Bayesian Forecasting of Multinomial Time Series Through Conditionally Gaussian Dynamic Models

Claudia CARGNONI, Peter MÜLLER, and Mike WEST

We consider inference in the class of conditionally Gaussian dynamic models for nonnormal multivariate time series. In such models, data are represented as drawn from nonnormal sampling distributions whose parameters are related both through time and hierarchically across several multivariate series. A key example—the main focus here—is time series of multinomial observations, a common occurrence in sociological and demographic studies involving categorical count data. However, we present this development in a more general setting, as the resulting methods apply beyond the multinomial context. We discuss inference in the proposed model class via a posterior simulation scheme based on appropriate modifications of existing Markov chain Monte Carlo algorithms for normal dynamic linear models and including Metropolis–Hastings components. We develop an analysis of time series of flows of students in the Italian secondary education system as an illustration of the models and methods.

KEY WORDS: Categorical data; Hierarchical dynamic model; Markov chain Monte Carlo; Posterior simulation; State-space model.

1. INTRODUCTION

Time series with multivariate, nonnormal sampling distributions arise in many areas of study, and multiple observations of categorical or ordinal count data are particularly prevalent in sociodemographic contexts. In general, when data at each time point are compositions of a total, the natural distribution of the observations is multinomial. In such cases, and especially when some or most counts are low, normal approximations are inappropriate. The new models and methods developed here are relevant to multivariate, nonnormal series quite widely, though much of the interest is related to motivating applications in modeling time variation in collections of such compositional data.

We refer to our new models as the class of conditionally Gaussian dynamic models. For a single time series of compositions, the observations at each time point follow conditional multinomial distributions whose defining parameters, after appropriate nonlinear transformations, are modeled according to a traditional dynamic linear model (West and Harrison 1989); this structure generates the “conditionally Gaussian” terminology. This provides a framework for flexibly modeling structured and stochastic time variation in multinomial proportions and for developing hierarchical model components to adequately represent relationships across categories. This framework naturally generalizes to provide models for multiple series of related vectors of compositions and to other multivariate, nonnormal contexts. Our extensions in these directions bring to dynamic models the traditional hierarchical modeling concepts of representing individual parameters in terms of stochastic deviations from underlying common structures. Such cross-sectional connections are critically relevant in analyzing problems when missing data are present in some of the series at some time points, and in making forecasts for future realizations of the series.

Our models can be viewed as flexible hierarchical generalizations and extensions of previous approaches involving multinomial data observed over time. The extensions are essentially threefold. First, our conditionally Gaussian models for individual vectors of nonnormal observations provide flexibility in modeling parametric relationships and the evolution of parameters over time not found in existing models. Second, our hierarchical model structures allow one to relate and synthesize models across two or more multivariate, nonnormal time series. Third, our very efficient Markov chain Monte Carlo (MCMC) schemes enable implementation of Bayesian analysis in a rather general framework.

Early work with multinomial time series from a Bayesian perspective involved analyses based on analytic approximations following transformation of multinomial data to approximate normality, as in the state-space models of Quintana and West (1988). Similar work, but using direct analytic approximations to analysis, was done by Fahrmeir (1992) and Fahrmeir and Kaufman (1991). A rather novel approach was taken by Grunwald, Raftery, and Guttrop (1993), who modeled multivariate series of proportions directly in the simplex, based on conditionally Dirichlet distributed vectors of multinomial probabilities, and introduced a novel “Dirichlet conjugate” distribution to develop structured and stochastic time evolution for these probabilities. Much of this can be seen as extension of early work with nonnormal models of West, Harrison, and Migon (1985) for univariate nonnormal time series. Some of the basic structure of our models is similar in spirit to the Grunwald et al. work, though it differs significantly in terms of the flexibility and generality of the treatment of models for parametric variation over time and in providing for relationships across several series of multinomial counts. Nonlinear transformations of multinomial cell probabilities are at each point in time represented by conditionally normal models, thus permitting flexibility and freedom to describe essentially arbitrary patterns of correlation across cells, unlike the overly constrained Dirichlet models. These parameters
are also modeled through time via dynamic linear models, thus permitting easy development of stochastic trend and regression components, and so forth. Further, we provide a framework for hierarchical structuring of the multivariate models, building in facilities for the modeling of cross-series structure. This relates closely to previous work in normal hierarchical models (as in Gamerman and Migon 1993), but translated into our nonnormal context.

In addition to introducing and developing useful new classes of time series models, we discuss computational approaches to analysis involving new developments in (MCMC) methods. Here we combine Gibbs sampling in normal dynamic linear models with Metropolis–Hastings ingredients to handle the inherently nonlinear/nonnormal components. Recent work in normal models has provided MCMC methods for a range of dynamic linear modeling contexts. The key algorithms for linear state-space models are due simultaneously to Carter and Kohn (1994) and Frühwirth-Schnatter (1994). Some of the extent of applicability of such methods is evident from further developments and applications of, for example, West (1995; 1996a,b). In particular, the “forward-filtering, backward-sampling” algorithm of these references provides an efficient approach to simulating the full joint posterior distribution of collections of state vectors over fixed time periods in normal dynamic linear models. We use this algorithm as part of a more general MCMC scheme in our models. Unlike other recent developments of MCMC in nonnormal and/or non-linear models, our framework is such that we do not need to resort to the usual “state-by-state” sampling in which individual state vectors are sampled conditional on their “neighbors” (see, e.g., Carlin, Polson, and Stoffer 1992, Gamerman 1995, and Shephard and Pitt 1995). This is a very relevant practical advantage of our simulation methods, as they thus are not hampered by the typically very slow MCMC convergence rates that result with such state-by-state simulation methods.

Section 2 introduces our class of conditionally Gaussian dynamic models for multinomial time series, following a preliminary discussion of the applied sociodemographic problem that originally motivated our study. Section 3 then describes implementation using a customized MCMC algorithm. Section 4 returns to analysis of data arising in our motivating application, concerning the analysis of student “flows” between grades and over school years in the Italian school system.

2. CONDITIONALLY GAUSSIAN DYNAMIC MODELS

2.1 A Motivating Case Study

Our motivating application concerns the policy-oriented problem of forecasting the number of high school students by grades in future school years in the Italian school system. Bernardi and Trivellato (1980) provided some basic models and relevant concepts in analyzing the mechanisms that regulate flows of students between grades and over years. For most grade levels, flows of students between consecutive years may be categorized into three groups: students that repeat the same grade in consecutive years, students that proceed to the following grade and do not leave the school, and students that leave the school. For simplicity, this third group actually represents the balance between the flow of students who enter and that of students who leave the system. Data on flow counts are computed from the official records of ISTAT (Istituto Nazionale di Statistica) for 1970–1993, as discussed by Bernardi and Trivellato (1980). Data are arranged into a trivariate time series for each grade, with the components being the numbers of students who repeat a grade, who proceed to the next grade, and who leave the school system. Hence we have a multinomial structure within each grade considered, and we are interested in carrying out a joint analysis for different series combining a collection of grades over a period of several years. The ISTAT data are available for different geographic aggregation levels (national data, regional data, provincial data, communal data), for the two genders, and for the different courses of the high school “cycle.” Hence, in addition to the grade disaggregation, other kinds of disaggregation could be considered, thus increasing the number of series to analyze.

Some plots of the marginal series of empirical flow proportions (one for each of the three series components) reveal similar patterns among series belonging to the same school level, or “cycle”: they have similar underlying trends but are shifted in levels. The common pattern can be justified by the existence of factors, not precisely identified, that are related to educational policy and affect the series in similar ways. These kinds of considerations motivated our choice of jointly modeling series assumed to be similar with respect to their transition probabilities, in a way that takes into account the existence of a common underlying trend that evolves in time. The empirical proportions for four grades of the high school cycle suggest a simple polynomial function to describe the underlying trend for both marginal series, but grade-specific and time-varying “offsets” might be needed to modify the trends within each individual grade. We next report on analysis of models with these and other features.

2.2 Conditionally Gaussian Dynamic Models

Suppose that we have I categorical time series $y_{it}, i = 1, \ldots, I, t = 1, \ldots, N$, where $y_{it}$ is a vector of $r$ cell counts $y_{it} = (y_{1it}, \ldots, y_{rit})'$ of a known total $n_{it}$ and $y_{i,t+1} = n_{it} - \sum_{j=1}^{r} y_{ijt}$ is the residual count in cell $r + 1$. The count vector $y_{it}$ is observed at time $t$, $(t = 1, \ldots, N)$ and is assumed to be conditionally multinomial with cell probabilities $\pi_{it}$: that is,

$$ (y_{it} | \pi_{it}, n_{it}) \sim M(\pi_{it}, n_{it}), $$

where $\pi_{it} = (\pi_{i1t}, \ldots, \pi_{irt})'$ with balance $\pi_{i,t+1} = 1 - \sum_{j=1}^{r} \pi_{ijt}$. Introduce an elementwise transformation of $\pi_{it}$ to a vector $\eta_{it}$ whose elements are real valued, with a view to defining distributions for $\pi_{it}$ indirectly by specifying multivariate normal models for $\eta_{it}$. So $\eta_{it} = (h(\pi_{i1t}), \ldots, h(\pi_{irt}))'$, $t = 1, \ldots, N$, where typically $h(\cdot)$ will be a logistic transformation in which the vector $\eta_{it}$
is composed of log-odds ratios, or an arcsin transformation 
\( \eta_{ijt} = 2 \arcsin \sqrt{\eta_{ijt}} \), or a similar transformation (Aitchison 1986). Based on this transformation and on an assumption of conditional independence of the series across indices \( i \), we now define the class of conditionally Gaussian dynamic models as follows:

a. First, the observations are independent across series with

\[
(y_{it}|\eta_{it}, n_{it}) \sim N(y_{it}|\eta_{it}, n_{it})
\]

and

\[
(y_{it}|\eta_{it}, n_{it}) \perp (y_{jt}|\eta_{jt}, n_{jt})
\]

for \( i, j = 1, \ldots, I \) and \( i \neq j \).

b. Second, the structural parameters are modeled individually across series as

\[
\eta_{it} = F_{it}\theta_t + v_{it} \quad \text{with} \quad v_{it} \sim N(v_{it}|0, V_t),
\]

where \( \theta_t \) is a \( p \)-dimensional vector of uncertain, time-varying state parameters representing both individual series-specific parameters and a set of parameters relating to an underlying "system," a common component across series. This hierarchical structure ties the series together but allows individual stochastic components through the individual \( r \)-vector term \( v_{it} \), and through individual regression and other effects represented in the \( r \times p \) design matrices \( F_{it} \) (assumed known). We combine elements across series, so that \( \eta_t = (\eta_{1t}', \ldots, \eta_{It}')' \) is the \( rI \times 1 \) vector of structural parameters for all \( I \) series at time \( t \), \( F_t = (F_{1t}', \ldots, F_{It}')' \) is the \( rI \times p \) design matrix for all \( I \) series at time \( t \), \( v_t = (v_{1t}', \ldots, v_{It}')' \) is the \( rI \times 1 \) vector of errors for the structural equations, and \( V = \text{diag}(V_1, \ldots, V_I) \) denotes the block diagonal covariance matrix combining \( V_1, \ldots, V_I \). Then for the full set of all \( I \) series, we have the equations

\[
\eta_t = F_t\theta_t + v_t \quad \text{with} \quad v_t \sim N(v_t|0, V_t),
\]

for each \( t \).

c. Third, the system equation

\[
\theta_t = H_t\theta_{t-1} + w_t \quad \text{with} \quad w_t \sim N(w_t|0, W)
\]

represents a standard dynamic linear model for the underlying system vector \( \theta_t \), with a specified \( p \times p \) state evolution matrix \( H_t \) and a stochastic "shock" \( w_t \) at time \( t \).

Model completion requires prior distributions for \( V, \theta_0, \) and \( W \) at time \( t = 0 \), based on existing initial information denoted by \( D_0 \). Various possibilities exist, depending on context and application. Fairly standard, conditionally conjugate forms, such as usually adopted in dynamic linear model (DLM) analyses (West and Harrison 1989) are used in the example in Section 4. Specifically, we assume that \( \theta_0, W, V_1, \ldots, V_I \) are mutually independent, with \( \theta_0 \) having a specified normal prior and the \( V_i \) having a common inverse Wishart prior. In our application \( W \) is block diagonal, and we use an inverse Wishart prior for each block.

Note that for \( \eta_{it}, t = 1, \ldots, N \), assumed known, and starting from the second level of hierarchy \( b \), the model corresponds to a normal dynamic linear model (NDLM) with unknown variances. At the first level of hierarchy \( a \), the series are kept distinct, and the cross-series “linkage” appears at the structural equation level through the terms \( F_{it}\theta_t \). The vector \( \theta_t \) contains parameters common to all the series, such as an underlying level and trend, and parameters specific to each series, such as individual random effects offsetting the underlying level and trend. The structure of the design matrices \( F_{1t}, \ldots, F_{It} \) and \( H_t \), as well as that of the \( \theta_t \) parameters, depend strictly on the kind of series involved and on the objectives of the analysis. Specification of the model through the choice of its defining components is entirely up to the modeler, as exemplified in the Italian school data analysis in Section 4.

3. POSTERIOR COMPUTATIONS

Here we describe an MCMC algorithm for full posterior analysis based on observing data on all \( I \) series over a fixed time period \( t = 1, \ldots, N \). Define the quantities \( y^N, \eta^N, \theta^N \) as follows: \( \eta^N = \{\eta_1, \ldots, \eta_N\}, y^N = \{y_1, \ldots, y_N\}, \) and \( \theta^N = \{\theta_0, \ldots, \theta_N\} \). We are interested in computing and summarizing the full joint posterior

\[
p(\eta^N, \theta^N, V, W|y^N)
\]

do and so via MCMC simulations. A direct Gibbs sampling approach (Gelfand and Smith 1990) involves iterative resampling from the complete conditional posterior distributions

\[
p(\eta^N|\theta^N, V, W, y^N) \leftrightarrow p(\theta^N|\eta^N, V, W, y^N) \leftrightarrow p(V, W|\theta^N, \eta^N, y^N).
\]

It turns out that the second and third conditionals here can in fact be sampled directly. The first cannot, and for that we embed a Metropolis–Hastings step—actually an independence chain step—in the Gibbs iterations (Tierney 1994). We provide details in the following sections.

3.1 Sampling From \( p(\theta^N|\eta^N, V, W, y^N) \)

Conditional on \( \eta_t, V, \) and \( W \), equations b and c in the model specify a standard dynamic linear model with known variance–covariance matrices. Note that \( \theta^N \) is conditionally independent of \( y^N, \) given \( \eta^N \). Thus standard theoretical results for DLM’s apply, and the conditional posterior for all \( N \) state vectors \( \theta^N \) may be sampled using the forward-filtering, backward-sampling algorithm (Carter and Kohn 1994; Frühwirth-Schnatter 1994). This is briefly detailed in Appendix A, and full details appear in the aforementioned references. As in the application of West (1995), we may modify the basic MCMC algorithm to avoid high-dimensional matrix inversions. This is done by partitioning the state vector \( \theta_t \) into subvectors, each of which is resampled conditional on the currently imputed values of the remaining subvectors. This can be most effective in cases where the specific model structure naturally identi-
flies appropriate partitions of \( \theta_i \). This is true in the specific model that we use in the application to the Italian schools data; the partitioning is described in Section 4. Alternative approaches, such as the Cholesky procedure suggested by Carter and Kohn (1994), may be used to avoid repeated large-scale matrix inversions, though we have not applied these in our study to date.

3.2 Sampling From \( p(V, W|\eta^N, \theta^N, y^N) \)

The posteriors for the variance matrices are essentially standard forms derived from the multivariate DLM for \( \eta^N \)—a multivariate extension of posterior forms of West (1995). We see that

\[
p(V, W|\eta^N, \theta^N, y^N) = p(W|\theta^N) \prod_{i=1}^{p} p(V_i|\eta^N, \theta^N) \quad \text{with} \quad p(W|\theta^N) \propto p(W) \left| W \right|^{-N/2} \exp(-\text{trace}(WT)) = \sum_{t=1}^{N} w_t w_t^T \text{ and } w_t \text{ is the observed evolution error } w_t = \theta_t - H_t \theta_{t-1}; \text{ and, for each } i = 1, \ldots, I, p(V_i|\eta^N, \theta^N) \propto p(V_i)|V_i|^{-N_{i}/2} \exp(-\text{trace}(V_i S_i)) = \sum_{t=1}^{N_{i}} v_{it} v_{it}^T \text{ with } v_{it} \text{ the observed residual } v_{it} = \eta_{it} - F_t \theta_{t-1}.
\]

Independent inverse Wishart priors for the \( V_i \) then imply inverse Wishart posteriors, which are easily simulated. Similarly, if \( p(W) \) is inverse Wishart, then so is \( p(W|\theta^N) \). A modification arises in models such as the model of Section 4. Here the state-space model for \( \theta_i \) is structured in terms of components that result in a block diagonal form for \( W \).

Then independent inverse Wishart (or, for scalar diagonal elements, inverse gamma) priors for the block variance matrices lead to conditionally independent inverse Wishart (or gamma) posteriors.

3.3 Updating Samples of \( \eta^N \)

Given \( \theta^N \) and the variance components, it is easily seen that both the prior and the likelihood functions for the \( \eta_{it} \) factorize completely, resulting in posterior independence of the \( \eta_{it} \) over all times \( t \) and across all series \( i \). Thus sampling reduces to independent simulations of the \( I \cdot N \) distributions

\[
p(\eta_{it}|\theta_t, V_t, y_{it}) \propto p(y_{it}|\eta_{it})p(\eta_{it}|\theta_t, V_t).
\]

The first term here is the likelihood from the multinomial model for \( y_{it} \); the second term is simply the multivariate normal prior \( \eta_{it} \sim \text{N}(\eta_{it}|F_t \theta_{t-1}, V_t) \). An independence chain method (Tierney 1994) is used to update the values of \( \eta_{it} \), as follows.

For each \( i \) and \( t \), we first derive a proposal distribution to generate candidate values of \( \eta_{it} \). We do this by approximating the multinomial likelihood by a normal form, \( \text{N}(\eta_{it}|\hat{\eta}_t, H^{-1}(\hat{\eta}_t)) \), where \( \hat{\eta}_t \) is the maximum likelihood estimate (MLE) and \( H(\hat{\eta}_t) \) is the Hessian matrix evaluated at the mode of the likelihood. Combined with the actual normal prior for \( \eta_{it} \), this leads to an approximation of the complete posterior distribution by a multivariate normal, namely

\[
\text{N}(\eta_{it}|m, Q), \quad m = [H(\hat{\eta}_t) + V_t^{-1}]^{-1}[H(\hat{\eta}_t)\hat{\eta}_t + V_t^{-1}F_t\theta_t] \text{ and } Q = [H(\hat{\eta}_t) + V_t^{-1}]^{-1}.
\]

We generate candidate values of \( \eta_{it} \) from this normal distribution, then use independence chain ideas (detailed in App. B), based on evaluations of the exact (unnormalized) posterior density, to accept or reject candidate draws. Note that under some choices of probability transformation \( h(t) \), the reverse transformation of \( \eta_{it} \) to multinomial probabilities may violate the unit sum constraint; in such cases, normal sampling subject to the constraint is needed and is trivially implemented by rejection of draws violating the constraint. This applies for each \( i \) and \( t \) independently, and represents an application of standard ideas of embedding Metropolis–Hastings steps within Gibbs sampling. Note that the approximation of the likelihood components is used only to develop simple, and effective, proposal distributions. The ergodic distribution of the resulting Markov chain for the full joint posterior simulation is exactly the desired posterior; the quality of the normal approximations will affect only the convergence rate for the defined Markov chain, not the form of the stationary distribution.

Finally, note that missing data are trivially accommodated in the analysis. If data are missing for any subset of series and time intervals, then the corresponding likelihood components are vacuous. As a result, the corresponding \( \eta_{it} \) are trivially updated; no data on \( \eta_{it} \) implies that the conditional posterior is simply the conditional multivariate normal prior, and this may be sampled directly. Our analysis of the Italian school data in Section 4 is subject to a modest amount of missing data.

3.4 Forecasting

Following posterior simulation, it is trivial to simulate forecast distributions that may be of interest in an application. We simply note that for each set of imputed parameters \( \eta^N, \theta^N, V, \) and \( W, \) we may simulate a sample from the posterior predictive distribution for future observations by successively sampling (a) “future” \( \theta_t \) values from the state-space model, followed by (b) the corresponding normal distributions for future \( \eta_t \) conditional on these state vectors, and then (c) the multinomial (or other) distributions of future data conditional on the sampled \( \eta_t \) vectors.

4. AN ILLUSTRATION IN ANALYSIS OF ITALIAN SCHOOL DATA

4.1 Data and Model Summary

We analyze a subset of the Italian school data introduced in Section 2.1, considering \( I = 4 \) high school grade levels (ages 14–18) in all schools in the province of Venice during 1970–1993. Data in series/grade \( i \) are represented by the 2-vector \( y_{it} \), with the first component being the number of students repeating grade \( i \) in year \( t + 1 \) and the second component being the number of students moving from grade \( i \) in year \( t \) to grade \( i + 1 \) in year \( t + 1 \). A third category contains the balance; that is, the students who leave school in the given year. These data are conditionally multinomial with totals \( n_{it} \) known at time \( t \). In our dataset, observed totals range from the low hundreds to several thousands of students, with some very low counts implying some extreme multinomial cell probabilities. In the years 1986–1988, most of the data are unavailable, but this is
easily accommodated as uninformative missing data in our analysis, as earlier mentioned.

Our analysis uses the arcsin transformation \( h(\pi) = 2 \arcsin \sqrt{\pi} \) to map the multinomial probabilities to the conditionally Gaussian/linear model parameters \( \eta_{it} \). Thus the 2-vector \( \eta_{it} \) comprises the transformed multinomial probabilities for the two categories just described. Changes over time in the multinomial probabilities are modeled on the transformed \( \eta \) scale with a smooth trend \( \mu_t \) over time, common to all grades, plus a grade-specific offset \( \gamma_{it} \) and a stochastic deviation \( \nu_{it} \). Specifically, we have \( \eta_{it} = \mu_t + \gamma_{it} + \nu_{it} \), where the 2-vector \( \mu_t \) represents an underlying, or average, level at time \( t \) for the distribution of student numbers in the two categories, and the 2-vector \( \gamma_{it} \) represents deviations away from this average level for grade \( i \), \((i = 1, \ldots, I)\). For identification, we constrain these offset parameters to zero sum by allowing the first \( I - 1 \) to vary freely and defining \( \gamma_{it} = -B_{i-1} \gamma_t \). We model time variation in the underlying levels using standard trend components, assumed to follow a (bivariate) third-order polynomial function (West and Harrison 1989, chap. 7). Specifically, the vector \( \mu_t \) evolves according to the system equation \( \mu_t = \mu_{t-1} + B_{t-1} + p_{t-1} + w_{it} \), where \( \mu_{t-1} + p_{t-1} + w_{it} \) and \( \nu_{it} \) evolve similarly. The vectors \( \gamma_{it} \) change through time according to a simple random walk: \( \gamma_{it} = \gamma_{it-1} + w_{it} \), where the \( w \) terms are mean 0, bivariate normal evolution errors. In the application context, observed variation in underlying multinomial probabilities is attributable to global control over the school system, including policy changes determining shifts in the proportions of students repeating grades, the systematic effects of curriculum changes and reform, and so forth.

We can write this model in the general notation of Section 2 as follows. First, the global state vector is defined as

\[
\theta_t = (\mu_t, \delta_t, \varepsilon_t, \gamma_t, \gamma_{t-1}, \ldots, \gamma_{1-t})',
\]

having \( 6 + 2(I - 1) = 2I + 4 \) elements. The state evolution equation has \( H_t = H \), a \((2I + 4)(2I + 4)\) matrix \( H = \text{block diag}(J, I) \) with block entries

\[
J = \begin{bmatrix}
1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 
\end{bmatrix}
\]

and \( 2(I - 1) \times 2(I - 1) \) identity matrix \( I \). The evolution variance matrix is also structured as \( W = \text{diag}(W_1, W_2) \), where \( W_1 \) is the \( 6 \times 6 \) variance matrix of \((\mu_t', \delta_t', \varepsilon_t')' \) and \( W_2 \) is the \((2I - 1) \times 2(I - 1) \) variance matrix of \((\gamma_t', \gamma_{t-1}', \ldots, \gamma_{1-t}')' \).

Each vector \( \theta_{it} \) is determined by \( \eta_{it} = F_t \theta_t + v_{it} \), where the time-independent \( \times 2(2I + 4) \) design matrix is

\[
F_t = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & \ldots & \ldots & 1 & 0 & 0 & \ldots & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & \ldots & \ldots & 0 & 1 & 0 & \ldots & 0 & 0 \\
\end{bmatrix}
\]

for \( i = 1, \ldots, I - 1 \) and

\[
F_1 = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & \ldots & -1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & \ldots & 0 & -1 \\
\end{bmatrix},
\]

and the \( v_{it} \) are mean 0 bivariate normal with variance matrices \( V_t \).

Analysis follows the scheme described in Section 3 but using the modification to the MCMC algorithm that samples subvectors of \( \theta_t \) conditionally, as mentioned earlier and in Appendix A. Specifically, we apply the conditional scheme mentioned in Appendix A based on partitioning \( \theta_t = (\theta'_{1t}, \theta'_{2t}) \), where \( \theta_{1t} \) is the 6-vector \((\mu_t', \delta_t', \varepsilon_t')\), and \( \theta_{2t} \) is the \((2(I - 1))\)-vector \((\gamma_{1t}, \ldots, \gamma_{1-I-t})'\).

4.2 Some Summary Inferences

Various analyses of the Italian student flows data have been performed and reported by Cargnoni (1996). Here we describe just one analysis of data from the first four “high cycle” grades in the province of Venice from 1970–1992. The required totals \( n_{it} \) at each time \( t \) after 1992 were generated by the model itself. In particular, for all grades after grade 1 of the “low cycle,” \( n_{it} \) is obtained by summing the flow of students who repeat grade \( i \) from time \( t - 1 \) to time \( t \), the flow of students who pass grade \( i - 1 \), and the flow of students who enter the system at time \( t \) for attending grade \( i \). The last number is obtained by estimating the rate \( r_{it} = \text{new}_{it}/n_{it} \); that is, the ratio of the number of new students at time \( t \) and grade \( i \) over the total number of students at time \( t \) and grade \( i \). The ratio \( r_{it} \) for year \( t \) is estimated by a weighted average (with increasing weights) over the last 5 years with observed data. A different procedure is used for grade 1 of the “low cycle,” where demographic forecasts for 6-year-old children need to be used as exogenous input to add to the flow of students who repeat the grade. (See Bernardi and Trivellato 1980 for more details.)

We applied the model of the previous section directly to the foregoing data, noting that all counts for 1986–1988 inclusive are missing. Initial prior distributions were specified as follows. For \( \theta_0 \), we adopt a prior variance covariance matrix \( C_0 \) with the same block diagonal structure as \( W \), and large variance elements reflecting a relatively uninformative prior distribution. We specified the inverse Wishart priors for the \( V_t \), and for the two blocks of \( W \) with small degree of freedom parameters.

Initial values for the parameters \((\eta^N, \theta^N, V, W)\) are needed to implement the MCMC analysis. We transformed the observed proportions \( y_{ij/st} / n_{ij/st} \) via \( 2 \arcsin \sqrt{\cdot} \) to provide initial values for all elements of \( \eta^N \), and used these same values to determine initial values for \( \theta^N \), and \( V \). To do this, we simply used standard normal theory analysis in the specified multivariate DLM’s for each grade with each \( \eta_{it} \) estimated with the transformed proportions. We ran these analyses (as in West and Harrison 1989, chap. 15) independently for each grade \( i \), and used the resulting posterior-generated values to initialize the full model analysis. We computed the series of parameters for the \( \mu_t, \delta_t, \varepsilon_t \) by averaging over the estimated parameters from the four analyses. We initialized the series of grade-specific offsets with the differences between the estimated level for the grade and the common level \( \mu_t \). We initialized the variance matrix \( V_t \) for each grade with the estimated variance of the observation equation for \( \eta_{it} \).
Of a total of 4,000 MCMC iterations, we used the final 2,000 to compute the posterior summaries in the accompanying figures. We addressed convergence issues graphically and using standard diagnostics, including those of Geweke (1992). Most series of simulated parameters stabilize almost immediately, suggestive of rapid convergence. There are two good reasons to believe this: First, we have excellent starting values available; second, the specific MCMC scheme is based on a very efficient scheme for DLM simulations, combined with the component for the $\eta_{t}$ that is a collection of conditionally independent parameters. In particular, we completely avoid convergence problems associated with very high correlations between parameters and variables in MCMC schemes for state-space models that sample state vectors one at a time given all other parameters.

A few specific summary inferences are illustrated in the figures. First, Figure 1 shows the series of the empirical transformed proportions for the four grades. The posterior mean trajectory of the underlying common trend—namely, the lead element of $E(\mu_{t}|y^{T})$ as a function of time $t$—is superimposed (dotted curves). The points plotted after 1992 represent the forecasted common trend for the series. Figure 2 shows the multivariate series for grade $i = 4$ alone. The two plotted series are the two components of the multivariate series of transformed proportions: The higher one corresponds to the students who passed the grade; the lower, to the students who repeat. The dots represent the empirical series, and the continuous curves correspond to the posterior means of $\eta_{t}$ over time; uncertainties in term of

![Figure 1. The First Four Grades of the High Cycle for the Province of Venice. Time series of the empirical transformed proportions for the students (a) who repeat and (b) who proceed to the following grade in each of grades 1 (solid line), 2 (light-dotted line), 3 (short-dashed line), and 4 (long-dashed line). The heavy dotted curves are the posterior estimates up to 1992 and forecasts from 1993–1998 for the common level of the group.](image1)

![Figure 2. Bivariate Series (Students Who Repeat and Proceed to the Following Grade) for the Fourth Grade of the High Cycle for the Province of Venice. The continuous line represents the estimated posterior mean for $\eta_{t}$, and uncertainties in terms of 90% equal tail intervals are shown. The dots represent the observations.](image2)

![Figure 3. Simplex Plot of the Multivariate Series of Proportions (Probabilities of the Students to Repeat, to Pass, or to Leave the School).](image3)
90% equal tail intervals are also plotted. Figure 3 displays a summary of the complete multivariate series for the common trend, where the vertices of the triangle represent the three extreme situations for the components of the series of probabilities; here the estimated series up to 1992 and five point forecasts are plotted. This figure shows improving performances of the students at the secondary school after the 1970s—the probability of leaving school declines, whereas the probability of passing grades increases. Figure 4 displays observed multivariate series of absolute flows of students who repeat grade 4 (the lower curve) and those who pass grade 4 (the higher curve), together with several point forecasts of future flows. Despite the increased probability of passing grade 4, the absolute flow of students who pass decreases, because the number of students enrolling into the “high cycle” decreases.

In addition to the direct estimation and forecasting of grade-specific transition probabilities and their underlying common components, a nice feature of the analysis is the ease with which we can generate step-ahead forecasts of actual future student flows between grades. This is of interest in feeding into policy considerations in the school system. Additional commentary on such issues, and on other questions of model generalization and evaluation, have been provided by Cargnoni (1996).

5. CONCLUDING COMMENTS

The class of conditionally Gaussian dynamic models is a rather general class of models for analysis and forecasting of a group of multivariate, nonnormal time series. Although developed here in the specific context of multinomial observation models, the underlying ideas and technical/computational developments are evidently more general, providing opportunity for future application with other distributional forms. The hierarchical structure of our models proves useful in representing random deviations of individual series away from an underlying common system, but with deviations related in structured ways through time and across series. Issues of computation are quite challenging, though the developments using MCMC here utilize most efficient schemes for simulation in DLM’s, made possible by the reduction of certain conditional distributions to multivariate normal forms arising in standard DLM theory. Additional developments of current interest include questions of model assessment and evaluation and specific issues of modeling “outlying” series (Cargnoni 1996) within the hierarchical framework. We expect to report on such developments in the near future.

APPENDIX A: SIMULATIONS FOR $\theta^N$

Consider the following representation of the posterior distribution for $\theta^N$:

$$p(\theta^N | \eta^N, V, W) = \prod_{t=0}^{N-1} p(\theta_t | \theta_{t+1}, \ldots, \theta_N, \eta^N, V, W).$$

The conditional independence structure of the DLM implies that

$$p(\theta_t | \theta_{t+1}, \ldots, \theta_N, \eta^t, V, W) = p(\theta_t | \theta_{t+1}, \eta^t, V, W)$$

and, by Bayes’s theorem,

$$p(\theta_t | \eta^t, V, W) \propto p(\eta^t | \theta_t, V, W)p(\theta_t | \eta^t, V, W).$$

The forward-filtering, backward-sampling algorithm used for sampling from the full posterior for $\theta^N$ exploits these results, as follows:

a. For $t = 1, \ldots, N$, compute the moments $m_t$ and $C_t$ of the normal posteriors $p(\theta_t | \eta^t, V, W)$, $t = 1, \ldots, N$ by applying the standard sequential updating results for normal DLM’s (West and Harrison 1989, chap. 4).

b. At $t = N$, sample the final state vector $\theta_N$ from the marginal distribution $p(\theta_N | \eta^N, V, W) = N(m_N, C_N)$.

c. Sequence through $t = N-1, \ldots, 0$, sampling from $p(\theta_t | \theta_{t+1}, \eta^t, V, W)$ at each time conditional on the latest value of $\theta_{t+1}$ just sampled. Following Carter and Kohn (1994), Frühwirth-Schnatter (1994), and West (1995), the required distribution here is an easily computed normal distribution.

The result is a draw $\theta_N, \theta_{N-1}, \ldots, \theta_0$ from the full conditional posterior.

A modified version is applied in our example, using the foregoing algorithm to iteratively resample subvectors of the $\theta_t$ conditional on the remaining subvectors (as in West 1995). This has attraction in reducing quite dramatic computational overheads by avoiding repeated inversions of large-scale matrices implicit in the direct algorithm. It operates as follows in the case of just two subvectors, as used in our example. Partition $\theta_t = (\theta_{1t}, \theta_{2t})$, say, and conformably partition the design matrices $F_t = (F_{1t}, F_{2t})$. For given values of the full set $\theta_{2t}$, $t = 0, \ldots, N$, the foregoing forward-filtering, backward-sampling algorithm applies to simulate the full set of $\theta_{1t}$ vectors by simply replacing $\eta_t$ with $\eta_t - F_{1t} \theta_{2t}$ for each $t$. Similarly, a second pass of the algorithm will simulate the full set of $\theta_{2t}$ conditional on the $\theta_{1t}$ by replacing $\eta_t$ with $\eta_t - F_{2t} \theta_{2t}$ for each $t$. In this way, two passes of the forward-filtering, backward-sampling algorithm are embedded within each Gibbs iteration. At iteration $r$, the values of the $\theta_{2t}$ from iteration $r-1$ are conditioned upon to simulate new values for the $\theta_{1t}$; in turn, these new values are conditioned upon to simulate new values for the $\theta_{2t}$, so completing the simulation of the $\theta_t$ at iteration $r$. 

Figure 4. The Fourth Grade of the High Cycle for the Province of Venice. The two flow marginal series were computed from the official ISTAT data, and forecasts after 1992 were computed with 90% equal tail intervals.
APPENDIX B: SIMULATIONS FOR $\eta_{it}$

The set of $\eta_{it}$ vectors over all $i$ and $t$ are updated independently using independence chain methods. For each $i$ and $t$, a candidate vector $\tilde{\eta}_{it}$ is drawn from the multivariate normal candidate distribution derived from the likelihood approximation discussed in Section 3.3. Denote by $g_{it}(\cdot)$ the density of this candidate generating normal distribution, and denote by $p_{it}(\cdot)$ the (unnormalized) density of the exact conditional posterior $p(\eta_{it}|\theta_i, V_i, y_{it})$ of that section. Then the Metropolis step in the MCMC analysis is as follows:

a. Sample a candidate value $\tilde{\eta}$ from $g_{it}(\cdot)$, subject to the constraint $h^{-1}(\tilde{\eta}_i) + \cdots + h^{-1}(\tilde{\eta}_t) < 1$ as needed.

b. Based on the current, most recently updated value $\eta_{it}$, evaluate

$$\alpha(\tilde{\eta}, \eta) = \min \left( \frac{p(\tilde{\eta})}{p(\eta_{it})}, \frac{g(\eta_{it})}{g(\tilde{\eta})} \right).$$

Then update by replacing $\eta_{it}$ with the candidate $\tilde{\eta}$ with probability $\alpha(\tilde{\eta}, \eta_{it})$, otherwise leaving $\eta_{it}$ unchanged.

The subunit sum constraint for the $r$ components of the $\eta_{it}$ vector is imposed at this stage of the analysis, as can be seen in step a of the algorithm. It is necessary to impose the constraint here unless the chosen transformation $h(\cdot)$ guarantees it.

[Received August 1995. Revised August 1996.]

REFERENCES


