

Analysis of hospital quality monitors using hierarchical time series models

Omar Aguilar and Mike West
ISDS, Duke University

ABSTRACT

The VA management services department invests considerably in the collection and assessment of data to inform on hospital and care-area specific levels of quality of care. Resulting time series of *quality monitors* provide information relevant to evaluating patterns of variability in hospital-specific quality of care over time and across care areas, and to compare and assess differences across hospitals. In collaboration with the VA management services group we have developed various models for evaluating such patterns of dependencies and combining data across the VA hospital system. This paper provides a brief overview of resulting models, some summary examples on three monitor time series, and discussion of data, modelling and inference issues. This work introduces new models for multivariate non-Gaussian time series. The framework combines cross-sectional, hierarchical models of the population of hospitals with time series structure to allow and measure time-variations in the associated hierarchical model parameters. In the VA study, the within-year components of the models describe patterns of heterogeneity across the population of hospitals and relationships among several such monitors, while the time series components describe patterns of variability through time in hospital-specific effects and their relationships across quality monitors. Additional model components isolate unpredictable aspects of variability in quality monitor outcomes, by hospital and care areas. We discuss model assessment, residual analysis and MCMC algorithms developed to fit these models, which will be of interest in related applications in other socio-economic areas.

1 Introduction

The performance monitoring system of the US Department of Veterans Affairs (VA) collects, reports and analyses data from over 170 hospitals. Policy interests lie in accurately estimating measures of hospital-level performance in key areas of health care provision, and in assessing changes over time in such measures to monitor impact of internal policy changes. Ultimately, these issues are related to the development of management and economic incentives designed to encourage and promote care provision at sustained and acceptable levels. As described in Burgess et al (1996), the quality monitor data are compiled annually and encompass a range of inpatient,

outpatient and long term care activities at each of the VA medical centers. Each hospital records data on the total numbers of individuals who were exposed to a specific and well-defined outcome in each monitor area, and the number for whom that outcome occurred. There is a related covariate, referred to as the *DRG predictor*, based on exogenous information providing some correction for hospital/monitor specific case-mix and characteristics of patient population profiles. Further details appear in Burgess, Christiansen, Michalak and Morris (1996 and in related unpublished work), who discuss aspects of data analysis and hierarchical modelling (Christiansen and Morris 1997) in this context. Our study is concerned with evaluating

- patterns of variability over time, in hospital-monitor and area-specific performance measures across a selection of quality monitors, and
- patterns of dependencies between sets of monitors, in addition to and in combination with assessment of time-variations.

Christiansen and Morris have developed a variety of Bayesian hierarchical models for the observed outcomes, including regressions on the DRG predictor and hospital-specific parameters drawn from a hospital population prior (see references above). From this basis, we explore multiple-monitor time series models to address the above key questions. We focus on three specific monitors introduced in Section 2 where we provide some basic data description and perspective. Section 3 reviews our new models; these are multiple monitor, binomial/logit models in which hospital-specific random effects are related through time via a multivariate time series model. In addition to systematic patterns of variation over time, the models include components of unpredictable variability in outcome probabilities. In Section 4 we describe summary inferences for all hospitals and monitors, investigation of aspects of model fit, and examples of additional possible uses of the models. We conclude with summary comments about the study, and an appendix briefly summaries model theory and computation.

Our work relates closely to what are now essentially standard approaches in health care outcomes research and institutional comparisons – hierarchical Bayesian models that allow for various components of heterogeneity involving nested random effects. A recent contribution and overview appears in Normand et al (1997), for example. Our work is novel in several methodological respects, and draws on developments in Cargnoni, Müller and West (1997) related to both latent time series structure and computational algorithms. The methods will prove useful to workers dealing with longitudinal data structures in various socio-economic fields. Finally, more extensive details on the data analysis and modelling summarised here appears in an on-line report by West and Aguilar (1997).

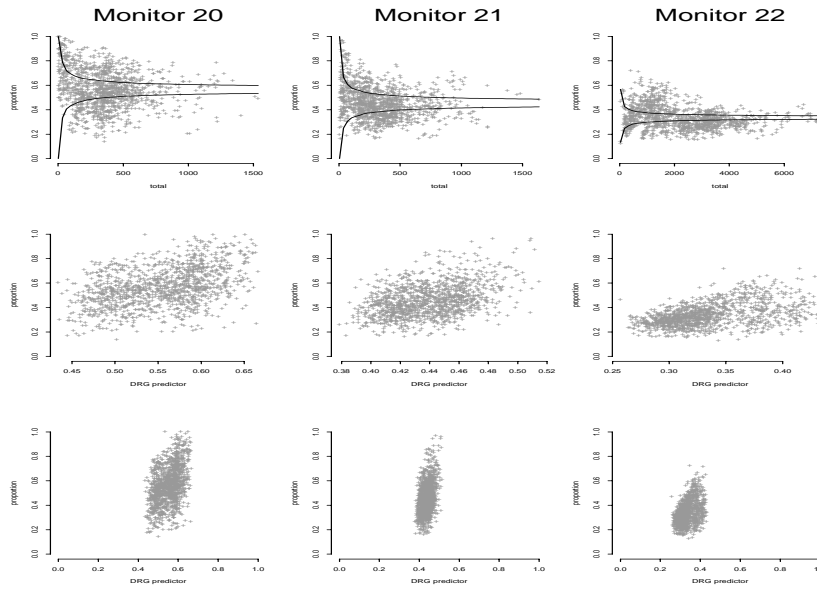


FIGURE 1. Observed and DRG-based predicted proportions on all three monitors and eight years of data (1988-95). The lower row is the DRG predictor on the probability scale.

2 Exploratory Data Analysis

The outcomes in the monitor areas represent annual numbers of individuals under a binary classification in an area of basic medical or psychiatric health care. The response recorded is the number of individuals who failed to return for an outpatient visit within 30 days of discharge out of the total number of annual discharges. Monitor M20 measures outcomes for *General Psychiatric*, M21 for *Substance Abuse Psychiatric*, and M22 for *Basic Medical and Surgical* care. Low return rates are indicative of low “quality” in these specific care areas. The data here covers years 1988 to 1995 for 152 hospitals having complete records.

Figure 1 displays the raw data on the three monitors separately, but combined over all eight years. There are $8 \times 152 = 1216$ observations per frame for the 8 years of data on $I = 152$ hospitals. The graphs plot the observed proportions of successes in each monitor against the total numbers of patients in each case, and then against the DRG-based predicted proportions. Super-imposed on each graph in the first row are approximate 99% intervals under marginal binomial distributions that assume “success” probabilities fixed at the overall average proportions for each monitor. Many observations lie outside these bands indicating considerable levels of over-

dispersion relative to binomial models. This extra-binomial variation is to be explained by models that describe how the individual probabilities vary across hospitals and across years, using a combination of regression on the DRG predictor and random effects.

There is an overall suggestion of decreasing levels of observed responses across the eight years (not displayed here; see Figure 4 of West and Aguilar 1997). This is most marked in M20 and, to a lesser extent M21. The average DRG values do not show decreasing patterns indicating that this is very likely a hospital system-wide feature, perhaps due to VA policy and/or general improvements in care provision over the years. Thus, we need to consider models where the overall levels of outcome responses across all hospitals and the DRG variable vary year to year.

3 Multivariate Random Effects Time Series Model

Consider monitors $j = 1, 2, 3$ in each year $t = 1, \dots, 8$ and for hospitals $i = 1, \dots, I = 152$. On the three monitors, we have observed outcomes $\mathbf{z}_{it} = (z_{i1t}, z_{i2t}, z_{i3t})'$, representing three conditionally independent binomial responses out of totals $\mathbf{n}_{it} = (n_{i1t}, n_{i2t}, n_{i3t})'$ and with “success” probabilities $\mathbf{p}_{it} = (p_{i1t}, p_{i2t}, p_{i3t})'$, respectively. The joint density is

$$p(\mathbf{z}_{it} | \mathbf{n}_{it}, \mathbf{p}_{it}) = \prod_{j=1}^3 \text{Bin}(z_{ijt} | n_{ijt}, p_{ijt}). \quad (3.1)$$

The p_{ijt} are hospital-specific parameters to be estimated, and the totals n_{ijt} are assumed uninformative about p_{ijt} . Our models for the p_{ijt} combine within-year random effects/hierarchical components with multivariate time series structure, now detailed.

3.1 Regression and hierarchical/random effects structure

For hospital i , the DRG-based predicted proportion of “successes” d_{ijt} is supposed to predict p_{ijt} on the basis of system-wide studies of patient case-mix profiles and historical data. Following Burgess et al (1996) we adopt a logistic regression as follows. Let $\mu_{ijt} = \log(p_{ijt}/(1 - p_{ijt}))$ and $x_{ijt}^* = \log(d_{ijt}/(n_{ijt} - d_{ijt}))$, and define $x_{ijt} = x_{ijt}^* - \bar{x}_{\cdot jt}^*$ where $\bar{x}_{\cdot jt}^*$ is the arithmetic mean of the x_{ijt}^* across all hospitals $1 = 1, \dots, I$. The logistic regression is $\mu_{ijt} = \beta_{0jt} + \beta_{1jt}x_{ijt}$ where the regression parameters β_{0jt} and β_{1jt} are unrestricted. In terms of the vectors $\boldsymbol{\mu}_{it} = (\mu_{i1t}, \mu_{i2t}, \mu_{i3t})'$, $\boldsymbol{\beta}_{0t} = (\beta_{01t}, \beta_{02t}, \beta_{03t})'$, $\boldsymbol{\beta}_{1t} = (\beta_{11t}, \beta_{12t}, \beta_{13t})'$, and matrices $\mathbf{X}_{it} = \text{diag}(x_{i1t}, x_{i2t}, x_{i3t})$, we have

$$\boldsymbol{\mu}_{it} = \boldsymbol{\alpha}_{it} + \mathbf{X}_{it}\boldsymbol{\beta}_{1t} + \boldsymbol{\nu}_{it} \quad (3.2)$$

where $\boldsymbol{\alpha}_{it} = (\alpha_{i1t}, \alpha_{i2t}, \alpha_{i3t})'$ and $\boldsymbol{\nu}_{it} = (\nu_{i1t}, \nu_{i2t}, \nu_{i3t})'$. For each monitor j and year t , the quantity α_{ijt} is an *absolute* hospital-specific random effect representing systematic variability that is related over time within each hospital. The ν_{ijt} represent residual, unpredictable variability, independent over time and across hospitals and monitors. The model assumes $\boldsymbol{\nu}_{it} \sim N(\boldsymbol{\nu}_{it} | \mathbf{0}, \mathbf{V})$ with monitor-specific variances v_1^2, v_2^2 and v_3^2 on the diagonal of the matrix \mathbf{V} , admitting cross-monitor dependencies through the covariances in \mathbf{V} .

Key to assessing quality levels are the *relative* random effects $\epsilon_{ijt} = \alpha_{ijt} - \beta_{0jt}$, i.e., hospital-specific deviations from the population level β_{0jt} . In terms of these quantities,

$$\boldsymbol{\mu}_{it} = \boldsymbol{\beta}_{0t} + \mathbf{X}_{it}\boldsymbol{\beta}_{1t} + \boldsymbol{\epsilon}_{it} + \boldsymbol{\nu}_{it} \quad (3.3)$$

where $\boldsymbol{\epsilon}_{it} = (\epsilon_{i1t}, \epsilon_{i2t}, \epsilon_{i3t})'$. This class of models accounts for variability over time in the hospital/monitor parameters $\boldsymbol{\beta}_{0t}$ and $\boldsymbol{\beta}_{1t}$ as well as the random effects $\boldsymbol{\alpha}_{it}$ that together will account for the high levels of observed extra-binomial variability.

The parameter $\boldsymbol{\beta}_{0t}$ represents the hospital system-wide average in corrected responses on the logit scale. Management policies across the VA system, and improvements (or otherwise) in care provision impacting all hospitals in similar ways contribute to changes in $\boldsymbol{\beta}_{0t}$ from year to year. We do not currently impose structure on the hospital/monitor population parameters $\boldsymbol{\beta}_{0t}$ and $\boldsymbol{\beta}_{1t}$. Predictive models, by contrast, would require evaluation of expert opinion about the reasons behind any inferred time evolution and the use of this in phrasing appropriate model extensions.

The $\boldsymbol{\epsilon}_{it}$ terms represent hospital-specific departures from the system-wide underlying level $\boldsymbol{\beta}_{0t}$. In Section 3.2 we model time series dependence over the years in the $\boldsymbol{\epsilon}_{it}$ quantities to explain the structured variability over time. However, time series models introduce partial stochastic constraints so that some of the evident variation in the logit parameters $\boldsymbol{\mu}_{it}$ will be unexplained by the regression and hospital-specific random effects $\boldsymbol{\epsilon}_{it}$. Hence the need for the residual random components $\boldsymbol{\nu}_{it}$.

3.2 Time series structure of random effects

Time series structure in the hospital-specific $\boldsymbol{\epsilon}_{it}$ is modelled via a vector autoregression of order one – or VAR(1) model. This is a natural, interpretable model incorporating the view that there should be stability in the $\boldsymbol{\epsilon}_{it}$ values within each hospital over such a short number of years. This stability represents true quality levels and any changes beyond this reflect unexplained random variations year to year due to the characteristics of the patient sample in each hospital. With such a short time span, more complex models are largely untenable. Moreover, the VAR(1) model has the desirable consequence that the annual marginal distributions of the

hospital-specific effects are the same across years. The model structure is

$$\boldsymbol{\epsilon}_{it} = \Phi \boldsymbol{\epsilon}_{i,t-1} + \boldsymbol{\omega}_{it} \quad (3.4)$$

over years t and independently across hospitals i within each year. Here $\Phi = \text{diag}(\phi_1, \phi_2, \phi_3)$ is the diagonal matrix of monitor-specific autoregressive coefficients. The $\boldsymbol{\omega}_{it}$ terms are innovations vectors, with $\boldsymbol{\omega}_{it} \sim N(\boldsymbol{\omega}_{it} | \mathbf{0}, \mathbf{U})$ conditionally independent over time. In any year t , we have the implied marginal distribution $\boldsymbol{\epsilon}_{it} \sim N(\boldsymbol{\epsilon}_{it} | \mathbf{0}, \mathbf{W})$; the within-year relative random effects are a random sample from a zero-mean normal distribution. This is consistent with a view of no global changes in the hospital population makeup, i.e., with variability in expected levels being essentially constant over the short period of years once the DRG predictor and any system-wide changes are accounted for through $\boldsymbol{\beta}_{1t}$ and $\boldsymbol{\beta}_{0t}$, respectively. Changes in relative performance of hospitals can therefore be assessed across years.

It follows that \mathbf{W} satisfies $\mathbf{W} = \Phi \mathbf{W} \Phi + \mathbf{U}$, so that correlation patterns in \mathbf{U} and \mathbf{W} , depend on the autoregressive parameters. In particular, for each monitor pair j, h we have covariance elements $\mathbf{W}_{jh} = \mathbf{U}_{jh} / (1 - \phi_j \phi_h)$. The matrix \mathbf{W} represents the variability in the systematic components of corrected quality levels across the entire hospital population, the related variability in changes in relative quality levels year-to-year, and the dependencies between such quality measures across the three monitors. The autoregressive parameters ϕ_j will generally be close to one, lying in part of stationary region $0 < \phi_j < 1$. Large values of ϕ_j imply high positive correlations between the $\boldsymbol{\epsilon}_{it}$ in a given hospital over the years. This is consistent with the view that a hospital that is generally “good” in a specific monitor/care in one year will have a high probability of remaining “good” the next year, and vice versa.

In terms of the absolute random effects $\boldsymbol{\alpha}_{it}$ we have a centred VAR(1) model

$$\boldsymbol{\alpha}_{it} = \boldsymbol{\beta}_{0t} + \Phi(\boldsymbol{\alpha}_{i,t-1} - \boldsymbol{\beta}_{0,t-1}) + \boldsymbol{\omega}_{it} \quad (3.5)$$

for $t > 1$, with yearly margins $N(\boldsymbol{\alpha}_{it} | \boldsymbol{\beta}_{0t}, \mathbf{W})$. Another feature to note concerns the time series structure of the combined hospital-specific random effects $\boldsymbol{\epsilon}_{it} + \boldsymbol{\nu}_{it}$ above. The addition of the residual/noise terms $\boldsymbol{\nu}_{it}$ to the VAR(1) process $\boldsymbol{\epsilon}_{it}$ modifies the correlation structure giving a VARMA(1,1) model with $N(\boldsymbol{\epsilon}_{it} + \boldsymbol{\nu}_{it} | \mathbf{0}, \mathbf{W} + \mathbf{V})$ yearly margins. Note that the overall levels of random effects variability, and the associated overall measures of cross-monitor dependencies, are represented through $\mathbf{W} + \mathbf{V}$. Our current model leaves \mathbf{V} and \mathbf{W} unrelated *a priori*, but the framework obviously permits the assessment of potential similarities in posterior inferences.

Finally, we assume constant values of Φ and \mathbf{U} in the time series components. This assumption could be relaxed to allow for differing variances across hospitals and/or years as may be desirable for other applications.

3.3 Prior distributions

Inference is based on posterior distributions for all model parameters and random effects under essentially standard reference/uninformative priors for: (a) the annual population parameters β_{0t} and β_{1t} , (b) the population residual variance matrix \mathbf{V} , and (c) the variance-covariance matrix \mathbf{U} ; the prior is completed with independent uniform priors for the autoregressive parameters ϕ_j on $(0,1)$.

4 Results for the VA Data

Various marginal posterior distributions from the multiple monitor analysis are reported and discussed here. First, Figure 2 provides summaries of the marginal posteriors for correlations and standard deviations in \mathbf{W} , \mathbf{V} and $\mathbf{W} + \mathbf{V}$. Here, and below, boxplots are centred at posterior medians, drawn out to posterior quartiles, and have notches at points 1.5 times the interquartile range beyond the edges of each box. These graphs indicate low overall correlations in each matrix. We focus on the key matrix $\mathbf{W} + \mathbf{V}$ that measures within-year, cross-monitor structure. Denoting posterior means by “hats” and writing \mathbf{E} for the column eigenvector matrix of $\hat{\mathbf{W}} + \hat{\mathbf{V}}$, we

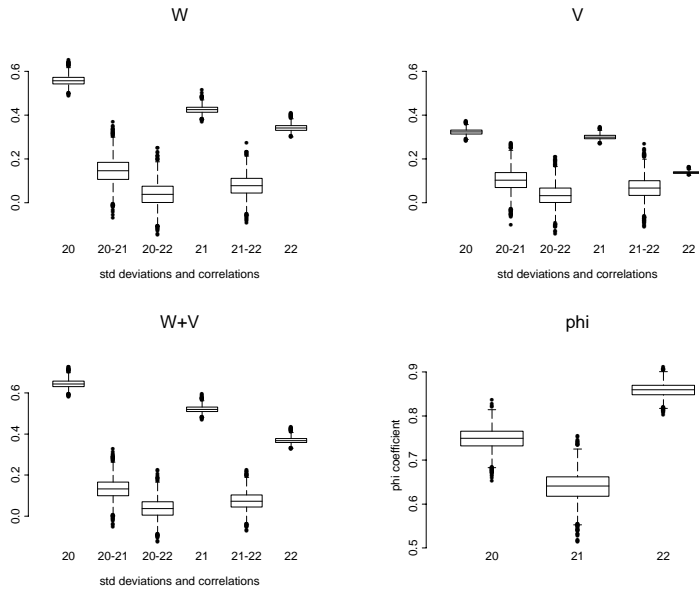
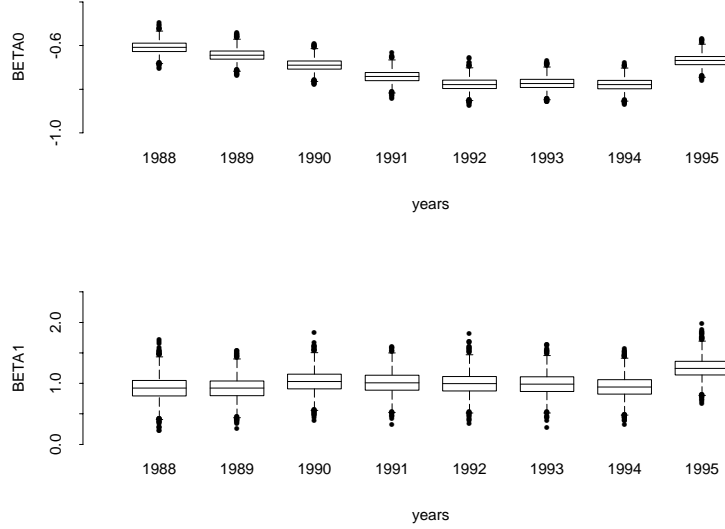


FIGURE 2. Posterior summaries for AR coefficients Φ and standard deviations and correlations of \mathbf{V} , \mathbf{W} and $\mathbf{V} + \mathbf{W}$.

FIGURE 3. Posterior summaries for β_{02t} and β_{12t} (Monitor 22) over the years.

have

$$\hat{\mathbf{W}} + \hat{\mathbf{V}} = \begin{pmatrix} 0.417 & 0.044 & 0.009 \\ 0.044 & 0.271 & 0.014 \\ 0.009 & 0.014 & 0.136 \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} 0.961 & -0.275 & -0.015 \\ 0.273 & 0.957 & -0.097 \\ 0.041 & 0.089 & 0.995 \end{pmatrix}.$$

This indicates correlations between M20 and M21 of around 0.13, between M20 and M22 of 0.04 and between M21 and M22 of 0.07, so supporting the suggestion that the correlation between M20 and M21 might be higher than any other combination, in view of the care areas of origination. The eigenvalues of $\hat{\mathbf{W}} + \hat{\mathbf{V}}$ are roughly 0.43, 0.26 and 0.13, so the principal components explain roughly 52%, 32% and 16% of variation; each of the eigenvectors is therefore relevant, and no data reduction seems appropriate. Posterior uncertainty about the variance matrices, and the eigen-structure, does not materially impact these qualitative conclusions. To exemplify this, the full posterior sample produces the following approximate posterior means and 95% intervals for the three eigenvalues of $\mathbf{W} + \mathbf{V}$: 0.42 (0.38-0.48), 0.25 (0.22-0.29), 0.13 (0.11-0.16), closely comparable to the estimates quoted above. Evidently, the eigenvector matrix \mathbf{E} is dominated by the diagonal terms, and all three are close to unity. Note that the eigenvector matrix would be the identity were the monitors uncorrelated. The first column represents an average of M20 and M21 dominated by the M20 psychiatric care component. The second column represents a contrast between M20 and M21 and the final column almost wholly represents M22 alone, and to the

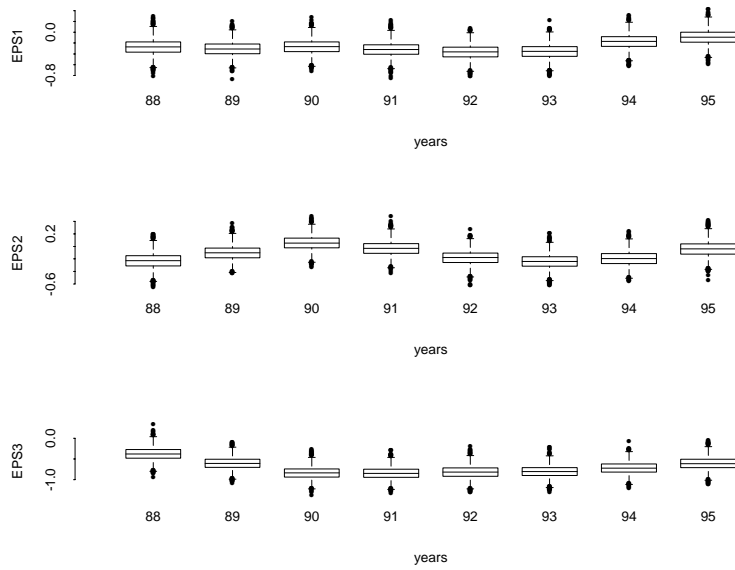


FIGURE 4. Posterior summaries for hospital-specific random effects ϵ_{ijt} on Monitor 22 over years t in hospitals $i = 2, 41$ and 92 .

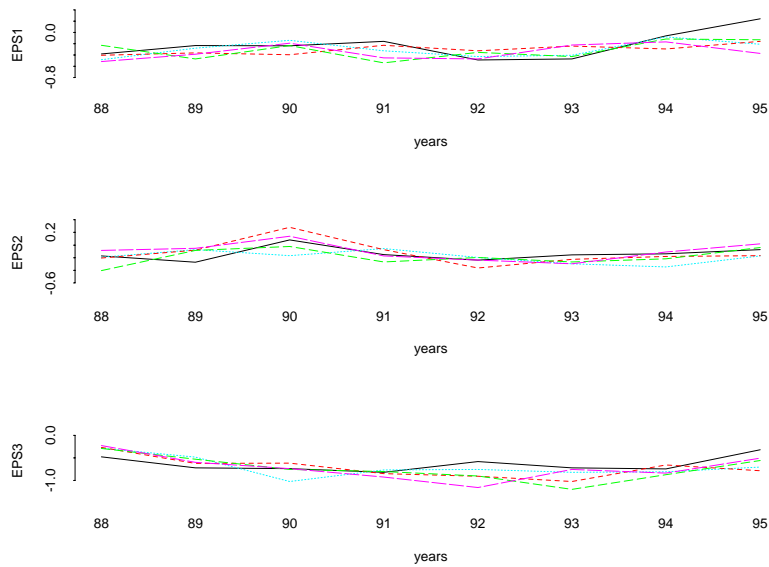


FIGURE 5. Selected posterior samples for hospital-specific random effects ϵ_{ijt} on Monitor 22 over years t in hospitals $i = 2, 41$ and 92 .

extent that the coefficients for M20 and M21 are non-ignorable, contrasts the two psychiatric care monitors with the general medical. The levels of correlation structure are clearly low for these specific monitors, perhaps surprisingly so for the first two in closely related care areas.

The lower right frame of Figure 2 provide summaries of the marginal posteriors for the three autoregressive parameters in Φ . These indicate highly significant dependence structures in each case, with inferred values of ϕ in the ranges 0.7 – 0.8 for M20, 0.6 – 0.75 for M21 and 0.8 – 0.9 for M22. The dependence in the random effects time series is high in each case, but there are apparent differences between M22 and the other two monitors, perhaps associated different health care areas.

There are meaningful differences in the β_0 parameters across the eight years in each of the three monitors. The main feature is a general decreasing trend in β_0 over the years for all three monitors, more markedly so for Monitors M20 and M21. This corresponds to generally increased probabilities of return for out-patient visits within 30 days of discharge (i.e., increased “quality”), and the apparent similarities between Monitors M20 and M21 are consistent again with the two being related areas of care. Posterior distributions for β_{03t} and β_{13t} across years t are displayed in Figure 3. For this monitor, M22, the level β_{03t} decreases over the years and levels off in 1993-4, but then exhibits an abrupt increase in 1995 that requires interpretation from VA personnel. The DRG regression coefficients β_{13t} are apparently stable over the years. They do exhibit real differences across monitors (not graphed), although the limited ranges of the DRG predictor variable limit the impact of this regression term on overall conclusions.

Posterior distributions for the variances v_j of the residual components ν_{ijt} indicate non-negligible values in comparison with the posteriors for the w_j . The v_j parameters are in the ranges of 0.3 – 0.37 for M20, 0.27 – 0.33 for M21 and 0.12 – 0.16 for M22. In terms of the variance ratio $v_j^2/(v_j^2 + w_j^2)$, the ϵ_{ijt} residuals contribute, very roughly, about 20 – 25% variation for M20, about 30-35% for M21, but only about 15% for M22.

Figure 4 displays posterior distributions for the relative random effects ϵ_{ijt} for three arbitrarily selected hospitals, those with station numbers 2, 41 and 92 for Monitor M22. Figure 5 displays five randomly chosen sets of posterior sampled values for these effects to give some idea of joint posterior variability. These summaries and examples highlight the kinds of patterns of variation exhibited by the random effects within individual hospitals – the plots indicate the smooth, systematic dependence structure over time that is naturally expected. Hospitals that have tended to be below the population norm in terms of its proportions of outcomes in recent years will be expected to maintain its below average position this year, so that the ϵ parameters of this hospital will tend to be of the same sign. Hospitals whose effects change sign at some point might be flagged as “interesting” cases for follow-up study.

4.1 Residual structure analysis

The model implies approximate normality of the standardised data residuals $e_{ijt} = (y_{ijt} - \mu_{ijt})/s_{ijt}$ where, y_{ijt} is the logit of the observed proportion z_{ijt}/n_{ijt} and s_{ijt} the corresponding approximate standard deviation. Posterior samples of the μ_{ijt} lead to posterior samples of the e_{ijt} that can be graphed to explore aspects of model fit, and misfit. Figure 6 illustrates this for M21 in 1995. The first four frames display one such sample of residuals—plotted against hospital number, against the n_{it} , in a normal quantile plot,

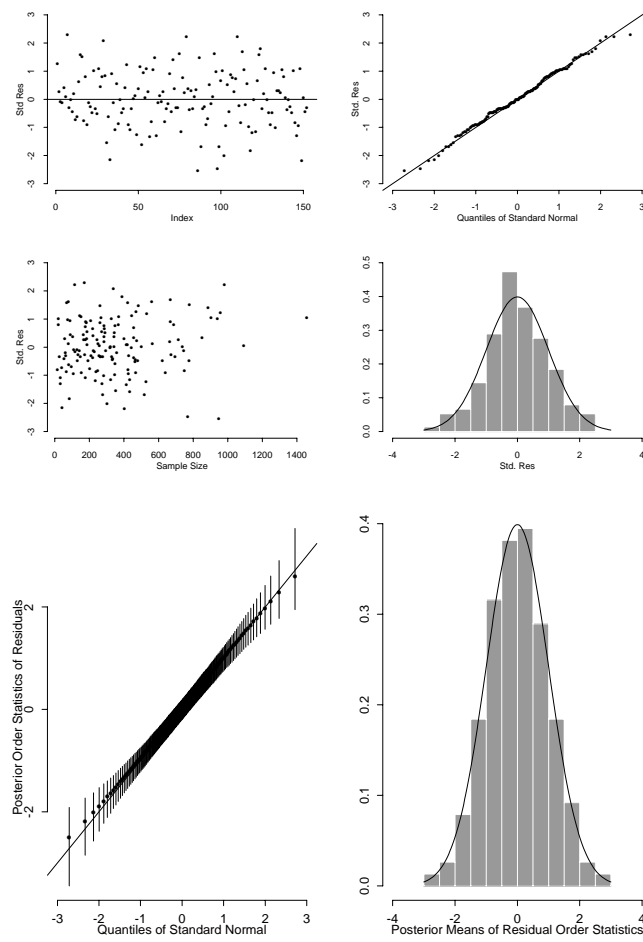


FIGURE 6. Draw from the posterior distribution of the observation residuals e_{ijt} across hospitals in M21 1995 (the four upper graphs). Posterior means of the ordered residuals e_{ijt} (the lower two graphs).

and finally a histogram with a normal density superimposed. The general impression is that of good conformity to normality, and this is repeated across many other samples of residuals, providing a measure of assurance of adequacy of this modelling assumption. The final two frames provide more global assessments. Here we explore the posterior means of the *ordered* observation residuals across all hospitals for M21 in 1995; in terms of a normal quantile plot with approximate 95% posterior intervals marked, and in terms of a histogram with the normal density superimposed. Again, adequacy of the normality assumption is indicated.

4.2 Summary Inferences for Monitor M21 in 1995

To illustrate additional uses of the model, we focus on M21 in 1995. Some summary posterior inferences appear in Figure 7, where a few specific hospitals are highlighted (with intervals drawn as dashed lines). Figure 7(a) displays approximate 95% intervals for the actual outcome probabilities p_{ijt} , ordering hospitals by posterior medians. Interval widths reflect posterior uncertainty which is a decreasing function of sample size. Hospitals with low n_{ijt} have wider intervals—hospitals 66, 86 and 114, for example. Figure 7(c) displays corresponding intervals for the $\epsilon_{ijt} + \nu_{ijt}$. The “low” hospitals have random effects lower than average, indicating that the model has adapted to the extreme observations. Adaptation is constrained by the model form and also by the high values of the DRG predictor. There is a general increasing trend in the random effects consistent with the ordering by outcome probabilities, though the pattern is not monotonic as the probabilities include the effects of the DRG predictor whereas the $\epsilon_{ijt} + \nu_{ijt}$ measure purely relative performance levels.

Figure 7(b) displays 95% posterior intervals for the *ranks* of the hospitals according to the p_{ijt} , and Figure 7(d) the intervals for ranks of the $\epsilon_{ijt} + \nu_{ijt}$. Evidently, the four or five hospitals with the highest (lowest) *estimated* outcome probabilities have very high (low) ranks, indicating that their *true* outcome probabilities are very likely to be among the largest (smallest) few across the system. Note that ranks based on p_{ijt} summarise absolute performance, impacted by patient-mix and other confounding factors, and ranks based on $\epsilon_{ijt} + \nu_{ijt}$ represent relative quality levels once these factors are accounted for via the model; the latter provide a firmer basis for assessing relative performance due to hospital-specific policies and practices. This is evident in the cases of hospitals 66, 86 and 114 noted above, for which appropriately lower rankings are indicated in Figure 7(d) than in the “unadjusted” rankings in Figure 7(b). Even then, there is high uncertainty about rankings for most hospitals, not only those with small sample sizes, reflecting the inherent difficulties in ranking now well understood in this and other areas (e.g., Normand et al 1997).

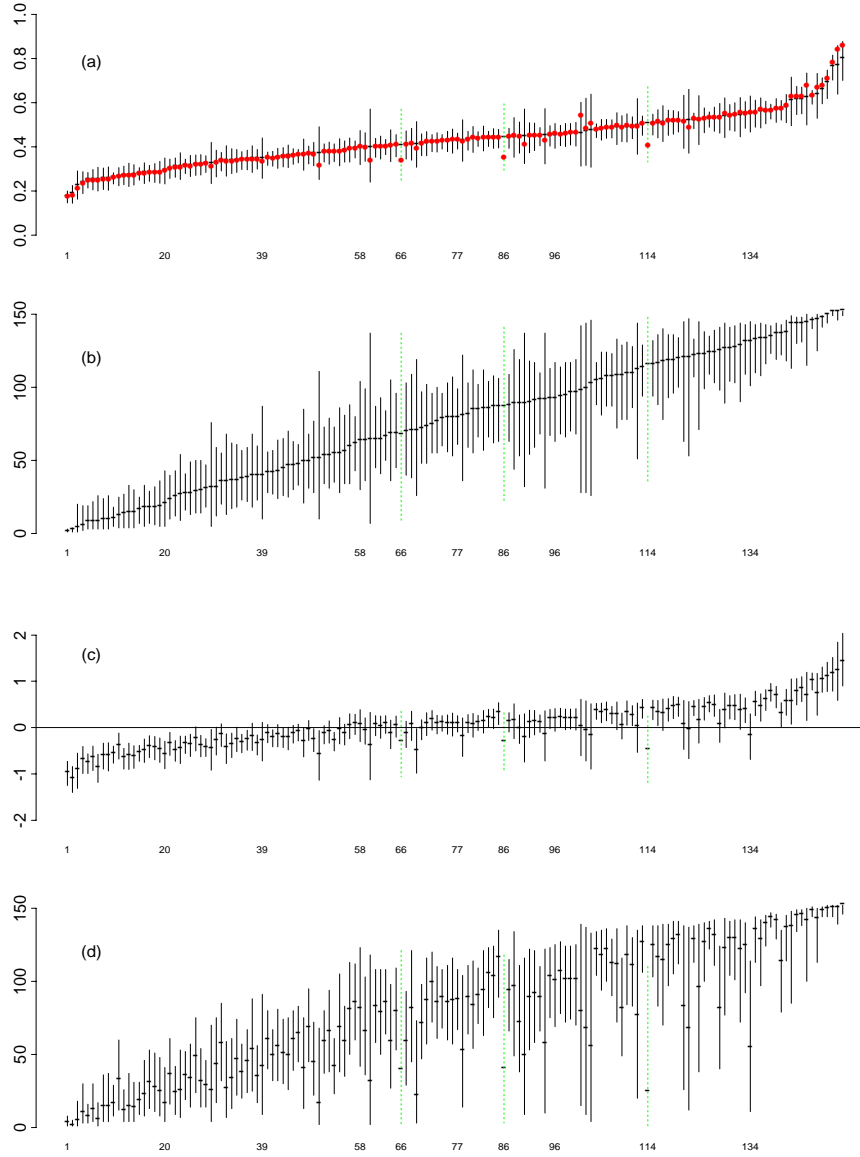


FIGURE 7. Posterior 95% intervals for all hospitals on M21 in 1995. (a) Outcome probabilities p_{ijt} (with dots marking the observed proportions), and (b) corresponding ranks of hospitals based on posterior for ordered p_{ijt} . (c) Composite random effects $\epsilon_{ijt} + \nu_{ijt}$, and (d) the corresponding ranks based on posterior for ordered random effects. The hospitals are graphed in order of posterior medians of p_{ijt} in each frame.

5 Summary Comments

We have presented a new class of multiple monitor, hierarchical random effects time series models to evaluate patterns of dependencies in series of annual measures of health care quality in the VA hospital system. A critical feature of our work has been the identification of several components of variability underlying within-year variation and across-year changes in observed quality levels. We split the hospital-specific variation in into two components: a partially systematic and positively dependent VAR component ϵ_{it} , and a purely unpredictable component ν_{it} . The latter component is non-negligible and contributes between 15-30% of the total random effects variance on the logit scale. Lower contributions in *general medical discharge* monitor than either of the psychiatric monitors. Hence hospital-specific levels of M22 are more stable over time and hence more predictable. Our multiple monitor time series models isolate changes over time and dependencies among such changes in the hospital-specific random effects across the three monitors. Though dependencies across monitors exist, they are apparently quite small. Summary graphs of posterior inferences for specific monitor:year choices provide useful insight into the distribution of outcome probabilities across the hospital system, about relative levels of performance, and about changes over time in such levels. There are evident changes in system-wide levels β_{0t} that require consideration and interpretation though such is beyond the data-analytic scope of our study. Display and assessment of posterior samples of model components provide insight in aspects of model fit and underpin the reported posterior inferences.

It should be clear that the models and computational methods (briefly detailed in the appendix below) may be applied in other contexts, and that the basic binomial sampling model may be replaced by other non-Gaussian forms as context demands. We expect that the work will be developed in such ways and that the models will find use in various other applications in the socio-economic arena.

Acknowledgements

This work was performed in collaboration with Jim Burgess and Ted Stefos of the VA Management Science Group, Bedford MA, and with important contributions and discussions with Cindy Christiansen, Carl Morris and Sarah Michalak. Our models and data exploration build on foundational contributions in hierarchical modelling of the hospital monitor data by Christiansen and Morris (see Burgess et al 1996). Corresponding author is Mike West, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251, USA; <http://www.stat.duke.edu>

Appendix: Model Theory and Computation

The full joint posteriors for all quantities $\{\mathbf{V}, \mathbf{U}, \Phi\}$ and $\{\beta_{0t}, \beta_{1t}, \alpha_{i,t}, \mu_{it}\}$ for all t was simulated via customised MCMC methods. The structure is related to that in Cargnoni, Müller and West (1997), although the methods involve substantial novelty as we are analysing new models. Results reported are based on a correlation-breaking subsample of 5,000 draws from a long chain of 100,000 iterates. The collection of conditional posteriors is briefly summarised here. Some are simulated easily, and are detailed without further comment; others require Metropolis-Hastings steps, which are noted as needed. By way of notation, for any set of parameters ξ , write ξ^- for the remaining parameters combined with the full data set \mathbf{Z} . The conditionals are as follows.

- For the β_{0t} we have conditional posteriors $N(\beta_{01}|\mathbf{b}_{01}, \mathbf{W}/I)$ and, for $t > 1$, $N(\beta_{0t}|\mathbf{b}_{0t}, \mathbf{U}/I)$ where

$$\mathbf{b}_{01} = \sum_{i=1}^I \alpha_{i1}/I \text{ and } \mathbf{b}_{0t} = \sum_{i=1}^I \{\alpha_{it} - \Phi(\alpha_{i,t-1} - \beta_{0,t-1})\}/I.$$

- For the β_{1t} we have conditionals $N(\beta_{1t}|\mathbf{b}_{1t}, \mathbf{B}_{1t})$ where

$$\beta_{1t} = \mathbf{B}_{1t} \sum_{i=1}^I \mathbf{X}'_{it} \mathbf{V}^{-1} (\mu_{it} - \alpha_{it}) \text{ and } \mathbf{B}_{1t}^{-1} = \sum_{i=1}^I \mathbf{X}'_{it} \mathbf{V}^{-1} \mathbf{X}_{it}.$$

- For \mathbf{V} we have the conditional inverse Wishart $Wi(\mathbf{V}^{-1}|8I, \mathbf{H})$ where $\mathbf{H} = \sum_{i=1}^I \sum_{t=1}^8 \nu_{it} \nu'_{it}$.
- For \mathbf{U}^{-1} we have conditional density

$$p(\mathbf{U}^{-1}|\{\epsilon_{it}\}, \Phi) \propto a(\mathbf{U}) Wi(\mathbf{U}^{-1}|7I, \mathbf{G})$$

with $\mathbf{G} = \sum_{i=1}^I \sum_{t=2}^8 (\epsilon_{it} - \Phi \epsilon_{i,t-1})(\epsilon_{it} - \Phi \epsilon_{i,t-1})'$ and $\mathbf{W} = \Phi \mathbf{W} \Phi + \mathbf{U}$. With the inverse Wishart component as a Metropolis-Hastings proposal distribution, a candidate value \mathbf{U}^* has acceptance probability $\min\{1, a(\mathbf{U}^*)/a(\mathbf{U})\}$ where $a(\mathbf{U}) = |\mathbf{W}|^{-I/2} \exp(-\text{tr}(\mathbf{W}^{-1} \mathbf{A})/2)$ and $\mathbf{A} = \sum_{i=1}^I \epsilon_{i1} \epsilon'_{i1}$.

- For Φ we have conditional posterior

$$p(\Phi|\{\epsilon_{it}\}, \mathbf{U}) \propto p(\Phi) N(\epsilon_{i1}|\mathbf{0}, \mathbf{W}) \prod_{t=2}^8 N(\epsilon_{it}|\Phi \epsilon_{i,t-1}, \mathbf{U})$$

where $\mathbf{W} = \Phi \mathbf{W} \Phi + \mathbf{U}$. Write $\phi = (\phi_1, \phi_2, \phi_3)'$ for the diagonal of Φ , and $\mathbf{E} = \text{diag}(\epsilon_{i,t-1})$. Then the density is proportional to

$p(\Phi)c(\Phi)N(\phi|\mathbf{f},\mathbf{F})$ where $\mathbf{f} = \mathbf{F} \sum_{i=1}^I \sum_{t=2}^8 \mathbf{E}'\mathbf{U}^{-1}\boldsymbol{\epsilon}_{it}$ and $\mathbf{F}^{-1} = \sum_{i=1}^I \sum_{t=2}^8 \mathbf{E}'\mathbf{U}^{-1}\mathbf{E}$. A Metropolis-Hastings step generates a candidate Φ^* from the truncated multivariate normal here, and accepts it with probability $\min\{1, c(\Phi^*)/c(\Phi)\}$ where

$$c(\Phi) = |\mathbf{W}|^{-I/2} \exp(-\text{tr}(\mathbf{W}^{-1}\mathbf{A})/2)$$

with $\mathbf{A} = \sum_{i=1}^I \boldsymbol{\epsilon}_{i1}\boldsymbol{\epsilon}'_{i1}$ and $\mathbf{W} = \Phi\mathbf{W}\Phi + \mathbf{U}$.

- The conditional for the $\boldsymbol{\alpha}_{it}$ is complicated but easily structured and sampled with dynamic linear modelling ideas. Write \mathbf{y}_{it} for the vector of logit transforms of the observed outcome proportions. Then, for each t ,

$$\begin{aligned} \tilde{\mathbf{y}}_{it} &= \boldsymbol{\alpha}_{it} + \boldsymbol{\eta}_{it} \text{ with } \boldsymbol{\eta}_{it} \sim N(\boldsymbol{\eta}_{it}|\mathbf{0}, \mathbf{V} + \mathbf{S}_{it}), \\ \boldsymbol{\alpha}_{it} &= \boldsymbol{\beta}_{0t} + \Phi(\boldsymbol{\alpha}_{i,t-1} - \boldsymbol{\beta}_{0,t-1}) + \boldsymbol{\omega}_{it} \end{aligned}$$

where $\tilde{\mathbf{y}}_{it} = \mathbf{y}_{it} - \mathbf{X}_{it}\boldsymbol{\beta}_{1t}$ and $\mathbf{S}_{it} = \text{diag}(s_{i1t}, s_{i2t}, s_{i3t})$ is the diagonal matrix of approximate data variances in the normal-logit model. This is a multivariate dynamic linear model with known variance matrices and state vector sequence $\boldsymbol{\alpha}_{it}$. Standard results for simulation in DLMS now apply, as in West and Harrison (1997, chapter 15).

- For $\boldsymbol{\mu}_{it}$ we have $p(\boldsymbol{\mu}_{it}|\boldsymbol{\mu}_{it}^-, \mathbf{z}_{it}) \propto p(\mathbf{z}_{it}|\mathbf{n}_{it}, \boldsymbol{\mu}_{it})p(\boldsymbol{\mu}_{it}|\boldsymbol{\mu}_{it}^-)$ where the likelihood function $p(\mathbf{z}_{it}|\mathbf{n}_{it}, \boldsymbol{\mu}_{it})$ is the product of the three binomial-logit functions, and $\boldsymbol{\mu}_{it}|\boldsymbol{\mu}_{it}^- \sim N(\boldsymbol{\mu}_{it}|\boldsymbol{\alpha}_{it} + \mathbf{X}_{it}\boldsymbol{\beta}_{1t}, \mathbf{V})$. A Metropolis-Hastings step generates a candidate $\boldsymbol{\mu}_{it}^*$ from the posterior based on the normal-logit approximation to the likelihood function. This delivers the proposal density $\boldsymbol{\mu}_{it}|\mathbf{y}_{it} \approx N(\boldsymbol{\mu}_{it}|\mathbf{m}_{it}, \mathbf{Q}_{it})$ where

$$\mathbf{Q}_{it} = (\mathbf{V}^{-1} + \mathbf{S}_{it}^{-1})^{-1} \text{ and } \mathbf{m}_{it} = \mathbf{Q}_{it}(\mathbf{V}^{-1}(\boldsymbol{\alpha}_{it} + \mathbf{X}_{it}\boldsymbol{\beta}_{1t}) + \mathbf{S}_{it}^{-1}\mathbf{y}_{it}).$$

The acceptance probability is $\min\{1, a(\boldsymbol{\mu}_{it}^*)/a(\boldsymbol{\mu}_{it})\}$ where $a(\cdot)$ is the ratio of the exact binomial to the approximate normal-logit likelihood.

References

- Burgess, J.F., Christiansen, C.L., Michalak, S.E., and Morris, C.N. (1996) Risk adjustment and economic incentives in identifying extremes using hierarchical models: A profiling application using hospital monitors, *Manuscript*, Management Science Group, U S Department of Veterans Affairs, Bedford MA.
- Cargnoni, C., Müller, P., and West, M. (1997) Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models, *Journal of the American Statistical Association*, **92**, 640-647.
- Christiansen, C.L., and Morris, C.N. (1997) Hierarchical Poisson regression modeling, *Journal of the American Statistical Association*, **92**, 618-632.
- Normand, S.T., Glickman, M.E., and Gatsonis, C.A. (1997) Statistics methods for profiling providers of medical care: Issues and applications. *Journal of the American Statistical Association*, **92**, 803-814.
- West, M., and Harrison, P.J. (1997) *Bayesian Forecasting and Dynamic Models*, (2nd Edn.), New York: Springer Verlag.
- West, M., and Aguilar, O. (1997) Studies of quality monitor time series: The V.A. hospital system, Report for the VA Management Science Group, Bedford, MA. *ISDS Discussion Paper #97-22*, Duke University. Available as `ftp://ftp.stat.duke.edu/pub/WorkingPapers/97-22a.ps`