Hidden Markov Models for Longitudinal Comparisons

STEVEN L. SCOTT*, GARETH M. JAMES* AND CATHERINE A. SUGAR*

Abstract

Medical researchers interested in temporal, multivariate measurements of complex diseases have recently begun developing *health state models* which divide the space of patient characteristics into medically distinct clusters. The current state of the art in health services research uses k-means clustering to form the health states and a first order Markov chain to describe transitions between the states. This fitting procedure ignores information from temporally adjacent observations and prevents uncertainty from parameter estimation and cluster assignments from being incorporated into the analysis. A natural way to address these issues is to combine clustering and longitudinal analyses using a hidden Markov model. We fit hidden Markov models to longitudinal data using Bayesian methods which account for all the uncertainty in the parameters, conditional only on the underlying correctness of the model. Potential lack of time homogeneity in the Markov chain is accounted for by embedding transition probabilities into a hierarchical model that provides Bayesian shrinkage across time. We illustrate this approach by developing a hidden Markov health state model for comparing the effectiveness of clozapine and haloperidol, two antipsychotic medications for schizophrenia. We find that clozapine outperforms haloperidol and identify the types of patients where clozapine's advantage is greatest and weakest. Finally, we discuss the advantages and disadvantages of hidden Markov models in comparison with the current methodology.

Key Words: inhomogeneous hidden Markov model, Markov chain Monte Carlo, health state model, k-means clustering, hierarchical model

1 Introduction

Applications in many fields, from market segmentation in business to health state modeling in medicine, involve dividing a population into contextually coherent subgroups. It is frequently desirable to understand how subjects move from one group to another over time, and in particular how transition patterns are affected by different treatments applied to members of the population. Various field-specific approaches have been developed to deal with such situations, for example Sugar *et al.* (1998) in health services research. However, these methods tend to be somewhat *ad hoc*, and can potentially be improved using likelihood procedures based on hidden Markov models (HMMs). HMMs assume that observations are generated from a mixture of distributions among which subjects move according to a latent Markov chain. By incorporating treatment data into the procedure for estimating the transition matrices one can obtain direct assessments of a treatment's effectiveness. This article applies HMMs to a health state modeling problem involving the comparison of two antipsychotic medications for schizophrenia and discusses the advantages and disadvantages of this methodology relative to the current medical approaches.

*Assistant Professors of Statistics, The Marshall School of Business, University of Southern California. The authors thank the referees for helpful comments.

Clinical trials typically measure different aspects of physical and mental well-being using health status instruments or questionnaires consisting of dozens of item responses. Traditionally such data are examined by performing univariate analyses on composite scores formed from the original responses. However, clinical trial investigators have recently turned to multivariate health state models to capture structural features in the data because the phenomena being studied are too complex to be described by univariate summaries. These models divide a population's sample space into medically coherent subgroups called health states. Clinical change is measured based on the probability of moving individuals between health states, rather than by a simple net increase or decrease in the mean of a univariate continuous scale. A treatment benefits patients in a given cluster if it has a high probability of moving them to a superior state or preventing them from moving to an inferior state. Health state models have numerous advantages. In particular, they lend themselves naturally to the assessment of long-run treatment effects via the estimation of stationary distributions, and they can be used in utility elicitation and cost-benefit analyses as the basis for making objective health policy decisions.

In the medical literature, the state of the art for fitting health state models uses the k-means clustering algorithm to produce hard assignments of patients to the nearest cluster center (Sugar *et al.*, 2004). The cluster assignments are then treated as known and used to estimate matrices of transition probabilities for different medications. The clustering approach is well suited to capturing complex relationships because it allows the data to choose the optimal locations of the health states. The clustering method, though easy to implement, has some potential limitations. The k-means algorithm implicitly assumes that the data are distributed as an equally weighted mixture of Gaussian distributions with identity covariance matrices. Thus the algorithm may perform poorly if mixtures of non-spherical or non-Gaussian distributions fit the data more naturally, or if different mixing weights are needed (see Banfield and Raftery, 1993, for example). Furthermore, the k-means health state model is fit using a two stage procedure: first the cluster centers are computed assuming independent observations and then transition matrices are estimated assuming that cluster means are known and that each subject belongs to the nearest cluster with probability 1. The two stage estimation procedure ignores potentially valuable information about a subject's cluster membership during other observation periods. It also prevents uncertainty about cluster means, cluster membership, and transition probabilities from correctly propagating through the model.

The preceding limitations can be addressed by modeling the data using a hidden Markov model. Because HMMs directly model the temporal aspect of the data they can borrow strength across nearby observations when estimating model parameters and classifying observations to states. HMMs are fit using likelihood-based procedures that simultaneously estimate the transition probabilities and the parameters of the mixture components. The Bayesian methods employed in this article allow arbitrary functions of HMM parameters to be estimated while automatically accounting for parameter uncertainty. Furthermore, the mixture components in an HMM belong to distributional families chosen by the modeler, so HMMs provide a very flexible way to fit the data. We model the data examined in this article using mixtures of multivariate t distributions, each with its own covariance matrix. The HMM described in this article is a strict generalization of the mixture model implicit in the k-means clustering algorithm, which we refer to as the k-means model.

Both the k-means and HMM approaches assume that transitions over time are governed by a time-homogeneous Markov process, an assumption which may be violated if the effect of a treatment changes as the study progresses. To address this concern we develop an inhomogeneous hidden Markov model, i.e. one in which different transition probabilities may apply for each observation period. To prevent an explosion in the number of parameters we model the rows of each period's transition matrix as draws from a

common Dirichlet distribution with parameters embedded in a Bayesian hierarchical model. The transition matrices in our inhomogeneous model benefit from Bayesian shrinkage, so that if the data show no evidence of inhomogeneity the inhomogeneous model collapses back to the homogeneous model. Shrinkage factors for the inhomogeneous model can be used to check whether the homogeneity assumption is reasonable.

The purpose of this article is to demonstrate the HMM approach to health state modeling and evaluate its potential advantages and disadvantages relative to the clustering method. We have fit HMMs to data from a comprehensive double-blind trial that compared the impact of haloperidol and clozapine, two medications for treating schizophrenia, on clinical outcomes, social, vocational and community functioning and societal costs (Rosenheck *et al.*, 1997). This data set has already been studied using a cluster-based health state model, which will allow us to make direct comparisons between the HMM and cluster methods. In Section 2 we provide a description of the data. Details of both a homogeneous and an inhomogeneous hidden Markov health state model are provided in Section 3. Section 4 gives results from the HMM fit to the schizophrenia data set. Finally, Section 5 provides a discussion of the relative merits of the clustering and HMM approaches. Details of the MCMC algorithms used to fit the model are left to an appendix.

2 Data

The schizophrenia data set contains 423 patients treated at 15 veterans health centers around the United States. The measurements consist mainly of scores on standard health status instruments measuring a broad spectrum of emotional, interpersonal, and physical functioning. Our analysis focuses on movement disorders that are typically induced by antipsychotic medications. We combined items from three commonly used instruments, the Abnormal Involuntary Movement Scale (AIMS) which measures tardive dyskinesia, i.e. unconscious movements, (Guy, 1976); the Barnes Akathesia Scale (BAS), which focuses on involuntary restlessness (Barnes, 1989); and the Simpson-Angus Scale (SAS), which deals with syndromes of pseudo-parkinsonism such as involuntary tremors, muscle stiffness, and salivation (Simpson and Angus, 1970). All these instruments use Likert scales to measure severity of symptoms with higher scores indicating a greater degree of impairment. Data were collected by trained research assistants at six time-points (baseline, 6 weeks, and 3, 6, 9, and 12 months). There was evidence of significant differences in ratings among the 15 study sites. To make the responses comparable we subtracted off the site effects, which were estimated by fitting mixed effects models to each question using patient response as the dependent variable, with time, treatment, and study site as independent variables.

The side effects data were 24 dimensional. To reduce the dimension of the data and to allow comparisons with previous analyses (e.g. Sugar *et al.*, 2004) we replaced the full data set with its first four principal components. Principal components also smooth over roughness inherent in the Likert responses to individual items, making mixtures of continuous distributions more reasonable. The choice of four components was made on both quantitative and qualitative grounds. We opted to include all dimensions for which the proportion of variance explained was higher than the average variance per dimension. This procedure yielded a small number of easily interpretable dimensions. The components represent, in order of variance explained, overall severity (PC1), a contrast between akathesia and tardive dyskinesia (PC2), extrapyramidal syndromes, as measured by the SAS, excluding akathesia (PC3), and a contrast between facial and extremity movements (PC4). The four principal components explained approximately 60% of the total variance. Clustering based on principal components has the potential to obscure cluster distinctions (Chang, 1983; Raftery, 2003). However because of the obvious medical interpretations attached to the principal compo-

nents we believe that the benefits from dimension reduction are likely to outweigh the potential risks in our particular case.

Patients within each study site were randomized to receive clozapine or haloperidol. Haloperidol is a standard treatment, while clozapine is a relatively new so-called atypical antipsychotic which is thought to show promise for reducing medication induced movement disorders. Because such disorders are ubiquitous side effects of antipsychotic medication, studies of this sort typically involve many patients switching treatments. During the study 105 subjects (24.8%) switched from one treatment medication to the other. Furthermore, 157 patients (37.1%) switched from either clozapine or haloperidol to a non-conventional treatment or went off medication altogether. While addressing this problem is not a central feature of the current article, frequent treatment switching clearly has implications for any analysis of this type of data. To simplify comparisons with earlier studies, we adopt the convention used in Sugar *et al.* (2004) for modeling treatment switches. Subjects who crossed over were analyzed on an as-treated basis. Subjects who went off all medications or switched to a non-conventional treatment were analyzed on an intent to treat basis, meaning that they remained in the group to which they were originally assigned. We also examined the data using a pure as-treated analysis, with patients who switched off both treatments counted as a third group. This had minor effects on some of our numerical estimates, but not on our qualitative conclusions regarding the relative merits of the two medications.

Data were available for 80% of planned follow up observations. Missing data were modeled as ignorable (Little and Rubin, 1987) largely because the forward-backward recursions used to fit the models in Section 3 make it easy to analytically integrate out ignorable but temporally dependent missing data. Patients with missing data tended to lack complete questionnaires rather than individual item responses. Most of the 420 missing observation times are due to patients who left the study. However, there were 41 patients who were unobserved for a single observation but subsequently returned. Eleven patients were unobserved for gaps of two observations or longer.

3 Longitudinal Hidden Markov Models

The hidden Markov models defined in this section differ from typical HMMs in two primary respects. First, different transition matrices are used to model subjects observed under different treatments. Second, because multiple subjects are observed at each time point, it is possible to fit an inhomogeneous model in which different Markov transition probabilities apply at each observation time. Section 3.1 defines the homogeneous model. Section 3.2 defines the inhomogeneous model.

3.1 Time Homogeneous Hidden Markov Models

Let $\mathbf{y}_{it}$ be the vector of observed responses from subject i at time $t \in \{1,\ldots,T\}$, when subject i is under treatment $k_{it} \in \mathcal{K} = \{1,\ldots,K\}$. In our case study $\mathbf{y}_{it}$ is a four dimensional vector of principal components. Our model assumes that responses are conditionally independent given a hidden state variable $h_{it} \in \mathcal{S} = \{1,\ldots,S\}$. Hence,

$$p(\mathbf{y}_{it}|h_{it} = s, \cdot) = \mathcal{T}(\mathbf{y}_{it}|\mu_s, \Sigma_s, \nu_s), \tag{1}$$

where the raised dot $\cdot$ in a probability distribution represents all other known and unknown quantities, and μ_s, Σ_s and ν_s respectively represent the mean vector, the "scatter matrix," and the scalar degrees of freedom parameter for the multivariate t distribution describing state s. We used the parameterization of the

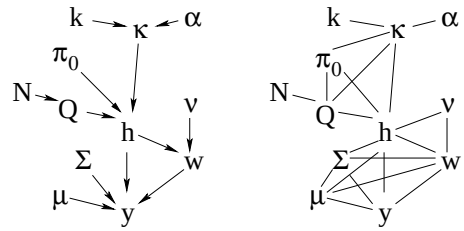

Figure 1: *Directed acyclic graph (left) and moral graph (right) describing the model. Each variable in the DAG is conditionally independent of its ancestors given its parents. Each variable in the moral graph is conditionally independent of all other variables given its neighbors. In the homogeneous model N is fixed. In the inhomogeneous model it is random. Other fixed hyperparameters are not shown.*

multivariate t distribution favored by Liu (1996), namely if $\mathbf{x}_{it} \sim \mathcal{N}(0,\Sigma)$, $w_{it} \sim Ga(\nu/2,\nu/2)$, and $\mathbf{x}_{it} \perp \perp w_{it}$, then $\mathbf{y}_{it} = (\mu + \mathbf{x}_{it}/\sqrt{w_{it}}) \sim \mathcal{T}(\mu,\Sigma,\nu)$. We opted to model responses using mixtures of multivariate t distributions instead of the more common mixtures of Gaussians because a small number of outlying observations had an undue impact on the variance matrices in Gaussian mixtures (see McLachlan and Peel, 2000, Chapter 7).

Subjects move through the state space according to a Markov chain with treatment dependent transition probabilities. The initial state distribution for subjects assigned to treatment k is $\pi_0^k(s) = p(h_{i1} = s|k_{i1} = k)$. Note that it is common in applications of HMMs to model the initial state distribution as the stationary distribution of the hidden Markov chain. This is true partly because most applications of HMMs involve a single long time series, but multiple subjects are needed to estimate π_0^k empirically. We model π_0^k as a separate parameter because we expect the distribution of subjects among states to evolve over time after treatments are administered. For patients who remain under treatment k from time $t-1$ to time t we define $Q^k(r,s) \equiv p(h_{it} = s|h_{it-1} = r, k_{it} = k_{it-1} = k)$. Transitions for subjects who switch treatments between observations $t-1$ and t are modeled using a mixture of the "pure" transition probabilities where the treatment proportions are the mixing weights. If $k_{it} \neq k_{it-1}$ and α_{it} is the (observed) proportion of time subject i spent under treatment k_{it} between observations $t-1$ and t then

$$q_{it}(r,s) \equiv p(h_{it} = s|h_{it-1} = r, k_{it}, k_{it-1}, \alpha_{it}) = \alpha_{it} Q^{k_{it}}(r,s) + (1-\alpha_{it}) Q^{k_{it-1}}(r,s). \tag{2}$$

It is computationally convenient to introduce a Bernoulli latent class indicator κ_{it} that decouples the mixture of transition probabilities in equation (2). That is, $p(\kappa_{it} = k_{it}|\mathbf{k},\alpha) = \alpha_{it}$, $p(\kappa_{it} = k_{it-1}|\mathbf{k},\alpha) = 1-\alpha_{it}$, and $p(h_{it}|h_{it-1},\kappa,\mathbf{k},\alpha) = Q^{\kappa_{it}}(h_{it-1},h_{it})$. In summary, the parameters of our model are $\theta = \{\mu_s, \Sigma_s, \nu_s, \pi_0^k, Q^k : s \in \mathcal{S}, k \in \mathcal{K}\}$. The observed data are $\mathbf{d}_{obs} = \{\mathbf{y}_{it}, k_{it}, \alpha_{it} : i = 1\ldots n, t = 1\ldots T\}$, and the latent data are $\mathbf{d}_{mis} = \{h_{it}, \kappa_{it}, w_{it} : i = 1\ldots n, t = 1\ldots T\}$. The relationships among these variables are illustrated in Figure 1.

We adopt a Bayesian approach to model fitting, which requires placing a prior distribution on model parameters. When available, standard independent conjugate priors are used. Specifically we use a Gaussian prior with mean m_s and variance Ω_s for μ_s, a Wishart prior with scalar degrees of freedom DF_s and sum of squares matrix SS_s for Σ_s^{-1} and Dirichlet priors with prior count vectors N_0^k and N_r^k for π_0^k and $Q^k(r,\cdot)$ (the rth row of Q^k) respectively. We model ν_s using the "uniform shrinkage prior" $p_0(\nu_s|z_{0s}) = z_{0s}/(z_{0s}+\nu_s)^2 I(\nu_s > 0)$ developed by Christiansen and Morris (1997). Note that p_0 is a normalized proper density function with median z_{0s}, but with no moments because of its heavy polynomial tail. Christiansen and Morris (1997) show

that p_0 has good frequency properties in a hierarchical Poisson regression model. It is relevant here because Christiansen and Morris's hierarchical model and the multivariate t used here are both defined through a latent gamma distribution with prior p_0 on its shape parameter. Hence the joint prior on θ is given by

$$p(\theta) = \left(\prod_{s \in \mathcal{S}} \mathcal{N}(\mu_s | m_s, \Omega_s) \mathcal{W}(\Sigma_s^{-1} | DF_s, SS_s) p_0(\nu_s | z_{0s}) \right) \prod_{k \in \mathcal{K}} \left(\mathcal{D}(\pi_0^k | N_0^k) \prod_{r \in \mathcal{S}} \mathcal{D}(Q^k(r,\cdot) | N_r^k) \right). \quad (3)$$

Equation (3) allows different hyperparameters for different treatments and mixture components, but in practice we choose identical priors for all k and s. Specifically we set $z_{0s} = 1$, $m_s = \mathbf{0}$, $\Omega_s = 1000I$, $DF_s = 6$, $SS_s = 6I$, and $N_r^k = N_0^k = \mathbf{1}$, where I is the identity matrix and $\mathbf{0}$ and $\mathbf{1}$ are vectors of 0's and 1's. These choices represent weak prior information while ensuring that the posterior distribution is proper.

3.2 A Hierarchical Inhomogeneous HMM

Because multiple subjects are present at each period, it is possible to estimate a different transition matrix for each pair of successive times using a hierarchical model that borrows strength across observations. Let Q_t^k be the matrix of transition probabilities for subjects under treatment k between observation times $t-1$ and t. If a subject switches treatments between $t-1$ and t then $Q_t^{k_{it-1}}$ and $Q_t^{k_{it}}$ are combined as in equation (2). We model $Q_2^k(r,\cdot), \ldots, Q_T^k(r,\cdot)$ as draws from a common Dirichlet distribution with parameter N_r^k, a vector of positive real numbers interpretable as prior counts. The joint prior for Q and N can be written

$$p(Q,N) = \prod_k \prod_r p(N_r^k) \prod_t \mathcal{D}(Q_t^k(r,\cdot) | N_r^k). \quad (4)$$

Equation (4) allows for Bayesian shrinkage across time, but elements of Q and N are independent across treatment and state indices. The hyperprior distribution $p(N_r^k)$ is defined by splitting $N_r^k = a_r^k \phi_r^k$ where a_r^k is a positive scalar controlling the variance of $\mathcal{D}(Q_t^k(r,\cdot)|N_r^k)$ and ϕ_r^k is a probability vector. That is, ϕ_r^k has elements $\phi_{rs}^k \in (0,1)$ with $\sum_s \phi_{rs}^k = 1$. We call a_r^k the *shrinkage parameter* and ϕ_r^k the *location parameter*. The full conditional distribution of $Q_t^k(r,\cdot)$ is $\mathcal{D}(N_r^k + n_t^k(r,\cdot))$, where $n_t^k(r,s)$ counts the number of transitions from state r to state s for treatment k between times $t-1$ and t. Thus one may interpret a_r^k as the number of prior observations present in the posterior distribution of $Q_t^k(r,\cdot)$. If a_r^k is large then $Q_2^k(r,\cdot), \ldots, Q_T^k(r,\cdot)$ will all be close to ϕ_r^k, in which case the model collapses back to the homogeneous form of Section 3.1. If a_r^k is close to zero then $Q_2^k(r,\cdot), \ldots, Q_T^k(r,\cdot)$ may vary substantially. We assume $p(a_r^k, \phi_r^k) = p_0(a_r^k|\zeta_0) p(\phi_r^k)$ where p_0 is the uniform shrinkage prior discussed in Section 3.1 and $p(\phi_r^k) = \mathcal{D}(\mathbf{1})$, the uniform prior on the S dimensional probability simplex. Transforming this prior back to the original scale introduces a Jacobian term of $(a_r^k)^{-(S-1)}$, so that the normalized prior distribution for N_r^k is

$$p(N_r^k) = \frac{\zeta_0 \Gamma(S)}{(\zeta_0 + a_r^k)^2 (a_r^k)^{S-1}}. \quad (5)$$

Small values of ζ_0 correspond to a prior belief in small amounts of shrinkage. We chose $\zeta_0 = 1$. Small changes in this value (e.g. $\zeta_0 = 2$) had no discernible effect on the posterior distribution.

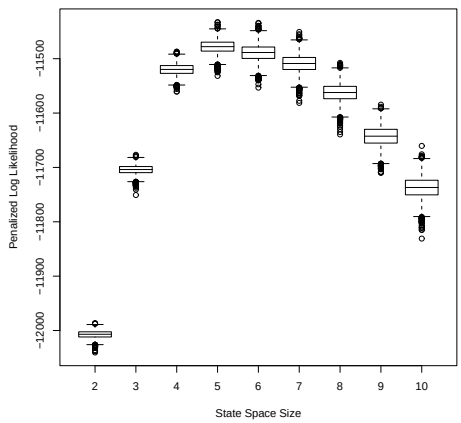

Figure 2: *Posterior distribution of log likelihood values produced by the MCMC sampler for models of different state space sizes after subtracting the BIC penalty* $k\log(n)/2$.

3.3 Posterior Computation

We sample the parameters of the models described in this Section from their posterior distribution given $\mathbf{d}_{mis}$ using an MCMC algorithm developed in Appendix B. The algorithm cycles between sampling from $p(\mathbf{d}_{mis}|\theta,\mathbf{d}_{obs})$ and sampling from $p(\theta|\mathbf{d}_{obs},\mathbf{d}_{mis})$. A key feature of our MCMC algorithm is a set of forward-backward recursions that allow $\mathbf{d}_{mis}$ to be drawn directly from $p(\mathbf{d}_{mis}|\theta,\mathbf{d}_{obs})$ without breaking it into multiple components (Scott, 2002). Conditioning on $\mathbf{d}_{mis}$ induces desirable independence properties in $p(\theta|\mathbf{d}_{obs},\mathbf{d}_{mis})$, so that the MCMC algorithm has only three components: $p(\mathbf{d}_{mis}|\mathbf{d}_{obs},\theta)$, $p(\mu,Q,\nu,\pi_0|\mathbf{d}_{mis},\mathbf{d}_{obs},N,\Sigma)$, and $p(\Sigma,N|\mathbf{d}_{mis},\mathbf{d}_{obs},\mu,Q,\nu,\pi_0)$. Each of these components further benefits from independence relationships which may be seen in the moral graph (Whittaker, 1990) shown in Figure 1. Gibbs updates are used for μ, Σ, Q, and π_0. Metropolis-Hastings updates are used for ν, and for N in the inhomogeneous model.

4 Case study

4.1 The Health States

The first task in developing the health state model is choosing S, the number of health states, based on empirical evidence and medical judgments. In essence, we are attempting to find a decomposition which provides a reasonable, medically interpretable, fit to the data. The natural Bayesian tool for choosing S is the posterior model probability, $p(S|\mathbf{d}_{obs})$. We implemented two methods for estimating this quantity, which is notoriously difficult to calculate. Chib's method (Chib, 1995; Chib and Jeliazkov, 2001) computes a direct Monte Carlo estimate of $p(S|\mathbf{d}_{obs})$ from the MCMC output. Alternatively, the Bayesian information criterion BIC, which applies a penalty to $\ell(\hat{\theta})$ the maximized log likelihood, can be used to obtain an asymptotic approximation to $p(S|\mathbf{d}_{obs})$ (Schwarz, 1978; Kass and Raftery, 1995). The BIC penalty is $k\log(n)/2$, where n is the number of observations and k is the number of free parameters in the model. Rather than maximize $\ell(\theta)$ we applied the BIC penalty to $\ell(\theta^{(t)})$, the sequence of log likelihood values associated with each MCMC draw of θ, producing the distribution of penalized likelihoods shown in the Figure 2. Note that we implement BIC on the homogeneous model because k, the effective number of parameters in the inhomogeneous model,

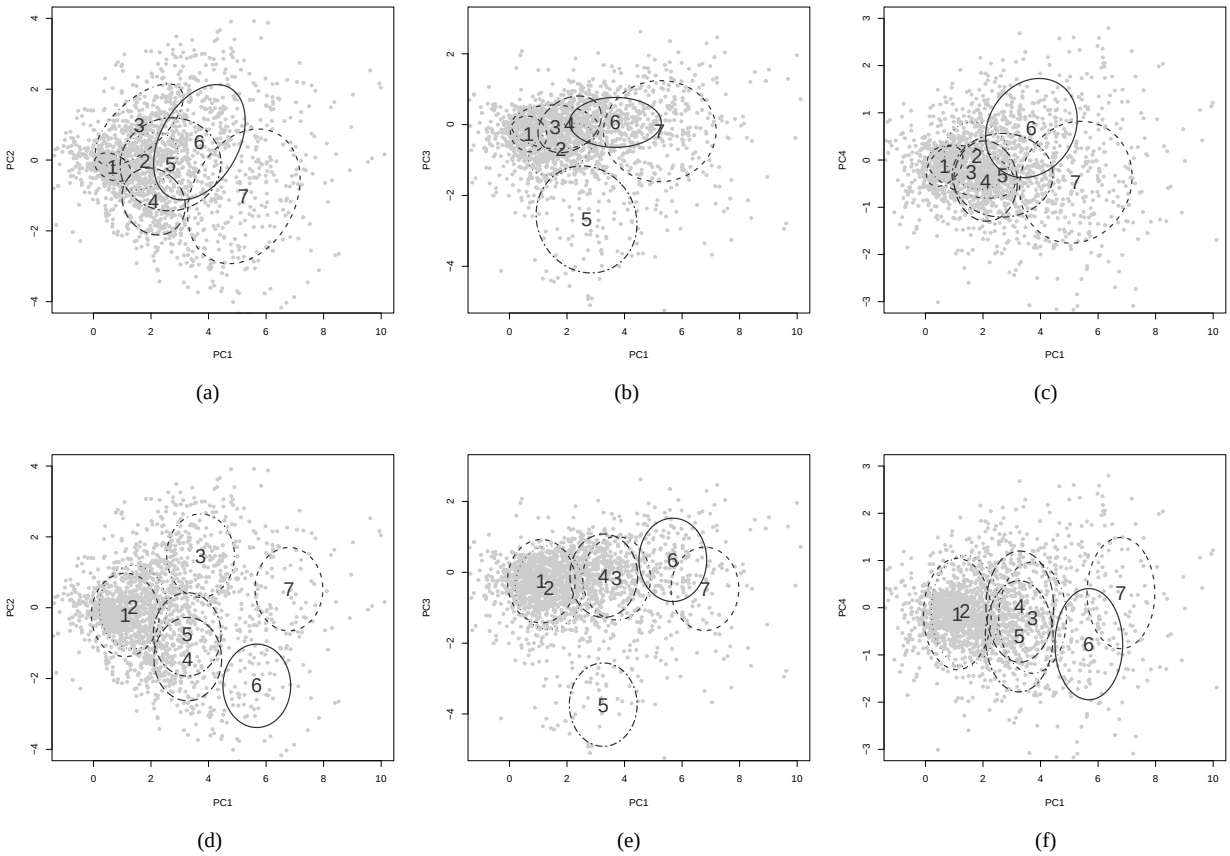

Figure 3: *Cluster centers and 50% probability ellipses for the seven state HMM (top row) and the k-means model (bottom row). Each panel excludes a small number of outliers.*

can not be determined exactly due to the Bayesian hierarchy (Spiegelhalter *et al.*, 2002). Chib's method suggested seven states while Figure 2 suggests four to seven states with a slight preference for five. We opted to fit the seven state model because it included a clinically distinct group that was absent from smaller models. In each case we ran the MCMC algorithm for 10,000 burn-in iterations, then we kept an additional 10,000 iterations. Models were initialized by setting all transition probabilities to $1/S$, setting all $\mu_s = \mathbf{0}$ and setting all Σ_s equal to large multiples of the identity. We checked convergence by monitoring time series plots of log likelihood for each model.

Figure 3 shows results from seven-state models fit using both our HMM and a finite mixture of Gaussian distributions with identity variance matrices, which serves as our proxy for the *k*-means procedure used by Sugar *et al.* (2004). Each panel of Figure 3 plots the posterior means of μ_s and Σ_s (represented by 50% probability ellipses) for each mixture component, along with the original data in the first four principal component dimensions. The "ellipses" for the *k*-means model would be circles if the axes in each plot were identically scaled. Note that state labels are arbitrary in all mixture models, including HMMs. Sometimes this can lead to a "label switching" phenomenon in the MCMC algorithm as the sampler jumps between $S!$ symmetric modes in the likelihood. Several authors have recently pointed out the danger of imposing artificial constraints on the parameters to create an identifiable likelihood function (Stephens, 2000; Celeux *et al.*, 2000; Frühwirth-Schnatter, 2001). We ran our algorithm with no such constraints, yet we saw no

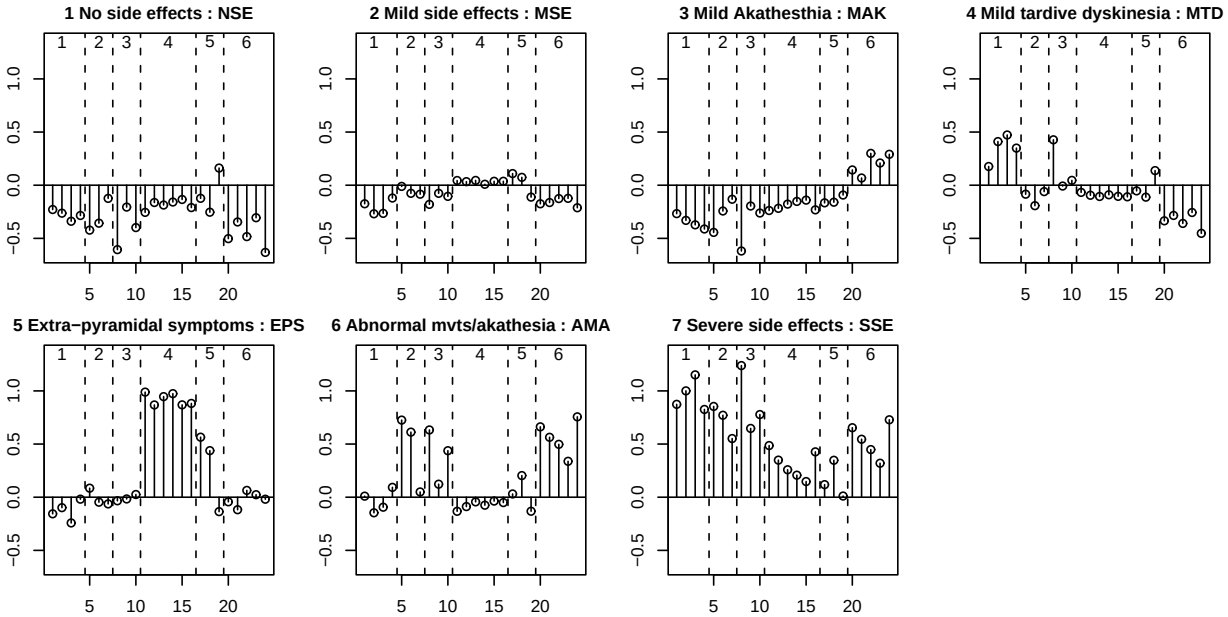

Figure 4: *The profile plots corresponding to a typical patient in each of the seven health states. The scores have been centered by subtracting off the global mean for each question. The six regions correspond to 1) facial/oral movements, 2) extremity and trunk movements, 3) global severity, 4) rigidity of gait, arms, head, 5) glabellar tap, tremor and salivation and 6) akathesia.*

evidence of label switching in the MCMC run for the seven state model, presumably because the $S!$ modes are well separated in the high dimensional parameter space. To check for label switching we examined S sets of boxplots representing the marginal posterior distributions of the four components of μ_s. In the presence of label switching one would expect to find a collection of states in which all four boxplots overlapped substantially. In our case, each state was clearly different from the others along at least one of the four dimensions. For descriptive purposes after the sampler finished we used PC1, a measure of a patient's overall distress, to construct a partially ordered labeling of the mixture components in which state 1 contains the healthiest patients and state 7 contains the patients with the most severe symptoms. PC2 contrasts akathesia (restlessness, positive scores) with tardive dyskinesia (involuntary movements, negative scores) and separates HMM states 3 and 4. A negative score on PC3 corresponds to extra-pyramidal symptoms such as problems with gait, rigidity, tremor, and salivation. State 5, which was absent from models with fewer than seven states, captures the observations with the most extreme values of PC3. The final principal component, PC4, is a contrast between facial movements and other movement difficulties. PC4 helps separate HMM state 6 from the other states. All the HMM states except 2 and 6 have posterior medians below 20 for ν_s, the t degrees of freedom parameter. States with small ν_s are capturing outliers that would otherwise be influential for μ_s and Σ_s in a Gaussian mixture (McLachlan and Peel, 2000). The ability to fit different variance matrices seems to help the HMM capture the triangular shape of the data. The more severe HMM states tend to have larger variances, while states 3 and 4 have rotated to capture observations along the edges of the plot. The HMM places a much smaller variance on state 1 than does the k-means model. Thus the HMM is more conservative than k-means about classifying observations into the healthiest state.

The information in Figure 3 is difficult to explain to clinicians because it is measured using principal components rather than the scale of the original 24 items. We can help clinicians interpret the health states by

using cluster profile plots that show the posterior mean response to each item for subjects in each state. Let x_{ijt} denote the observed response of subject i to item j at observation time t. Let $\pi_{it}(s)$ denote the probability, averaging over θ, that subject i is in state s at time t, which is available from the MCMC algorithm (Scott, 2002, Section 3). A cluster profile plot displays $\bar{x}_{js} = \sum_i \sum_t x_{ijt}\pi_{it}(s) / \sum_i \sum_t \pi_{it}(s)$. A medical doctor can examine cluster profile plots like those in Figure 4 and provide brief medical descriptions of each state. For instance, the typical patient in "no side effects" (NSE) has below average scores on all but one of the items, indicating relative health. The opposite can be said for the "severe side effects" (SSE) state. The "mild side effects" (MSE) state has a typical patient with average scores on most questions, but slightly higher scores on extra-pyramidal symptoms. The other four states each pick out the medical conditions "mild akathesia" (MAK), "mild tardive dyskinesia" (MTD), "extra-pyramidal symptoms" (EPS) and "abnormal movements/akathesia" (AMA). We feel confident using these states for our final model because each of the seven groups corresponds to a medically distinct health state. Otherwise we would have combined medically redundant states into larger clusters. Henceforth we will refer to the seven states by their three letter abbreviations. More detailed descriptions of the cluster profiles are provided in Appendix A.

4.2 Analysis of Longitudinal Treatment Effects

The preceding results are all from the homogeneous hidden Markov model, although the inhomogeneous model identified nearly identical health states. The inhomogeneous model allows one to measure the stability of the transition probabilities in the underlying Markov chain, which can be understood through the shrinkage parameters a_r^k. Figure 5(a) shows boxplots describing the marginal posterior distributions of $\log_{10} a_r^k$. The posterior medians of the shrinkage parameters are typically 100 or more for most states in both treatments. Recall that a_r^k represents the number of prior observations present in the full conditional distribution of $Q_t^k(r,\cdot)$, so the very large values of a_r^k indicate that the model has shrunk almost entirely back towards the homogeneous model. Bayesian shrinkage is typically measured in terms of shrinkage factors between 0 and 1 (Morris, 1983). Shrinkage factors for this model are defined as $B_r^k(t) = a_r^k/(a_r^k + n_t^k(r,+))$, where $n_t^k(r,+) = \sum_s n_t^k(r,s)$, the total number of transitions out of state r between $t-1$ and t for subjects on treatment k. Posterior medians of $B_r^k(t)$ are plotted in Figures 5(b) and (c). During the first transition the SSE state for clozapine had a posterior median shrinkage factor of .63, by far the lowest for either treatment. Most other transition probabilities had posterior median shrinkage factors above .8, with roughly half of the clozapine figures above .9. The consequence of such large shrinkage factors is that the transition probabilities $Q_t^k(r,s)$ are essentially the same for all t. The only substantial evidence of inhomogeneity is shown in Figure 5(d), which plots marginal posterior distributions for $Q_t^1(7,7)$, the probability that a clozapine subject in state SSE at time $t-1$ remains in SSE at time t. A low probability is medically desirable because it indicates that patients are likely to leave the worst state (SSE) for a better one. Figure 5(d) suggests that the first probability between baseline and six weeks was somewhat lower than the other periods, indicating that clozapine's effect on the sickest patients is felt immediately. The inhomogeneous effect is slight, but it was present for all choices of S that we considered (up to 10). By contrast the transition probabilities for haloperidol patients shown in Figure 5(e) appear to be homogeneous, as did all other sets of transition probabilities for both medications.

The high degree of shrinkage means that the inhomogeneous model is very close to the homogeneous model, with the possible exception of the first interval between observations. At first glance this is a somewhat surprising result given that the first two time intervals are half the length (6 weeks) of the other three

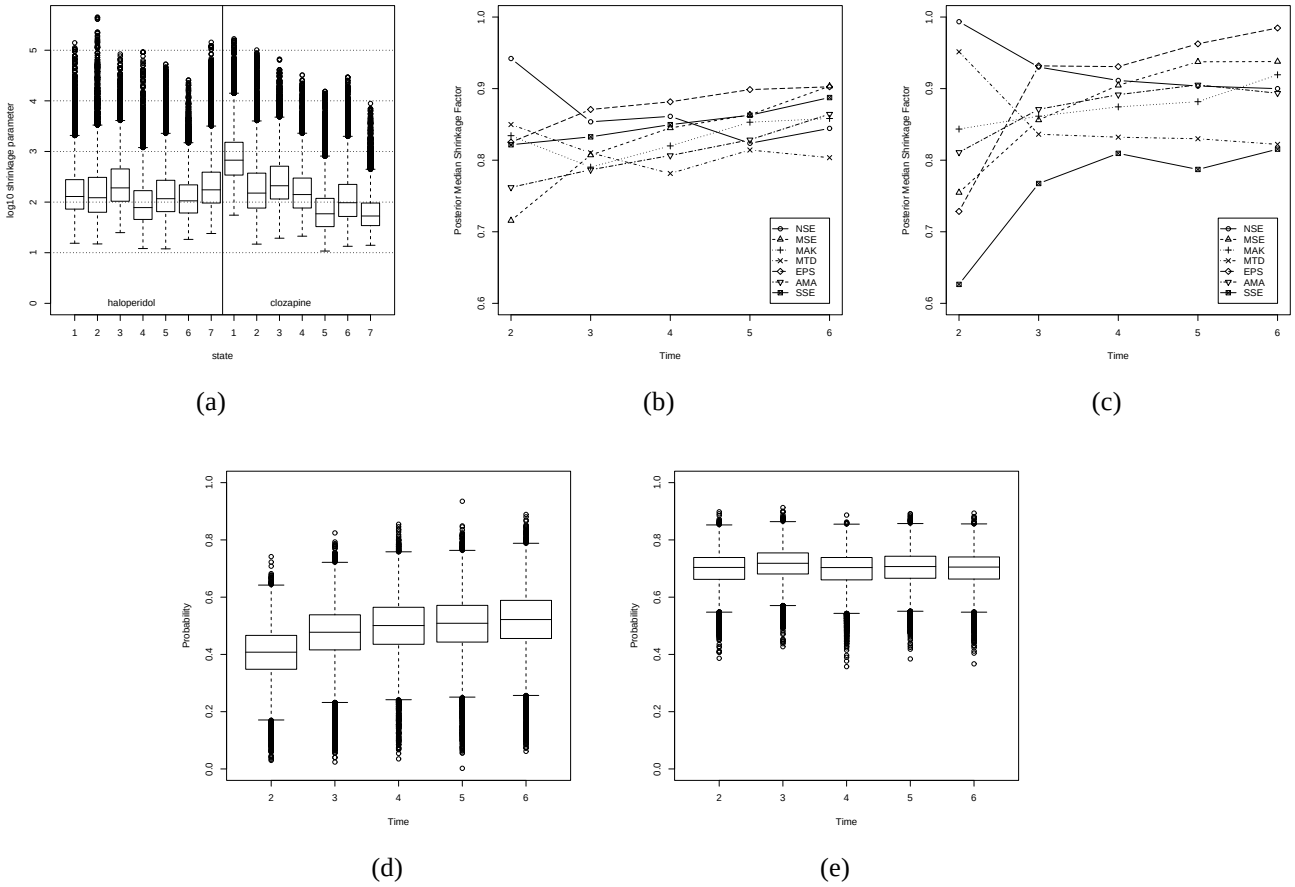

Figure 5: *Shrinkage in the inhomogeneous model. (a) Marginal posterior distributions of* $\log_{10}$ *shrinkage parameters* a_r^k*. (b) Posterior median shrinkage factors for haloperidol. (c) For clozapine. (d) The posterior distribution of* $Q_t^1(7,7)$*, the probability of a patient remaining in SSE at each of the five transition times for clozapine. (e)* $Q_t^0(7,7)$ *for haloperidol.*

intervals. However, there are medical reasons to expect more rapid transitions early in the study, which is why it was designed with early measurements at 6 week intervals. For example, the study period for a patient began when they were released from the hospital and one might expect this change in clinical care to have an important short term effect. It appears in this case that the shorter intervals between observations roughly offset the more rapid transitions to produce data consistent with a homogeneous model. This suggests that it would be inappropriate to account for the different durations between observations using a continuous time homogeneous HMM for these data. The remainder of this Section only considers the homogeneous model.

We can compare the effectiveness of clozapine versus haloperidol in terms of $\pi_t^k(s)$, the proportion of patients under treatment k in state s at time t. Figure 6 plots the posterior means of $\pi_t^k(s)$ for both medications at each of the six observation times. Figure 6 also plots the posterior mean of $\pi_\infty^k(s)$, the stationary distribution of Q^k for each medication. As one would hope, at baseline (0M) there is very little difference between the two medications. However, as early as the six week time point (6W) we note substantial differences. At six weeks, clozapine patients have approximately an 80% probability of belonging to one of the four best health states NSE, MSE, MAK and MTD and a 45% chance of falling in the two best states NSE

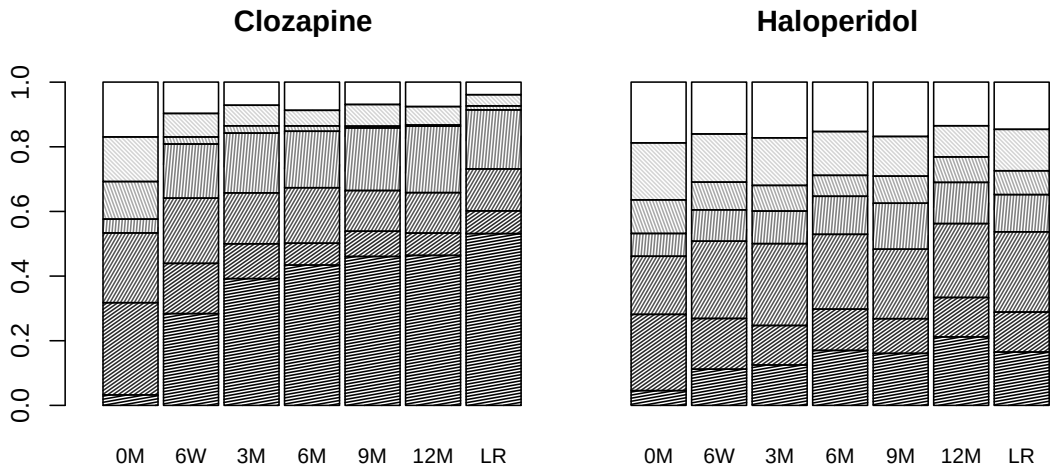

Figure 6: *Posterior expected proportions of patients in each health state for clozapine and haloperidol. The two groups of bars correspond to the six observed times plus the long run stationary distribution. The order of the states from bottom to top is NSE, MSE, MAK, MTD, EPS, AMA and SSE.*

or MSE. In comparison haloperidol patients have only 60% and 27% chances of falling in these groupings. The most dramatic change for clozapine patients is seen in the first six weeks. However the proportion of clozapine patients in NSE continues to climb, with the long run fraction greater than 50%. In comparison haloperidol patients experience relatively small gains. While there is a small increase over time in the proportion of haloperidol patients in the best health state NSE, the fraction in SSE remains fairly stable. This implies that the patients in the worst health states are not helped by haloperidol. Another dramatic difference between the medications is in EPS, which is essentially eliminated by clozapine but shows no improvement with haloperidol. The long run and 12 month distributions are similar for both haloperidol and clozapine, indicating that the patients appear to be close to stationarity after one year. The only state other than NSE whose proportion under clozapine grows over the course of the study was MTD.

Similar effects can be seen in the Markov transition probabilities displayed in Table 1. The transition probabilities reveal that a clozapine patient has a much higher probability of remaining in NSE than a haloperidol patient, a lower probability of remaining in any negative state except MTD, and a much lower probability of remaining in EPS. The probability of a clozapine patient transitioning into MTD is higher than that of a haloperidol patient for all states except NSE and MAK. Therefore clozapine does not induce MTD on healthy patients any more than does haloperidol, so MTD is not a side effect of clozapine in that sense. Rather, it appears to be a destination state for patients who fail to reach the more favorable state NSE.

Figure 7 captures the uncertainty about Figure 6 by plotting the marginal posterior distributions of $\pi_t^1(s) - \pi_t^0(s)$ for each state and observation time. For example, the plot for NSE indicates relatively similar proportions for both medications at baseline, the first box plot, but higher proportions of clozapine patients at six weeks. The difference in proportions increases at three months and then stays relatively stable over time. At three months and beyond there is a very high posterior probability that the proportion of clozapine patients in NSE is at least 20% greater than for haloperidol patients. Thus one can feel very confident that

From:	Clozapine							Haloperidol						
NSE	0.809	0.042	0.076	0.056	0.005	0.008	0.005	0.488	0.109	0.232	0.057	0.042	0.052	0.020
MSE	0.351	0.255	0.146	0.161	0.015	0.048	0.024	0.226	0.300	0.199	0.085	0.053	0.106	0.032
MAK	0.359	0.071	0.451	0.060	0.011	0.034	0.015	0.134	0.097	0.595	0.059	0.018	0.074	0.023
MTD	0.130	0.060	0.062	0.645	0.016	0.037	0.050	0.112	0.123	0.063	0.479	0.030	0.077	0.116
EPS	0.115	0.216	0.097	0.146	0.149	0.060	0.217	0.033	0.132	0.052	0.045	0.619	0.040	0.079
AMA	0.105	0.136	0.138	0.161	0.020	0.382	0.058	0.051	0.120	0.149	0.063	0.021	0.504	0.093
SSE	0.050	0.079	0.069	0.202	0.044	0.066	0.491	0.017	0.041	0.040	0.089	0.023	0.085	0.706
To:	NSE	MSE	MAK	MTD	EPS	AMA	SSE	NSE	MSE	MAK	MTD	EPS	AMA	SSE

Table 1: *Posterior means of transition probabilities for clozapine and haloperidol.*

clozapine is providing a genuine short term and long run overall improvement relative to haloperidol. Figure 7 also provides strong evidence of lower rates of clozapine patients in the MAK, EPS, AMA and SSE states, and similarly strong evidence of elevated levels in MTD. The differences do not appear as large for the other states, notably EPS, because they have fewer overall members than NSE. The differences in EPS would appear larger if they had been standardized by the total state size. Note that the differenced long run stationary distributions are population inferences based solely on the posterior distribution of Q^0 and Q^1. The other differences in proportions are in-sample inferences for the 423 subjects in our data set, which are less variable.

5 Discussion and Conclusions

In this article we used a hidden Markov model to analyze multivariate longitudinal data comparing the side effects experienced by patients with schizophrenia under two different medications. There was evidence that the population could be described by seven states of health ranging from no to severe side effects. Based on the fitted model we found very strong evidence that clozapine produces a larger and more rapid improvement in side effects than haloperidol, a standard antipsychotic treatment. There was also strong evidence of long term improvements with 60% of clozapine patients predicted to exhibit no or mild side effects compared to fewer than 30% for haloperidol. One of the advantages of a health state model over more standard univariate approaches is the ability to easily determine not just whether an overall improvement has occurred but also the types of improvement. For example we found evidence that clozapine was very effective at treating akathesia and extra-pyramidal symptoms, but less effective at treating tardive dyskinesia.

Our most compelling methodological advance is the hierarchical inhomogeneous model. As with other hierarchical models, the inhomogeneous HMM allows data to decide the extent of the compromise between fitting each period's transition probabilities independently and fitting a global transition matrix for the entire model. Longitudinal data are required to fit such a model, as multiple transitions need to be observed during each interval.

Our findings reinforce and extend those of Sugar *et al.* (2004) who analyzed the same data set using a health state model based on k-means clustering. Sugar *et al.* fit a six state model with states very close to those in the bottom row of Figure 3, but with states 1 and 2 merged. Section 4.1 notes the differences between the states fit by the HMM and the k-means model, but they are similar enough to see a rough correspondence. Sugar *et al.* observe longitudinal effects similar to our Figure 6. However, the HMM approach offers several advantages over the k-means approach. Cluster based methods involve assuming that

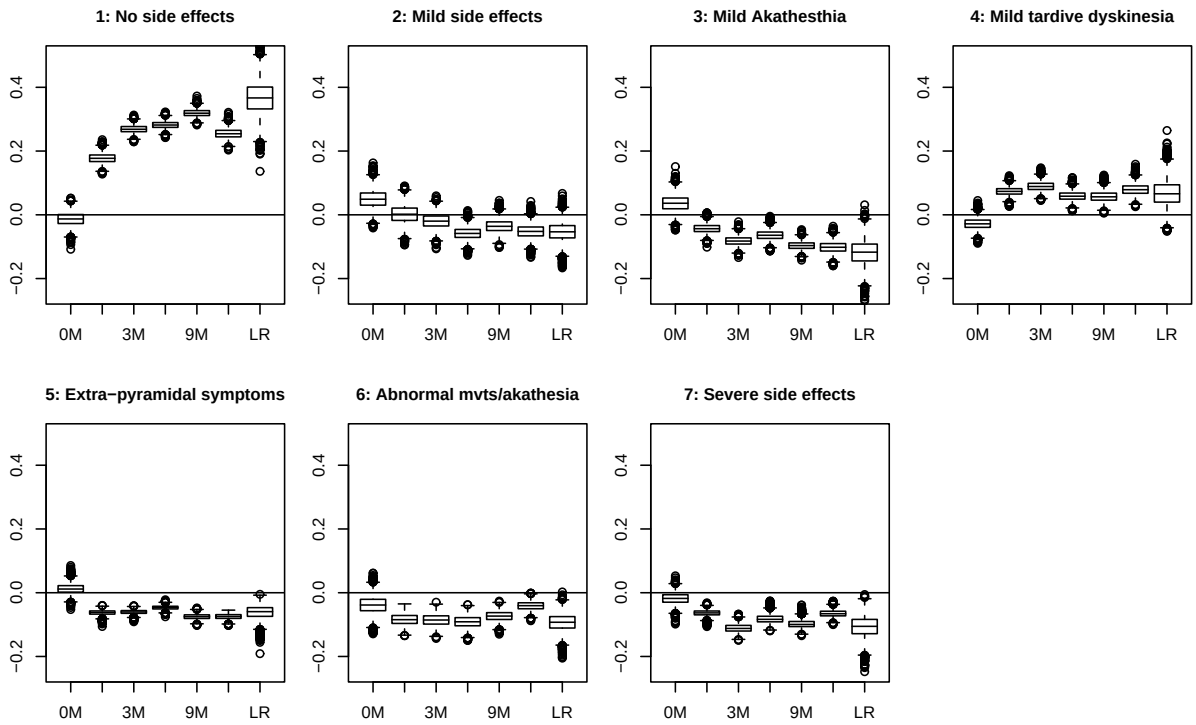

Figure 7: *Boxplots illustrating the posterior distributions of the difference between the proportions of clozapine patients and haloperidol patients in each state at each time.*

each observation's health state membership is known rather than estimated, introducing potential bias into the analysis. By contrast HMM parameters estimated using Bayesian methods automatically incorporate all sources of uncertainty, conditional on the model being correct. In addition Bayesian methods provide automatic measures of uncertainty even for complicated functions of the parameters like the differences between stationary distributions in Figure 7. As part of their model checking Sugar *et al.* performed a hypothesis test for inhomogeneity and found no evidence to reject a homogeneous model. However, through the use of shrinkage factors, our hierarchical model actually supports the discrete time homogeneous model rather than simply failing to provide evidence against it. Finally, by allowing varying covariance matrices and multivariate t distributions HMMs provide a more flexible fit to the data than the k-means procedure.

HMMs allow the health states and longitudinal effects to be simultaneously estimated borrowing strength from both. HMMs also allow the classification of an individual to a health state to depend on the state they belonged to in the previous time period. Moreover, uncertainty estimates for the HMM are not conditional on hard assignments of subjects to clusters. We see a few potential drawbacks to the HMM fit in this article, relative to the approach of Sugar *et al.* (2004). First, a potential issue in convincing people in health services research to adopt this approach is the way that health states are defined using HMM. Typically, medical doctors define health states via the boundaries of a partition of the space of patient characteristics. In the HMM a health state is a latent class, and people in different classes can have similar characteristics. From a medical perspective this approach involves treating the state as a genotype and the characteristics as a phenotype where the traditional approach would define the genotype via an individual's characteristics. While the HMM health state definition is reasonable, and in some ways more natural, there may be some resis-

tance to its adoption. Second, if the model were applied injudiciously it is possible that the larger number of parameters required to fit each state (relative to k-means) could lead to overfitting. Weak but proper priors for Σ_s centered on the identity matrix can help reduce this risk. Finally, the lack of widely available software for fitting HMMs has prevented their widespread adoption. However, examples of generally useful HMM code are beginning to appear in online libraries. We expect this to hasten the implementation of HMMs for health state modeling.

A Interpretation of health states

While it is not possible to form an exact ranking of the states in terms of severity of symptoms it appears that NSE represents the patents with fewest symptoms followed, in no clear order, by the patients in MSE, MAK and MTD. The patients in EPS and AMA are in yet worse shape and those patients in SSE exhibit the worst disorders.

No side effects (1: NSE) These patients are below average on almost all the side effects questions so they are relatively speaking in good shape.

Mild side effects (2: MSE) These patients are somewhat below average on the tardive dyskinesia and akathesia questions and slightly above average on items relating to extrapyramidal symptoms.

Mild akathesia (3: MAK) These patients have scores comparable to the NSE group on all questions except the akathesia scale where they are worse than average.

Mild tardive dyskinesia (4: MTD) These patients have average scores on the Simpson-Angus and below average scores on the akathesia questions. However, they have high scores on several of the AIMS questions corresponding to tardive dyskinesia.

Extra-pyramidal syndromes (5: EPS) These patients are close to average in every area except the first eight Simpson-Angus questions on which they are significantly worse than average. The Simpson-Angus Scale deals with syndromes of pseudo-parkinsonism, involuntary tremors and stiffness of muscles, and salivation.

Abnormal movements and akathesia (6: AMA) These patients have more severe akathesia problems than the MAK group and high scores on several of the AIMS questions corresponding to abnormal movements.

Severe side effects (7: SSE) These patients have well above average scores on almost all the questionnaire items and have significant side effects disorders.

B MCMC fitting procedures

This Section defines the MCMC algorithm used to sample from $p(\theta|\mathbf{d}_{obs})$ by alternately sampling from $p(\mathbf{d}_{mis}|\mathbf{d}_{obs},\theta)$ and $p(\theta|\mathbf{d}_{obs},\mathbf{d}_{mis})$. We sample the latent data from $p(\mathbf{h},\mathbf{w},\kappa|\theta,\mathbf{d}_{obs})$ directly, without breaking it into multiple MCMC components. The forward-backward recursions for HMMs (see Scott, 2002, for a review) are employed to draw $\mathbf{h}$ from $p(\mathbf{h}|\theta,\mathbf{d}_{obs})$, averaging over $(\mathbf{w},\kappa)$. Then $(\mathbf{w},\kappa)$ are drawn from

$p(\mathbf{w},\kappa|\theta,\mathbf{d}_{obs},\mathbf{h}) = p(\mathbf{w}|\theta,\mathbf{d}_{obs},\mathbf{h})p(\kappa|\theta,\mathbf{d}_{obs},\mathbf{h})$. Sampling $(\mathbf{w},\kappa)$ is particularly easy because the elements of $\mathbf{w}$ are independent in $p(\mathbf{w}|\theta,\mathbf{d}_{obs},\mathbf{h})$, as are the elements of κ in $p(\kappa|\theta,\mathbf{d}_{obs},\mathbf{h})$. The full conditional distribution for w_{it} is $p(w_{it}|h_{it}=s,\cdot) = Ga((\nu_s+p)/2,(\nu_s+\Delta_{it})/2)$ where p is the dimension of $\mathbf{y}_{it}$ and $\Delta_{it} = (\mathbf{y}_{it}-\mu_s)^T\Sigma_s^{-1}(\mathbf{y}_{it}-\mu_s)$. The full conditional distribution for κ_{it} concentrates on k_{it} and k_{it-1} with $p(\kappa_{it}=k_{it}|\cdot) = p_{1it}/(p_{1it}+p_{2it})$ where $p_{1it} = \alpha_{it}q^{k_{it}}(h_{it-1},h_{it})$ and $p_{2it} = (1-\alpha_{it})q^{k_{it-1}}(h_{it-1},h_{it})$. Most of the time $\alpha_{it}=1$ in which case $\kappa_{it}=k_{it}$ with probability 1.

For the homogeneous model, all parameters except ν_s have closed form full conditional distributions which are independent across s. We update these parameters using Gibbs draws from their full conditional distributions. Let $w_s^+ = \sum_{it} w_{it}I(h_{it}=s)$ and $y_s^+ = \sum_{it} w_{it}\mathbf{y}_{it}I(h_{it}=s)$. Then $p(\mu_s|\cdot) = \mathcal{N}\left\{A^{-1}\left(\Omega_s^{-1}m_s+\Sigma^{-1}y_s^+\right),A^{-1}\right\}$ with $A = (\Omega_s^{-1}+w_s^+\Sigma^{-1})$. Let $V_s^+ = \sum_{it} w_{it}(\mathbf{y}_{it}-\mu_s)(\mathbf{y}_{it}-\mu_s)^T I(h_{it}=s)$ and $n_s = \sum_{it} I(h_{it}=s)$. The full conditional for Σ_s^{-1} is $p(\Sigma_s^{-1}|\cdot) = \mathcal{W}(DF_s+n_s,SS_s+V_s^+)$. The rows of Q^k are independent across states and treatments with $p(Q^k(r,\cdot)|\cdot) = \mathcal{D}(N_r^k+n^k(r,\cdot))$. Similarly, $p(\pi_0^k|\cdot) = \mathcal{D}(N_0^k+\mathbf{n}^k)$ where $\mathbf{n}^k$ is a vector with elements $n_{rs}^k = \sum_i I(h_{i1}=r)I(k_{i1}=k)$.

We employ the Metropolis Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970; Chib and Greenberg, 1995) to sample ν_s from $p(\nu_s|\cdot)$ using proposals based on an approximation to $p(\log\nu_s|\cdot)$. Let m and v denote the mean and variance of the asymptotic normal approximation to $p(\log\nu_s|\cdot)$ derived in Appendix C.1. We propose deviates from $\log\nu_s^* \sim f(\log\nu_s^*|\cdot) = \mathcal{T}(m,v,3)$ and accepts them according to a standard Hastings probability. The t distribution with 3 degrees of freedom provides a proposal distribution with heavier tails than the target distribution, which helps prevent the sampler from becoming trapped in low probability regions (Mengersen and Tweedie, 1996).

The sampling algorithm for the inhomogeneous model differs from the homogeneous model in three respects, two of which require only trivial modifications. First, Q_t^k replaces Q^k when constructing q_{it} in the forward-backward recursions. Second, the inhomogeneous model has more transition probabilities which must be sampled from their full conditional distributions (given in Section 3.2). Third, an MCMC component must be added to sample from $p(N_r^k|\cdot)$.

Recall $a_r^k = \sum_s N_{rs}^k$ and $\phi_r^k = N_r^k/a_r^k$. Let $\delta_{r1}^k = \log a_r^k$, let $\delta_{rs}^k = \log(N_{rs}^k/N_{r1}^k)$ for $s>1$, and let $\delta_r^k = (\delta_{rs}^k)$. There is a one-to-one correspondence between N_r^k and δ_r^k with $\phi_{rs}^k = [I_{s1}+(1-I_{s1})\exp(\delta_{rs}^k)]/(1+\sum_{s'=2}^S \exp(\delta_{rs'}^k))$. Let M and V denote the mean vector and variance matrix of the multivariate normal approximation to $p(\delta_r^k|\cdot)$ developed in Appendix C.2. A proposal deviate is generated as $(\delta_r^k)^* \sim \mathcal{T}(M,V,3)$. The deviate is either promoted according to a Metropolis-Hastings probability or else N_r^k remains unchanged during the current iteration.

Care must be taken when sampling (Q,N) because $Q_t^k(r,s)=0$ is an absorbing state. That is, there can be pairs of states within a treatment for which $n_t^k(r,s)=0$ for some t in a given iteration of the sampler. The zero count leads to a draw of $Q_t^k(r,s)\approx 0$, which is a problem because the sufficient statistic for $p(N_r^k|\cdot)$ is the geometric mean of $Q_2^k(r,\cdot),\ldots,Q_T^k(r,\cdot)$. For any t, $Q_t^k(r,s)=0$ has infinite weight in the geometric mean, and thus in $p(N_r^k|\cdot)$. Thus $Q_t^k(r,s)\approx 0$ forces $N_{rs}^k\to 0$, which increases the probability mass near zero for all $Q_2^k(r,s),\ldots,Q_T^k(r,s)$ and exacerbates the problem on the next iteration. The absorbing state can be eliminated by truncating the support of $p(N_r^k)$ to enforce $N_{rs}^k > N_0$ for all s. In practice, truncating the support of the prior equates to simply rejecting Hastings proposals with any $N_{rs}^k \le N_0$. We set $N_0=1$ so that the "worst case" prior for $Q_t^k(r,\cdot)$ is the uniform prior.

C Approximations

C.1 Approximating $p(\log \nu_s|\cdot)$

Suppose $w_1,\ldots,w_n \overset{iid}{\sim} Ga(\nu/2,\nu/2)$ and $p(\nu) = z_0/(z_0+\nu)^2$. Let $w^+ = \sum_i w_i$, $u = \sum_i \log w_i$, and $\eta = \log\nu$. The log posterior density is

$$\log p(\eta|w) = K + \frac{n\nu}{2}\log(\nu/2) - n\log\Gamma(\nu/2) + (\nu/2-1)u - w^+\nu/2 + \log z_0 - 2\log(z_0+\nu) + \log\nu$$

where K is a normalizing constant, and the extra $\log\nu$ at the end is the log of the Jacobian $d\nu/d\eta = \nu$. Standard asymptotic theory (e.g. Le Cam and Yang, 2000) implies $p(\eta|w) \approx \mathcal{N}(\hat\eta, -1/h(\hat\eta))$, where $\hat\eta = \arg\max \log p(\eta|w)$ and $h(\eta) = \partial^2 \log p(\eta|w)/\partial\eta^2$. Basing the approximation on η rather than ν speeds convergence to normality. The derivatives of $\log p(\eta|w)$ with respect to η, which are useful in obtaining $\hat\eta$, are most easily computed using the chain rule. Let

$$g^* = \frac{\partial \log p(\eta|w)}{\partial\nu} = \frac{n}{2}\left[\log(\nu/2)+1-\psi(\nu/2)+\frac{u-w^+}{n}\right] - \frac{2}{z_0+\nu} + 1/\nu$$
$$h^* = \frac{\partial^2 \log p(\eta|w)}{\partial\nu^2} = \frac{n}{2}\left[\frac{1}{\nu}-\frac{1}{2}\psi'(\nu/2)\right] + \frac{2}{(z_0+\nu)^2} - 1/\nu^2.$$

Then by the chain rule $\partial\log p(\eta|w)/\partial\eta = g^*\nu$ and $h = (h^*\nu + g^*)\nu$.

C.2 The Posterior Distribution of Dirichlet Parameters

Suppose $\mathbf{q} = (q_1,\ldots,q_n)$ with $q_i \overset{iid}{\sim} \mathcal{D}(N)$ where q_i is an S dimensional probability vector $q_{is}, s \in \mathcal{S}$, and N is an $S-$vector of positive real elements N_s interpretable as counts. Let $a = \sum_s N_s$, and $\phi_s = N_s/a$. Define $\delta_1 = \log a$, $\delta_s = \log(\phi_s/\phi_1)$ for $s > 1$, and let $\delta = (\delta_1,\ldots,\delta_S)$. The Dirichlet likelihood function is

$$p(\mathbf{q}|\delta) = \prod_{i=1}^n \mathcal{D}(q_i|N) = \Gamma^n(a)\prod_{s=1}^S \frac{T_s^{N_s-1}}{\Gamma^n(N_s)}$$

where $T_s = \prod_{i=1}^n q_{is}$ are sufficient statistics. The prior distribution for δ is $p(\delta) = p(N(\delta))|\det J|$. The term $|\det J|$ is the absolute value of the determinant of the Jacobian matrix $\partial N/\partial\delta$, with elements

$$J_{rs} = \frac{\partial N_s}{\partial\delta_r} = N_s[I_{r1} + (1-I_{r1})(I_{rs}-\phi_r)] = N_s K_{rs}(\phi_r), \tag{6}$$

where $I_{rs} = 1$ if $r = s$, and $I_{rs} = 0$ otherwise. Notice that $K_{rs}(\phi_r)$ is a linear function of ϕ_r. One may write $\log p(\delta|\mathbf{q}) = C + \log p(\mathbf{q}|N(\delta)) + \log p(N(\delta)) + \log|\det J|$ where C is a normalizing constant. As $n\to\infty$, $p(\delta|\mathbf{q}) \to \mathcal{N}(\hat\delta, -H^{-1})$ where $\hat\delta = \arg\max\log p(\delta|\mathbf{q})$ and $H = \partial^2\log p(\delta|\mathbf{q})/\partial\delta\partial\delta^T$. Derivatives of $\log p(\delta|\mathbf{q})$ are useful for obtaining $\hat\delta$. It is easiest to differentiate the first two terms with respect to N, then transform the derivatives using the chain rule. Let $f(N) = \log p(\mathbf{q}|N) + \log p(N)$, $g^* = \partial f/\partial N$, and $H^* = \partial^2 f/\partial N\partial N^T$. Then $g = \partial f/\partial\delta = Jg^*$. The Hessian matrix with respect to δ can be computed from H^*, g^*, J, and the second order Jacobian $J^{(2)}$, a triply indexed array with elements

$$J^{(2)}_{rsm} = \frac{\partial^2 N_m}{\partial\delta_r\partial\delta_s} = \frac{\partial J_{sm}}{\partial\delta_r}. \tag{7}$$

Substituting (6) into (7) yields $J^{(2)}_{rsm} = J_{rs}K_{sm}(\phi_m) + N_s K_{sm}(D_{rm})$, where $D_{rm} = \partial\phi_m/\partial\delta_r = \phi_r(1-I_{r1})(I_{rm} - \phi_m) = \phi_r L_{rm}(\phi_m)$. The Hessian matrix with respect to δ is

$$H^0 = \frac{\partial f}{\partial\delta\partial\delta^T} = JH^*J^T + J^{(2)} \cdot g^*$$

where $J^{(2)} \cdot g^*$ is a matrix whose (r,s) element is $\sum_{m=1}^{S} J^{(2)}_{rsm} g^*_m$. The final set of derivatives involve differentiating $|\det J|$ with respect to δ, which can be accomplished according to formulas given by Harville (1997), Section 15.9, equation 9.3. The formula for computing the Hessian of $|\det J|$ requires the third order Jacobian

$$J^{(3)}_{irsm} = \frac{\partial J_{rsm}}{\partial\delta_i} = J^{(2)}_{irs}K_{sm}(\phi_m) + J_{rs}K_{sm}(D_{im}) + J_{is}K_{sm}(D_{rm}) + N_s K_{sm}(D^{(2)}_{irm}) \tag{8}$$

where $D^{(2)}_{irm} = \partial D_{rm}/\partial\delta_i = D_{ir}L_{rm}(\phi_m) + \phi_r L_{rm}(D_{im})$.

References

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Barnes, T. R. E. (1989). A rating scale for drug-induced akathesia. *British Journal of Psychiatry* **154**, 672–676.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture prior distributions. *Journal of the American Statistical Association* **95**, 451, 957–970.

Chang, W.-C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics* **32**, 267–275.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician* **49**, 327–335.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-hastings output. *Journal of the American Statistical Association* **96**, 453, 270–281.

Christiansen, C. L. and Morris, C. N. (1997). Hierarchical Poisson regression modeling. *Journal of the American Statistical Association* **92**, 618–632.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association* **96**, 453, 194–209.

Guy, W. (1976). *Abnormal involuntary movements. In: Guy W, ed. ECDEU assessment manual for psychopharmacology*. Rockville, Md.: National Institute of Mental Health (DHEW publication no. ADM 76-338.).

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

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

Le Cam, L. M. and Yang, G. L. (2000). *Asymptotics in Statistics: Some Basic Concepts*. Springer-Verlag.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis With Missing Data*. John Wiley & Sons.

Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data. *Journal of the American Statistical Association* **91**, 1219–1227.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, New York.

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics* **24**, 101–121.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications (c/r: P55-65). *Journal of the American Statistical Association* **78**, 47–55.

Raftery, A. E. (2003). Discussion of 'Bayesian clustering with variable and transformation selections'. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds., *Bayesian Statistics 7*, 266–270. Oxford University Press.

Rosenheck, R., Cramer, J., Xu, W., Thomas, J., Henderson, W., Frisman., L., Fye, C., and Charney, D. (1997). A comparison of clozapine and haloperidol in hospitalized patients with refractory schizophrenia. *New England Journal of Medicine* **337**, 809–815.

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

Scott, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association* **97**, 337–351.

Simpson, G. M. and Angus, J. W. S. (1970). A rating scale for extrapyramidal side effects. *Acta Psychiatr Scand Suppl* **212**, 11–19.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (Pkg: p583-639). *Journal of the Royal Statistical Society, Series B, Methodological* **64**, 4, 583–616.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B, Methodological* **62**, 795–810.

Sugar, C. A., James, G. M., Lenert, L. A., and Rosenheck, R. (2004). Discrete state analysis for interpretation of data from clinical trials. *Medical Care* **42**, 183–196.

Sugar, C. A., Sturm, R., Sherbourne, C., Lee, T., Olshen, R., Wells, K., and Lenert, L. (1998). Empirically defined health states for depression from the SF-12. *Health Services Research* **33**, 911–928.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.