnostic device is able to detect the superiority of Model B over Model A. In my analyses I used the same standardized variables x*,* y *and* z *and the same prior specification as in Spiegelhalter et al. (1996b). In particular, I set* $\alpha \sim N(0, 10^{-6})$*,* $\beta \sim N(0, 10^{-4})$*,* $\gamma \sim N(0, 10^{-6})$*, and* $\delta \sim N(0, 10^{-4})$*, independently within models, with the AAR error structure following the basic specification of Example 3.1, except that the shape and scale parameters of the gamma distribution for* $1/\sigma_\epsilon^2$ *were set equal to* 10^{-4}*.*

In this example there is no natural ordering of the observations, not even a partial ordering, such as that induced by the taxonomy of the mammals data set of Example 5.1. If one fits Models A or B using the original ordering in the data file, the posterior distribution of the AAR parameter ϕ *turns out to put considerable mass around zero and no evidence of lack-of-fit is uncovered in either case. The posterior distributions of*

	Model A		Model B	
Ordering	Expectation	95% Prob. Int.	Expectation	95% Prob. Int.
Original	−0.02	(−0.35, 0.32)	−0.15	(−0.49, 0.20)
Based on $x - z$	0.35	(0.00, 0.72)	−0.01	(−0.35, 0.32)
Based on Residuals	0.26	(−0.10, 0.63)	−0.13	(−0.46, 0.19)

Table 2: Posterior Estimates of ϕ For the Pines Data Example.

ϕ corresponding to the two fits are summarized in the first line of Table 2.

Here, however, the main concern is a comparison of the fit provided by the two models and the ordering of the observations should relate to this comparison. The models differ in the predictor they use, so the relative quality of the fit will be determined by differences in the two predictors. This suggest simply ordering the data according to the increasing values of the differences $x_i - z_i$ (alternatively, an ordering based on relative differences, normalized for the size of the predictors, could also be used). The posterior distributions of ϕ corresponding to the two fits of the ordered data are summarized in the second line of Table 2. While the posterior distribution of ϕ in Model B continues to put considerable mass around zero, the posterior distribution of ϕ in Model A is shifted to the right of zero, thus providing evidence of lack-of-fit. Model A fits poorly systematically at those points where the differences between the x and y predictors are large, a feature that AAR diagnostic can readily detect.

A similar behavior of the posterior distributions of ϕ can be observed when the observations are ordered according to the average differences of the residuals for the models with independent errors (or, equivalently, according to the average differences of the fitted values). As a convenient shortcut, because of the noninformative nature of the prior distributions, the observations can be ordered according to the differences of the residuals from the least squares fits. This approach yields the posterior distributions for ϕ summarized in the third line of Table 2. These summaries still indicate that Model B provides a superior fit, albeit not as strongly as the summaries in the second line of the table. The ordering based on the differences of the residuals is particularly useful in situations when the models being compared contain several, possibly different, predictors.

Example 5.3. *In this example I examine the ability of the AAR diagnostic device to uncover aspects of model inadequacy in the context of a complex hierarchical model for a large set of response time data. The data were collected in a series of recognition memory trials conducted on four subjects over ten non-consecutive days.*

For each trial, a subject was initially asked to study a list of 40 words randomly selected from a database of 2337 common English words. The subject was then presented with a sequence of 40 words, 20 selected from the study list and 20 selected from words in the database not included in the study list. The words were displayed sequentially on a computer monitor and the subject was asked to strike one of two keys on the keyboard depending on whether she thought the word was an "old" word contained in the study *list or a "new" word not contained in the study list. The times in milliseconds elapsed between the appearances on the words on the screen and the keystrokes were recorded, along with an indicator of the accuracy of the responses. Each subject participated in two consecutive trials on each day. There were thus a total of $4 \times 2 \times 10 = 80$ trials, each contributing 40 response times.*

The hierarchical model considered here is a refinement of the model presented in Peruggia, Van Zandt, and Chen (2002). For the current analysis I used a two parameter Weibull likelihood to model the shifted response times obtained by subtracting a value equal to 95% of the minimum response time for each list from all of the 40 response times for that list. (A three parameter Weibull likelihood could be used to model the unshifted response times.) Let $\mathtt{RT}_{i,d,l,w}$ denote the shifted response times, where $i, 1 \leq i \leq 4$, indexes subject, $d, 1 \leq d \leq 10$, indexes day, $l, 1 \leq l \leq 2$, indexes list, and $w, 1 \leq w \leq 40$, indexes word. Conditional on $r_{i,d,l}$ and $\lambda_{i,d,l,w}$, the $\mathtt{RT}_{i,d,l,w}$ are assumed to follow independently a Weibull distribution with shape parameter $r_{i,d,l}$ and scale parameter $\lambda_{i,d,l,w}$, for $1 \leq w \leq 40$.

The logarithm of the scale parameters is endowed with a regression structure that includes random effects $\alpha_{i,d,l}$ primarily intended to model different levels of subject specific learning as days go by. The regression also includes fixed effects for the nature of the words ("old" vs. "new"), the accuracy of the responses ("right" vs. "wrong"), and their interaction. In summary, I assumed that $\ln(\lambda_{i,d,l,w}) = \alpha_{i,d,l} + \beta_1\, \mathtt{I.Old}_{i,d,l,w} + \beta_2\, \mathtt{I.Right}_{i,d,l,w} + \beta_3\, (\mathtt{I.Old}_{i,d,l,w} \times \mathtt{I.Right}_{i,d,l,w}) + \eta_{i,d,l,w}$, with $\mathtt{I.Old}$ denoting a 0-1 indicator of an old word and $\mathtt{I.Right}$ denoting a 0-1 indicator of a correct response. For the random effects I assumed $\alpha_{i,d,l} \sim N(\mu_d, \sigma^2_\alpha)$, independently, where, also independently, $\mu_d \sim \mathrm{N}(0, 10^{-4})$ and $1/\sigma^2_\alpha \sim \mathrm{Gamma}(10^{-1}, 10^{-1})$. For the fixed effects coefficients I assumed $\beta_1, \beta_2, \beta_3 \sim \mathrm{N}(0, 10^{-3})$, independently. For the shape parameters I assumed that, independently, $r_{i,d,l} \sim Exp(\theta_d)$ where, also independently, $\theta_d \sim Exp(\theta_0)$ and $\theta_0 \sim Exp(10^{-3})$.

The error terms $\eta_{i,d,l,w}$ act as diagnostic devices for the stated model of conditional independence of the 40 response times within each list, following the basic AAR structure used throughout the article with AAR coefficients $\phi_{i,d,l}$ that depend on subject, day, and list. In this example, I specified a more diffuse prior for the variance of the first innovation than in the previous examples by assuming $1/\sigma^2_1 \sim Gamma(0.1, 0.1)$ rather than $1/\sigma^2_1 \sim Gamma(2.5, 10.0)$. I made this choice because, with 80 different lists, the data can contribute much information about the distribution of the first innovation. The AAR structure is specified with respect to a given ordering of the response times within a list. In the diagnostic analysis that follows I will consider two separate orderings, the one corresponding to the sequence in which the words were presented to the subjects and the one corresponding to a rearrangement in which the "new" words are made to precede the "old" words, while preserving the original ordering within each group.

For a basic diagnostic analysis, I quickly scanned the histograms of the 80 sets of 1,000 realizations from the marginal posterior distributions of the AAR parameters $\phi_{i,d,l}$ generated using WinBUGS with a burn-in period of 20,000 and a subsequent thinning rate of 1 in 500. Most distributions put considerable mass around zero but there are

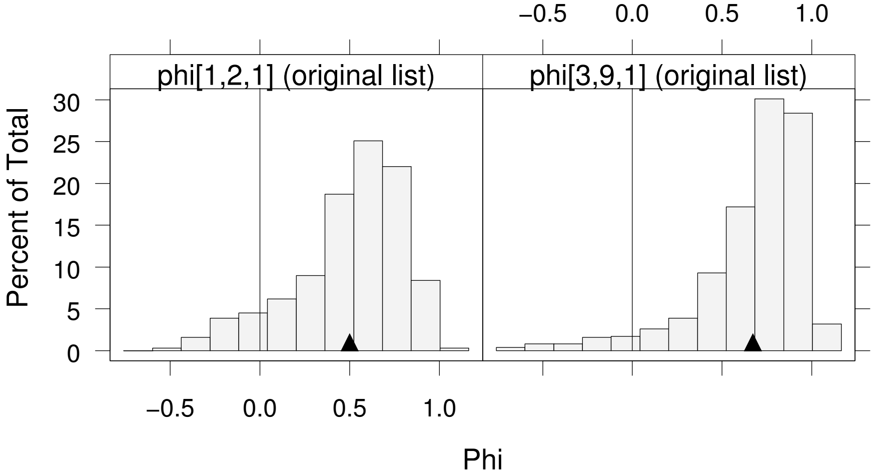


Figure 9: Histograms of the AAR Parameters for Two Word Lists in Example 5.3. The triangular symbols represent the estimated posterior means of the AAR parameters.

several interesting cases in which the shapes of the posterior distributions of the AAR parameters suggest inadequacies in the fit for certain word lists. For the purpose of illustration I will now look at some of these cases.

The histograms of Figure 9 represent estimates of the posterior densities of $\phi_{1,2,1}$ *and* $\phi_{3,9,1}$. *The posterior distribution of* $\phi_{3,9,1}$ *puts 0.97 of its mass to the right of zero and has a mean of 0.67. The posterior distribution of* $\phi_{1,2,1}$ *shows similar qualitative features (if not quantitatively as pronounced), putting 0.91 of its mass to the right of zero and having a mean of 0.50. To understand what causes the mass of the posterior distributions of the two autoregressive parameters to be so heavily shifted to the right of zero, I computed residuals from the Bayesian fit as follows. With the superscript* (k), $1 \leq (k) \leq 1,000$, *indexing the k-th MCMC parameter draw, let* $\lambda^{(k)}_{i,d,l,w} = \exp\left[\alpha^{(k)}_{i,d,l} + \beta^{(k)}_1 \, \texttt{I.Old}_{i,d,l,w} + \beta^{(k)}_2 \, \texttt{I.Right}_{i,d,l,w} + \beta^{(k)}_3 \left(\texttt{I.Old}_{i,d,l,w} \times \texttt{I.Right}_{i,d,l,w}\right)\right]$ *and let* $E\left(W^{(k)}_{i,d,l,w}\right)$ *denote the mean of a Weibull random variable with shape parameter* $r^{(k)}_{i,d,l}$ *and scale parameter* $\lambda^{(k)}_{i,d,l,w}$. *The residual for the response time* $\texttt{RT}_{i,d,l,w}$ *is then computed as* $\texttt{res}_{i,d,l,w} = \texttt{RT}_{i,d,l,w} - (1,000)^{-1} \sum_{k=1}^{1,000} E\left(W^{(k)}_{i,d,l,w}\right)$. *Note that these are observation-level residuals which, although related, do not correspond directly to the AAR error terms appearing at a higher level of the hierarchy.*

The two panels of Figure 10 display the residuals for the response times of subject 1, day 2, list 1 and subject 3, day 9, list 1. After adjusting for the values of the covariates, the earlier response times for subject 3, day 9, list 1 appear to be slower than the model would predict (the residuals are mainly positive) and the later response times appear to be faster (the residuals are mainly negative). This systematic departure from a random

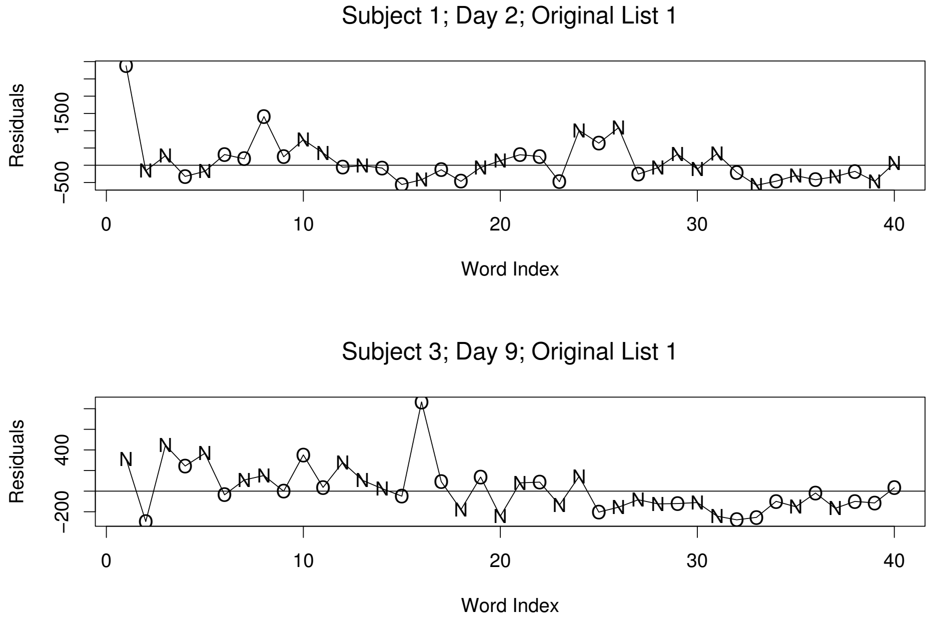


Figure 10: Residuals for the Response Times based on the Two Word Lists Considered in Figure 9. A symbol 'N' denotes a "new" word and a symbol 'O' denotes an "old" word.

pattern in the residuals is accurately detected by the AAR parameter diagnostic. The residuals for subject 1, day 2, list 1 also appear to exhibit a systematic pattern, corresponding to four or five alternating stretches of longer and shorter response times. The pattern in the set of residuals for subject 3 appears to be stronger than the pattern in the set of residual for subject 1 and this difference is reflected in the fact that the posterior distribution of $\phi_{3,9,1}$ *is shifted further to the right than the posterior distribution of* $\phi_{1,2,1}$.

Another interesting case is the one illustrated in Figures 11 and 12, dealing with subject 3, list 6, day 1. The left panel of Figures 11 summarizes the posterior distribution of $\phi_{3,6,1}$ *when the RTs are ordered according to the sequence in which the words were originally presented. The histogram is essentially symmetric about zero and provides no indication of lack-of-fit. Contrast this to the right panel which summarizes the posterior distribution of* $\phi_{3,6,1}$ *when the RTs are rearranged so that those corresponding to the "new" words are made to precede those corresponding to the "old" words, while preserving the original ordering within each group. Now the posterior mass is heavily shifted to the right of zero. Inspection of the residual plots of Figure 12 provides an explanation for this difference. The top panel, corresponding to the original ordering, does not present any systematic pattern, but the bottom panel suggests that responses to the new words were systematically faster than the model would predict and responses to*

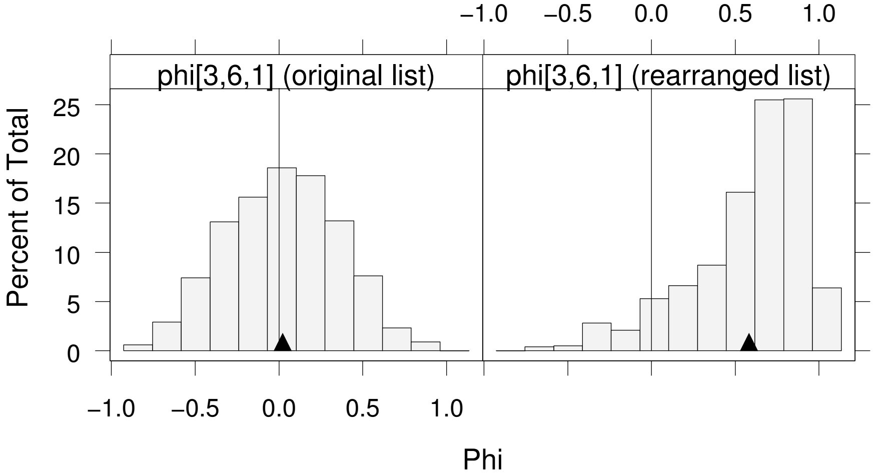
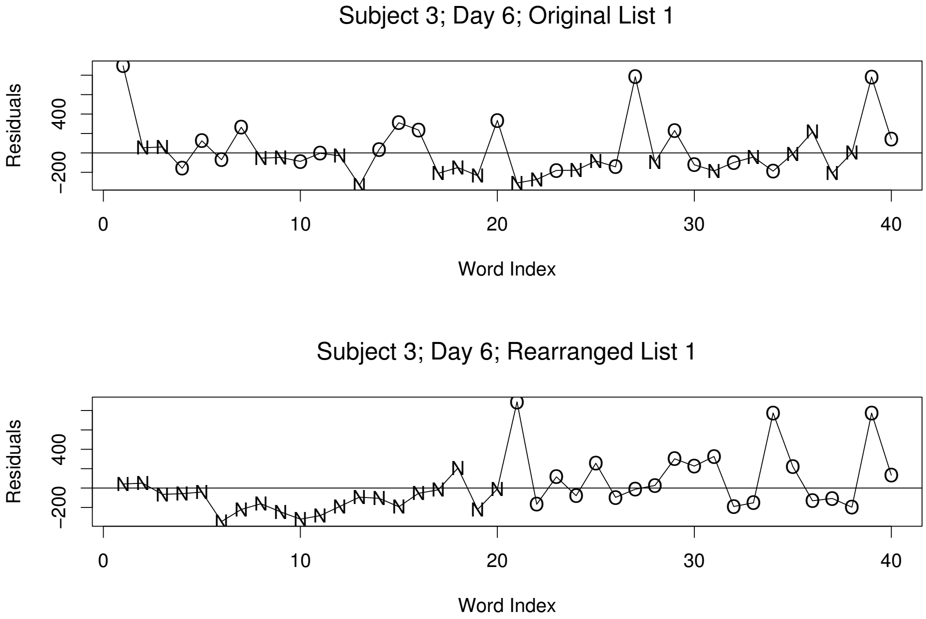


Figure 11: Histograms of the AAR Parameters for Two Orderings of a Word List in Example 5.3. The triangular symbols represent the estimated posterior means of the AAR parameters.

the old words were systematically slower.

Diagnostics based on the AAR parameter can be carried out to investigate jointly various aspects of the model fit. As an example, consider the Bland-Altman plot of Figure 13 (Bland and Altman 1986). There, using a threshold value of $t = 0.5$ *and for each subject by day combination, I plot the difference between the tail probabilities* p_t *defined in Equation (5) for the AAR parameters of the first and of the second list against the average tail probability for the AAR parameters of the two lists. Different plotting symbols are used to identify the four subjects. Several interesting features become apparent. First, the level of variation in the differences appears to increase with the value of the average. Second, the seven largest values of the absolute differences all correspond to positive differences (the seven points in the upper right corner of the plot). This is consistent with the possible presence of unmodeled disturbances in the response times for the first list during a "settling in" phase. Such disturbances appear to weaken as the subjects become more comfortable with the task to be performed.*

In this example the AAR error devices provide a simple and reliable means of scanning the 80 word lists to examine the fit of the model. The histograms of the posterior MCMC draws of the AAR parameters and the tail probabilities defined in Equation (5) are much quicker to derive and easier to interpret than a direct examination of the observation-level residuals from the model fit, and, as noted in Section 4, carry additional information about the strength of the lack-of-fit. Once lack-of-fit is suspected for a certain list a more thorough analysis, including a direct examination of the residuals, can pinpoint more specifically the nature of the inadequacies in the model. In addition to those described above, there are a variety of additional model deficiencies and unchar-


Subject 3; Day 6; Original List 1

Subject 3; Day 6; Rearranged List 1

Figure 12: Residuals for the Response Times based on the Two Orderings of the Word List Considered in Figure 11. A symbol 'N' denotes a "new" word and a symbol 'O' denotes an "old" word.

acteristic observations that the diagnostic device is able to uncover. For example, there are a few cases in which the distribution of the AAR parameter is shifted to the left of zero, corresponding to a systematic pattern of the residuals that alternate rhythmically between positive and negative values.

The complexity of these data makes modeling a challenge. Overall the specified model is adequate, but there are a few lists for which the fit is not entirely satisfactory. Even though the individual experimental tasks are relatively short (there are only 40 words in each list) and long term trends and dependencies are not the norm, in some cases the assumption of conditional independence is seemingly violated. For example, it appears as if subject 3, in her reaction to the words in list 1 on day 9, is trying to compensate for a slow start in the first half of the list by speeding up her responses to the words in the second half of the list. This and other types of uncommon features cannot be captured by the basic hierarchical model, but the introduction of the AAR error structure acts as an effective screening tool for uncovering them.

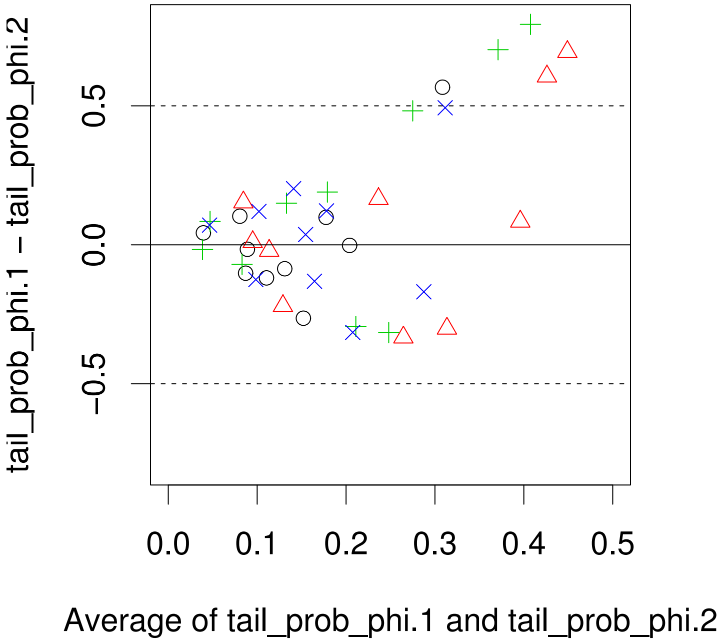


Figure 13: Bland-Altman Plot of the Tail Probabilities of the AAR Parameters for List 1 vs. List 2 in the RT Example. Four different plotting symbols are used to distinguish between points corresponding to the four subjects.

6 Discussion

In this article I have proposed a general procedure based on AAR errors for detecting lack-of-fit in hierarchical Bayes model. The principal appeal of the procedure lies in its flexibility and its ease of implementation. All examples presented in this article were in fact fit using publicly available software without any need to write customized programs.

The most popular Bayesian tools for performing model comparison and assessing goodness-of-fit—such as, for example, Bayes factors (Kass and Raftery 1995) and their generalizations (Berger and Pericchi 1996; O'Hagan 1995), posterior predictive model checks (Gelman, Meng, and Stern 1996), and the nice, general-purpose method based on a Bayesian version of the χ^2 test for goodness-of-fit proposed in Johnson (2004)—do not examine the residuals directly. The behavior of the residuals, however, can shed much light on the fit of a model and suggest directions for refinement if lack-of-fit is detected. The strength of the proposed method is that the AAR device can be applied to assess the behavior of residuals at different levels of the hierarchy, probing different aspects of the fit, without much conceptual or practical difficulty.

Unlike other diagnostics for model fit based on embedding (such as the decision theoretical model elaboration strategy of Carota, Parmigiani, and Polson 1996 and its application to the detection of autocorrelation of the disturbances in regression models described in Carota 1998, 2005), there is no expectation that the model providing the embedding (the model with AAR errors in this case) be adequate. The sole purpose of the AAR error device is to detect if, and in what ways, the embedded model with independent errors is inadequate. In this respect, even though more elaborate models of dependence could be considered, the first-order autoregressive structure is all that is needed because it is sensitive to misspecifications of conditional means and is easy to understand. In my experience, initialization of the first error term η_1, though arbitrary, does not have a great impact on the inferential conclusions, as long as the prior distribution of the variance σ_1^2 is not too diffuse.

The potential benefits of the method when applied to complex hierarchical models is clearly illustrated in Example 5.3. There, I introduced, at a higher level of the hierarchy, a separate AAR error device for each of 80 sets of multidimensional observations (the 80 sets of RTs for each of the 80 word lists). The posterior distributions of each AAR parameter is a low dimensional summary that can quickly detect if the model of conditional independence provides an inadequate fit to the RTs for the corresponding list. Scanning such summaries to assess and quantify lack-of-fit (cf. Figures 9 and 11) is much easier than scanning observation-level residual plots (cf. Figures 10 and 12).

The AAR device is based on orderings of the observations. To detect lack-of-fit, as illustrated in the various examples, it is important not to find an ordering for which the AAR parameter is significant (after all there will always be such an ordering) but to find a *meaningful* ordering for which this is true. While a routine implementation of the method should always consider orderings based on the size of predicted values and of covariate values, other orderings might be suggested by application-specific considerations.

References

Berger, J. O. and Pericchi, L. R. (1996). "The Intrinsic Bayes Factor for Model Selection and Prediction." *Journal of the American Statistical Association*, 91: 109–122. 839

Bland, J. M. and Altman, D. G. (1986). "Statistical Method for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, i: 307–310. 837

Carlin, B. P. and Chib, S. (1995). "Bayesian Model Choice Via Markov Chain Monte Carlo Methods." *Journal of the Royal Statistical Society, Series B, Methodological*, 57: 473–484. 832

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis (2nd ed.)*. Boca Raton: Chapman & Hall/CRC. 818

Carota, C. (1998). "A Diagnostic for Autocorrelation of the Disturbances in Regression Models." *Journal of the Italian Statistical Society*, 3: 257–266. 818, 840

— (2005). "Symmetric Diagnostics for the Analysis of the Residuals in Regression Models." *Biometrika*, 92: 787–799. 818, 840

Carota, C., Parmigiani, G., and Polson, N. G. (1996). "Diagnostic Measures for Model Criticism." *Journal of the American Statistical Association*, 91: 753–762. 839

Durbin, J. and Watson, G. S. (1950). "Testing for Serial Correlation in Least Squares Regression I." *Biometrika*, 37: 409–428. 818

— (1951). "Testing for Serial Correlation in Least Squares Regression II." *Biometrika*, 38: 159–178. 818

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis, Second Edition*. Boca Raton, FL: Chapman & Hall/CRC Press. 818, 821

Gelman, A., Meng, X.-L., and Stern, H. (1996). "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies (Disc: P760-807)." *Statistica Sinica*, 6: 733–760. 839

Johnson, V. E. (2004). "A Bayesian χ^2 Test for Goodness-of-Fit." *Annals of Statistics*, 32: 2361–2384. 839

Judge, G. G., Griffiths, W. E., Hill, R. C., and Lee, T.-C. (1980). *The Theory and Practice of Econometrics*. John Wiley & Sons. 818

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90: 773–795. 839

MacEachern, S. N. and Peruggia, M. (2002). "Bayesian Tools for EDA and Model Building: A Brainy Study." In Gatsonis, C., Kass, R. E., Carlin, B., Carriquiry, A., Gelman, A., Verdinelli, I., and West, M. (eds.), *Case Studies in Bayesian Statistics, Vol. 5*, 345–362. New York: Springer-Verlag. 828, 829

O'Hagan, A. (1995). "Fractional Bayes Factors for Model Comparison (Disc: P118-138)." *Journal of the Royal Statistical Society, Series B: Methodological*, 57: 99–118. 839

Peruggia, M., Van Zandt, T., and Chen, M. (2002). "Was it a Car or a Cat I Saw? An Analysis of Response Times for Word Recognition." In Gatsonis, C., Kass, R., Carriquiry, A., Gelman, A., Higdon, D., Pauler, D., and Verdinelli, I. (eds.), *Case Studies in Bayesian Statistics, Vol. 6*, 319–334. New York: Springer-Verlag. 834

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects Models in S and S-PLUS*. New York: Springer-Verlag. 829

Sacher, G. A. and Staffeldt, E. F. (1974). "Relation of Gestation Time to Body Weight for Placental Mammals: Implications for the Theory of Vertebrate Growth." *The American Naturalist*, 108: 593–615. 828

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1996a). *Bayesian Inference Using Gibbs Sampling, Version 0.5, (version ii)*. Cambridge, UK: MRC Biostatistics Unit. 818

— (1996b). *BUGS Examples Volume 2, Version 0.5, (version ii)*. Cambridge, UK: MRC Biostatistics Unit. 832

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2003). *WinBUGS User Manual, Version 1.4*. Cambridge, UK: MRC Biostatistics Unit. 818

Weisberg, S. (1985). *Applied Linear Regression (2nd ed.)*. New York: John Wiley & Sons. 817

Williams, E. (1959). *Regression Analysis*. New York: Wiley. 832

Zellner, A. and Tiao, G. C. (1964). "Bayesian Analysis of the Regression Model with Autocorrelated Errors." *Journal of the American Statistical Association*, 59: 763–778. 818

Acknowledgments

This material is based upon work supported by the National Science Foundation under Awards No. SES-0214574, SES-0437251 and DMS-0605052. The author thanks Peter Craigmile, Ilenia Epifani, Steve MacEachern, Trisha Van Zandt, an Associate Editor and an anonymous referee for their helpful comments.