Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

INFERENCE ON RANDOM COEFFICIENT MODELS FOR HAPLOTYPE EFFECTS IN DYNAMIC MUTATIONS USING MCMC.

Richard M. Huggins[1], Guoqi Qian[1] & Danuta Z. Loesch[2]
[1]Department of Statistical Sciences, La Trobe University, Bundoora, 3083, Australia.
[2]School of Psychological Science, La Trobe University, Bundoora, 3083, Australia.

Abstract

Modern genetic research involves complex stochastic processes and difficult inference problems and research in this area is necessarily a collaboration with geneticists and biologists. The major difficulty is in defining stochastic processes which are biologically meaningful yet amenable to analysis. To illustrate this we examine a class of random effects models for dynamic mutations. Dynamic mutations characterize several inherited disorders in humans. In these disorders a mutated segment of the gene typically increases in size as it is transmitted from generation to generation until the gene fails. Biological interest is in the effect of various genetic markers on the rate of expansion of the mutated segment and to what extent these markers describe alternate pathways. We concentrate on the widely studied fragile X disorder as there are data sets available for statistical analysis. We use heirarchical Bayes models fitted via MCMC methods to examine some data and hence determine a class of random coefficients, branching, time series models which have applications in genetical research.

1 Introduction

The realistic modelling of the stochastic processes occurring in biology quickly leads to analytically intractable models. However, these models may be of enormous practical importance. Here we examine a model for the recently discovered phenomenon of dynamic mutations, that lead to a number of genetic disorders. In these disorders rather than an on-off allele being transmitted from a parent to their offspring the mutation itself changes upon transmission. The mutation arises through expansion at a region of the gene, i.e. the insertion of extra genetic material in the form of trinucleotide

repeats at that site. This induces instability at that site with the size of the expansion usually increasing from generation to generation. Ultimately the expansion becomes large enough to affect the functioning of the gene resulting in phenotypic expression of the disorder. A feature of these dynamic mutations is that they are clinically undetectable until the expansion crosses a threshold and phenotypically abnormal individuals arise.

The fragile X Syndrome is the most widely studied dynamic mutation. It is a common X-linked genetic disorder associated with intellectual disability. The fragile X syndrome arises as the result of progressive CGG trinucleotide repeat expansion in the FMR1 gene, which culminates in the failure of gene expression. The mechanism for the expansion of the repeat size has not been determined. The fragile X syndrome is classified according to the number of CGG repeats. The premutation (55-200 repeats) and full mutation (200+ repeats) categories are unstable and some instability has also been reported within grey zone (35-55 repeats) alleles. Previously the rate of expansion of CGG repeats on transmission has usually been modelled as a function of the size of CGG repeat sequence on the maternal X chromosome. However, some molecular characteristics of the gene other than the size of the CGG repeat in this chromosome have recently been shown to affect the rate of expansion. These genetic covariates are certain combinations of microsatellite markers (haplotypes) flanking the repeat sequence, and the number and position of the AGG triplets which normally interrupt CGG repeats at intervals of 9-10 units. Eichler et al. (1996) have recently hypothesised that at least two different mutational pathways leading to the fragile X syndrome related to those genetic covariates are in action in the population. However, the data obtained by Dana et al (2000) from an Afro-American population showed that this hypothesis should be viewed from a broader evolutionary perspective.

Several mathematical models for the transmission of fragile X have been proposed. These include comprehensive transmission models (Ashley and Sherman 1995, Morton & Macpherson 1992, Morris et al 1995), iterated branching models (Gawel and Kimmel 1996) and Bernoulli counting process models (Bat et al 1997). The model of Ashley and Sherman (1995) accounts for the dynamics of fragile X mutation by assuming the multi-step expansion of CGG repeats (meiotic expansion in either sex, and mitotic expansion restricted to somatic alleles of maternal origin). They also included population assumptions, such as selection against full mutation, as well as molecular mechanisms, such as the loss of AGG interspersions, which in their model constitutes the initial mutation in the FMR1 gene. These previous models were biologically meaningful but were difficult to analyse statistically and tended to become outdated as more data on molecular characteristics of the gene came to hand.

Huggins, Loesch & Sherman (1998) introduced a multi-step non-linear time series models for the transmission of dynamic mutations which could be analysed using standard statistical methods. They used a random intercepts model to examine the effects of haplotypes or genetic markers on the transmission of the mutation. Here we extend that model in several directions, with an emphasis on the effect of the genetic markers on the rate of transmission. Firstly, we use cubic splines to model the relationship between the length of the repeat sequence in offspring X chromosomes to that in the parent X chromosome. Secondly we allow the rate of transmission to be random, rather than just a random intercept.

Available data consists of observations on several generations of affected families as, due to the lengths of human generations, it is not possible to collect DNA to construct longer chains. However, one purpose of our modelling is to be able to model distantly related affected individuals who are known to have a common ancestor. We illustrate the inference method using published data of Murray et al (1997) who report a variety of haplotypes which may be related to the rate of expansion in fragile X. The biological meaning of our results shall be discussed elsewhere, as will formal tests of hypotheses.

2 Notation

We label the observed haplotypes by $h = 1, 2, \ldots, H$. We suppose there are N_h families with haplotype h, labelled hf, $f = 1, \ldots, N_h$. Let n_{hf} denote the number of parents that produce observed offspring in family hf, labelled hfk, $k = 1, \ldots, n_{hf}$. Also, let w_{hfk} be the number of offspring of individual hfk. The offspring of individual hfk are denoted by $hfkl$, $l = 1, \ldots, w_{hfk}$. Note that individuals may be labeled both as parents and as offspring of their parents. This notation aids in constructing the conditional densities required below.

In this paper, inference is conducted conditional on the initiating individual in each family, the observed family structure and the observed haplotypes. Further, the probands, or the individuals initially detected with the disorder in each family, have been omitted. This was done on order to reduce the ascertainment bias as the probands were usually detected by their having an extreme phenotype and hence genotype. Let $\mathbf{Y}$ be the vector of observations of the CGG triplet repeat lengths on all offspring. Namely, $\mathbf{Y} = \{Y_{hfkl} : (h, f, k, l) \in S\}$, where S is defined by $S = \{(h, f, k, l) : h = 1, 2, \cdots, H;\ f|h = 1, 2, \cdots, N_h;\ k|hf = 1, 2, \cdots, n_{hf};\ l|hfk = 1, 2, \cdots, w_{hfk}.\}$. Here $f|h$ is understood as the value of f given h.

3 The Hierarchical Model

Our full model for the length of the CGG triplet repeat sequence in offspring l of individual k is of the general form

$$g(Y_{hfkl}) = X_{hfk}\{\beta + (\beta_h + \beta_f^{(h)})I_F(hfk)\} + \delta_F I_F \zeta_{hfkl} + \sigma\varepsilon_{hfkl}, \qquad (3.1)$$

$h = 1,\ldots,H$, $f|h = 1,\ldots,N_h$, $k|hf = 1,\ldots,n_{hf}$, $l|hfk = 1,\ldots,w_{hfk}$, where the ζ_{hfkl}'s and ε_{hfkl}'s are independent normal random variables with 0 means and variance 1, X_{hfk} is a design matrix and $I_F(hfk)$ takes the value 1 if individual hfk is female and zero otherwise. The response $g(Y_{hfkl})$ is taken to be $\log\log Y_{hfkl}$ in this paper as this gave errors that appeared normally distributed, although other choices are possible. We suppose β_h follows a common probability distribution for all individuals with haplotype h and $\beta_f^{(h)}$ follows a common distribution for all individuals in family f within haplotype h. The introduction of the sex effect allows us to mimic the Ashley-Sherman (1995) model and the subsequent model of Huggins et al (1998) which were two step models with an initial step common to transmissions from males and females, and a second step that only occurs in transmissions from females. This is difficult to directly model in a linear framework so instead we allow different rates in transmissions from males and females. Available data, see the plots of Huggins & Loesch (1998), suggest there is little difference between the haplotypes or families in transmissions from males and a single vector β is used to model this transition.

The conditional mean $X_{hfk}\{\beta + (\beta_h + \beta_f^{(h)})I_F(hfk)\}$ of $g(Y_{hfkl})$ as a function of the parameters and $\log Y_{hfk}$ was modelled using cubic splines (Dierckx 1993). The corresponding B-splines are contained in X_{hfk}. As the present work is exploratory ten degrees of freedom were used in the cubic splines to allow flexible modelling. This gave similar results to models with different degrees of freedom (6, 8 and 14). Let m denote the number of columns of X_{hfk}. The main interest is in the posterior distributions of β_h and $\beta_f^{(h)}$.

We will use the following prior distributions for the parameters $\beta, \beta_h, \beta_f^{(h)}, \sigma^2$ and δ_F^2:

1. $\beta \sim b_3(\beta|\beta_0, \Sigma_0) = DMVN(\beta_0, \Sigma_0)$,
2. $\beta_h \sim b_2(\beta_h|\Sigma_{h0}) = DMVN(0, \Sigma_{h0}), \quad h = 1,\ldots,H$,
3. $\beta_f^{(h)} \sim b_1(\beta_f^{(h)}|\Sigma_{f0}) = MVN(0, \Sigma_{f0}), \quad f|h = 1,\ldots,N_h,\ h = 1,\ldots,H$
4. $\sigma^2 \sim b_4(\sigma^2|\nu,\lambda) = = \mathrm{IGa}(\nu/2, \nu\lambda/2)$,
5. $\delta_F^2 \sim b_5(\delta_F^2|\nu_F,\lambda_F) = \mathrm{IGa}(\nu_F/2, \nu_F\lambda_F/2)$,

which are all assumed to be independent. Here MVN denotes the multivariate normal distribution, IGa the inverse Gamma distribution (e.g. Robert 1994, p.153). The values of ν, λ, ν_F and λ_F do not have any significant effect on the analysis so will be specified in the study (see Appendix A). The hyper-prior distributions for the hyper-parameters in the above prior distributions are taken to be:

1. $\beta_0 \sim b_0(\beta_0|b,\Sigma) = DMVN(b,\Sigma)$,
2. $\Sigma_0 \sim W_m^{-1}(\xi_0 + m + 1, \xi_0 R_0), \quad \xi_0 \geq m$
3. $\Sigma_{h0} \sim W_m^{-1}(\xi_{h0} + m + 1, \xi_{h0} R_{h0}), \quad \xi_{h0} \geq m,$
4. $\Sigma_{f0} \sim W_m^{-1}(\xi_{f0} + m + 1, \xi_{f0} R_{f0}), \quad \xi_{f0} \geq m,$

which are also assumed to be independent. Here $W_m^{-1}(\xi_\cdot + m + 1, \xi_\cdot R_\cdot)$ denotes an inverted Whishart distribution for an $m \times m$ random matrix $\Sigma_\cdot$ with degrees of freedom $\xi_\cdot + m + 1$ and positive definite matrix $R_\cdot$. It can be shown that $E(\Sigma_\cdot^{-1}) = R_\cdot^{-1}$ and $E(\Sigma_\cdot) = \{(\xi_\cdot - m - 1)\xi_\cdot R_\cdot\}^{-1}$ if these exist (Muirhead 1981, p.113). The values of $b, \Sigma, \xi_0, R_0, \xi_{h0}, R_{h0}, \xi_{f0}$ and R_{f0} will be specified in Appendix A. We denote the last three densities by $w_0(\Sigma_0|\xi_0, R_0)$, $w_{h0}(\Sigma_{h0}|\xi_{h0}, R_{h0})$ and $w_{f0}(\Sigma_{f0}|\xi_{f0}, R_{f0})$ respectively.

In order to simplify the presentation, we use Θ and γ respectively to denote the parameters and hyper-parameters in the hierarchical model defined here. Namely, $\Theta = (\beta^t, \vec{\beta}_h^t, \vec{\beta}_f^t, \sigma^2, \delta_F^2)^t$ and $\gamma = (\beta_0, \Sigma_0, \Sigma_{h0}, \Sigma_{f0})$, where $\vec{\beta}_h = (\beta_1^t, \cdots, \beta_H^t)^t$, $\vec{\beta}_f = (\vec{\beta}_f^{(1)t}, \cdots, \vec{\beta}_f^{(H)t})^t$ and $\vec{\beta}_f^{(h)} = (\beta_1^{(h)t}, \cdots, \beta_{N_h}^{(h)t})^t$, $h = 1, \cdots, H$. Note that Θ is a vector of dimension $(1 + H + \sum_{h=1}^H N_h)m + 2$ and γ is an $m \times (1 + 3m)$ matrix.

The equation (3.1) together with the prior distribution of Θ and hyper-prior distribution of γ comprise a hierarchical random effects model for the CGG triplet repeat sequence. In order to make inference about Θ and γ, and in particular β_h and $\beta_f^{(h)}$ from the data, we need the posterior distributions of β_h and $\beta_f^{(h)}$. Because of their mathematical complexity we will not be able to make inferences directly from these posterior distributions. Rather we will first generate a sample from these distributions and then make inferences from the sample. The Markov chain Monte Carlo (MCMC) technique has shown to be powerful in simulating a sample from those probability distributions which are analytically intractable. Recall that the basic idea of MCMC is to use a relatively simple transition probability kernel to generate a Markov chain in such a way that its invariant distribution is that from which we want to generate a sample. After a sufficient number of generations of the chain have been simulated, subsequent generations can be regarded as a sample from the target distribution. Two most basic algorithms used in MCMC methods are Gibbs sampling and Metropolis-Hastings

algorithm. We use a mixture of these two algorithms which is called MCMC block-at-a-time algorithm (Chib and Greenberg, 1995).

Several methods are available (Robert 1998) for monitoring the convergence to stationarity of the resulting Markov chain. We employ the Gelman-Rubin statistic (Gelman and Rubin 1992) which measures the correlation between the within- and between-chain variations of the simulated multiple Markov chains. When the Gelman-Rubin statistic becomes very close to 1 (usually less than 1.2 or 1.1 is enough in practice), the chains can be regarded as having achieved convergence. Once a sample has been generated from the joint posterior distribution, it can be examined to detect the random effects of haplotypes and families. The sample can also be used to simulate the posterior predictive distribution for the response variable CGG repeats which, after comparison with the true CGG observations, allows determination of the goodness of fit of the model. The various joint and conditional densities and the MCMC procedure used to fit the model are described in the appendices.

4 Results

The data of Murray et al (1997) contain observations on CGG repeat sequences for 124 individuals and their parents after the probands were omitted. The remaining 124 individuals are the offspring of 86 observed parents who come from 57 different families, nested in 18 different haplotypes. In addition, 10 of the 86 observed parents are fathers who between them have 18 offspring. When applying the hierarchical model of section 3, the parameter vector Θ contains 762 components and the hyper-parameter γ is a 10×31 matrix. For each of these parameters we generated 3 Markov chains of length 10,000 from which the posterior samples were formed.

We first check the convergence of the simulated Markov chains. It was found that for the current model at most 2400 transitions would be sufficient for the simulated Markov chain to be stationary. In the simulations the slowest convergence occurred at the Markov chain of $\Theta(88) = \beta_8(8)$(component 8 of haplotype 724) among all parameters and at $\gamma(6,17) = \Sigma_{h0}[6,6]$ (the 6th diagonal element of Σ_{h0}) among all hyper-parameters. Figure 1 displays the sub-chain of a simulated Markov chain (by filter $[k/50]$, i.e., taking every 50th value) for each of $\beta_8(8)$ and $\Sigma_{h0}(6,6)$ and their Gelman-Rubin statistic value sequences based on the 3 parallel chains.

To form a posterior sample for each parameter or hyper-parameter, we discard the first half of each of the 3 generated chains and combined the rest into one sequence; then we filter this sequence by the operator $[k/50]$ and take the resulting sequence of length 300. Shown in Figure 2 are histogram densities of the posterior samples for the 10 components of the haplotype

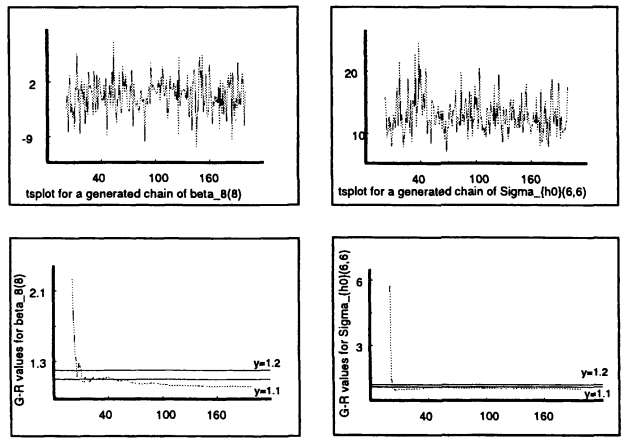

Figure 1: *Row 1 gives Markov chains generated for* $\beta_8(8)$ *and* $\Sigma_{h0}(6,6)$. *Row 2 gives the corresponding Gelman-Rubin statistic values based on the three parallel chains.*

8 effect β_8 (the 81st–90th elements of Θ, i.e. $\Theta[81:90]$ and the 10 components of the family effects $\beta_1^{(2)} = \Theta[261:270]$ (effects of the first family in the second haplotype which are the 261–270th components of Θ). These histograms are also typical of other haplotype and family effects.

The haplotype effects can be analysed by examining whether or not there is any heterogeneity among the marginal posterior distributions of $\{\beta_h(i), h = 1, \cdots, 18\}$ $(i = 1, \cdots, 10)$. Shown in the first column of Figure 3 are the Xbar-chart and S-chart with 95% confidence level for posterior samples of $\{\beta_1(8), \cdots, \beta_{18}(8)\}$, and the histogram density and the QQ-plot for the corresponding posterior means. The other two columns of Figure 3 are for $\{\beta_1(6), \cdots, \beta_{18}(6)\}$ and $\{\beta_1(3), \cdots, \beta_{18}(3)\}$. Figure 3 gives evidence of haplotype effects in the expansion process.

The family effects given the haplotype effects can be similarly analysed based on the marginal posterior distributions of $\vec{\beta}_f(i)$ $(i = 1, \cdots, 10)$. The three columns in Figure 4 give the results of these effects from the posterior samples of $\vec{\beta}_f(8)$'s, $\vec{\beta}_f(4)$'s and $\vec{\beta}_f(1)$'s respectively. Again family effects are evident in Figure 4.

The goodness of fit of the model can be assessed by comparing the response observations $g(\mathbf{Y})$ to their posterior predictive distributions (Gelman et al 1995, sec. 6.3). For each response observation $g(Y_{hfkl})$ a sample of size 300, regarded as the replicates of $g(Y_{hfkl})$ which could have been observed,

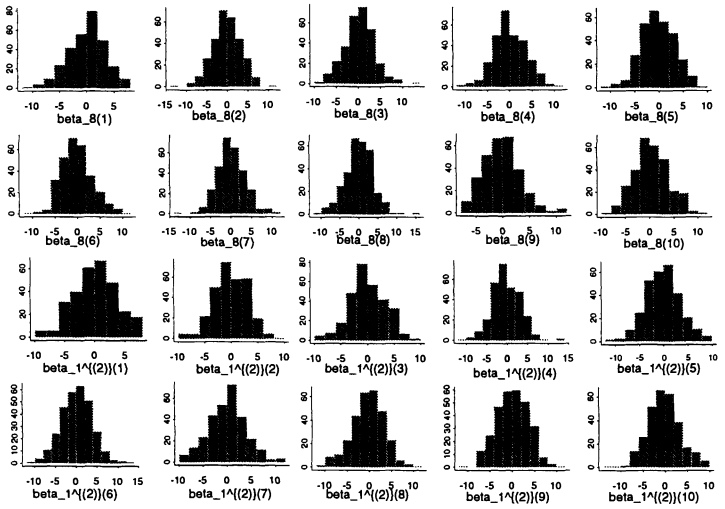

Figure 2: *Histograms for all the 20 components of* β_8 *and* $\beta_1^{(2)}$ *effects.*

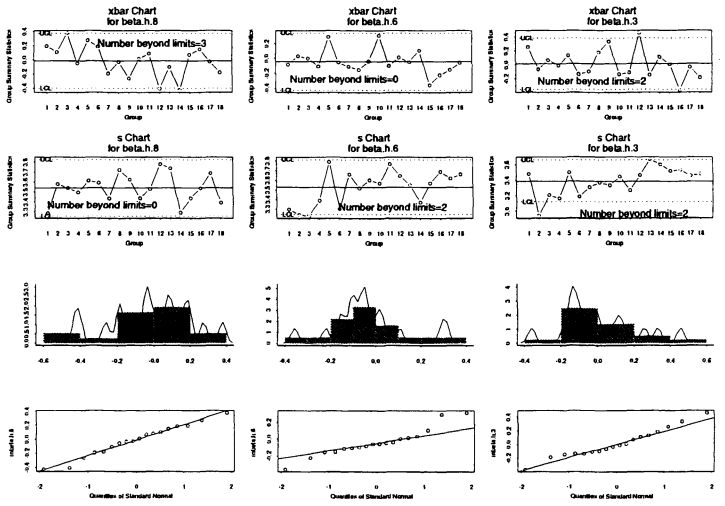

Figure 3: *Effects of haplotype factors evaluated at components 8, 6 and= 3.*

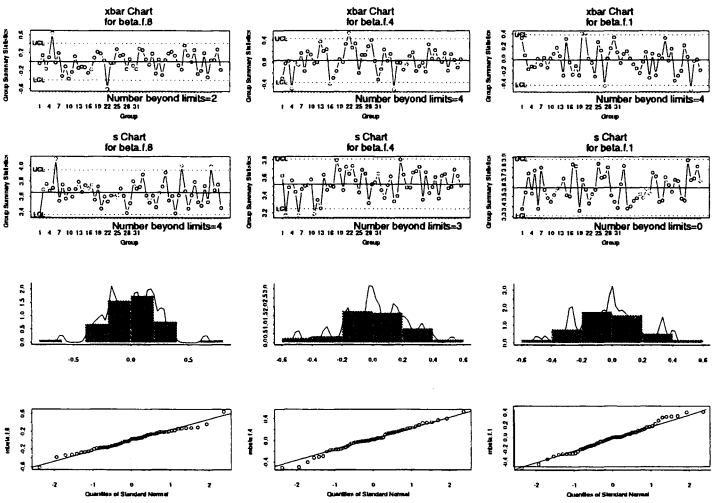

Figure 4: *Effects of family factors evaluated at components 8, 4 and 1.*

was generated from its posterior predictive distribution. Then the Bayes p-value was calculated for each sample, which is the proportion of the replicates that are greater than $g(Y_{hfkl})$. The histogram density for the replicates of $\log\log Y_9$, which is typical for all 124 response observations, is given in Figure 5. A histogram of the 124 Bayes p-values is also given in Figure 5. Since most of the Bayes p-values are very close to 0.5 (all but one are between 0.4 and 0.563), it implies that the response observations are typical under the posterior predictive distributions. In addition, we provide in Figure 5 the plots of $\overline{\log\log \mathbf{Y}^{new}}|\mathbf{Y}$ — the sample means of the posterior predictive distributions against= $\log\log \mathbf{Y}$, and $E(\log\log \mathbf{Y}|\bar{\Theta}^{pos\cdot})$ — the fitted response values using the posterior means of the distributions of the β's against $\log\log \mathbf{Y}$. These plots show the reasonable fit of the proposed model to the data which is also supported by the values of the goodness of fit statistic computed from (2) below. Actually it was found that $||\log\log \mathbf{Y} - \overline{\log\log \mathbf{Y}^{new}}|\mathbf{Y}||^2 = 14.10$ and $||\log\log \mathbf{Y} - E(\log\log \mathbf{Y}|\bar{\Theta}^{pos\cdot})||^2 = 7.26$. However, the histogram density in Figure 5 shows that the posterior predictive distribution of $\mathbf{Y}$ has a quite wide range and a large variance, particularly considering the nature of a $\log\log(\cdot)$ transformation for the response Y. This implies that the proposed model has a large variation in its predictability although it is not the main interest of this paper. One could possibly obtain better predictions if another transformation $g(Y)$ or a discrete model such as a Poisson distribution for Y were used.

Finally we propose to evaluate the goodness of fit of the model by the

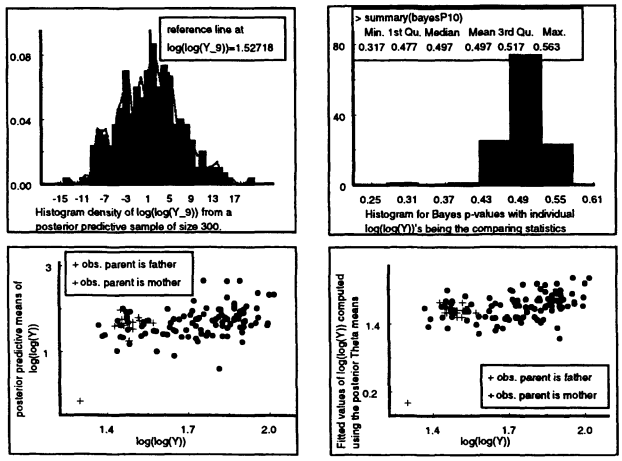

Figure 5: *Model assessment based on posterior predictive distributions.*

following statistic

$$T = \frac{1}{N}\sum_{i=1}^{N}\sum_{hfkl}\frac{(Y_{hfkl} - E(Y_{hfkl}|\Theta_i) =)^2}{Var(Y_{hfkl}|\Theta_i)} \tag{4.1}$$

where $\Theta_1, \cdots, \Theta_N$ is a posterior sample of Θ. Although no rigorous justification is available, it sounds sensible to approximate T by a χ^2_{124} distribution without being too erroneous. Based on the posterior Θ sample generated in our study we found $T = 87.26$, which corresponds to 0.5-percentile of a χ^2_{124} and thus indicates that the proposed model gives an adequate fit.

5 Discussion

Many of the advances in inference for stochastic processes over the past few decades have arisen from branching processes and time series models. A population model for the transmission of dynamic mutations combines time series models with branching processes, for, as well as (3.1) holding the number of offspring w_{hfk} of individual hfk has a distribution that depends on both the sex of individual hfk and the value of Y_{hfk}. The outstanding problem is to find conditions under which the population model has a stationary distribution.

We have concentrated on developing and fitting time series type models for the transmission of a dynamic mutation given the observed haplotypes

and family structures. The model can be viewed as an extension of the simpler bifurcating autoregressive models for cell lineages. The work on cell lineage studies starting with Cowan (1984), Cowan & Staudte (1986) and others considers the bifurcating autoregressive process for cell lineage studies. Cell lineage trees arise from the binary splitting of cells and the processes of interest are time series, typically low order autoregresive models, down lines of descent with correlated errors for sister cells and in later models (Huggins & Basawa 1999) correlations between individuals in the same generation. Bui & Huggins (1999) considered random effects models for the bifurcating autoregressive(1) model. The processes of interest here have a random number of offspring rather than just the two of the bifurcating autoregressive process and the models down each line of descent are more sophisticated than the ARMA type processes. Moreover, individuals with a given identified haplotype may differ at other gene sites, which naturally gives rise to the random effects models. Hence we take the transition rates to be random, which are common for individuals in the same families and there is communality between unrelated individuals with the same haplotype.

Markov models are typically used to model the transmission of a characteristic from parent to offspring. That is, the genetic make up of the offspring only depends on that of the parents. The indiscriminant use of such Markov models may be unrealistic in the branching situations that occur in family studies. For example, in related work on cell lineage studies, Staudte et al (1996) determined that the correlation between cousins in the data examined by them was larger than what was predicted by AR(1) type Markov models. In Markovian AR(1) models for cell lineages, correlations drop off as powers of the mother-daughter correlation θ. That is, mother-daughter correlation is θ, the= sister-sister correlation is θ^2 (plus environment effects) the cousin–cousin correlation is θ^4 etc. Huggins & Basawa (1999) proposed several models which allow higher correlations between cousins and other more distant relatives including an AR(2) type model as well as random effects models of generation effects. This situation may also occur in dynamic mutations unless genetic covariates that affect the expansion rates are taken into account. This motivated our hierarchical random effects model.

The models examined here are based on transmissions over one or two generations and may not reflect the full complexity of the dynamic mutations. Moreover, the processes are sparsely observed and as the process is unobserved in its initial stages, there is little information concerning the mutation rate currently available. The nested random effects, which depend on haplotypes that are themselves evolving according to some stochastic process, add another level of complexity. However, the haplotypes occur on a portion of the gene which does not appear to code for a protein hence they should not be exposed to selection so that their evolution should be depen-

dent only on random mutations and rare recombinations and be simpler than that of the the FMR1 gene itself.

Future work will involve the collection and modelling of data on distantly related affected individuals, where the evolution of the haplotypes will need to be considered. The strength of the relationship between individuals will be determined either through family trees or through genetic markers with known mutation rates.

It is not necessary to use MCMC to estimate random effects although it was convenient in the present example. For example, Park & Basawa (2000) have developed an optimal estimating function approach. However, they have not yet fully developed the inferential procedures.

The use of cubic splines has advantages in that the models are linear and may to some extent be regarded as non-parametric. Moreover, it was not necessary to introduce a threshold as in Huggins et al (1998). However, the non-linear models of Huggins et al (1998) did have some biological advantages in that they more naturally modelled the expansions in the meiotic and mitotic phases. In the non-linear case it was possible to let the parameters in the model depend on the size of the parents repeat sequence which is more difficult in linear models.

The increasing interest in molecular genetics will result in new and complex stochastic processes. However, as in cell lineage studies and the example considered here, these new processes are combinations of familiar simpler stochastic processes with perhaps extended error structures. Nevertheless, it appears certain that many new problems in statistical inference and applied probability will arise in this area.

Acknowledgements This research was supported by NICHID grant # HD36071 and an Australian Research Council small grant.

References

[1] Ashley, E.A. & Sherman, S.L (1995) Population dynamics of a meiotic/mitotic expansion model for the fragile X syndrome. *Am. J. Hum. Genet.* **57:** 1414–1425.

[2] Bui, Q.M. & Huggins, R.M. (1999) The random coefficient bifurcating autoregression model in cell lineage studies. *J. Stat. Plan. & Inf.* **81**; 253–262.

[3] Bat O, Kimmel M, & Axelrod DE (1997) Computer simulation of expansions of DNA triplet repeats in the fragile X syndrome and Huntington 92s disease. *J.Theor.Biol.* 188: 53-67

[4] Chib, Siddhartha and Greenberg, Edward (1995) Understanding the Metropolis-Hastings Algorithm *American Statistician* V. **49**, No. 4 327-335.

[5] Cowan, R. (1984). Statistical concepts in the analysis of cell lineage data. *Proceedings of the 1983 Workshop on Cell Growth and Division.*, 18–22, LaTrobe University.

[6] Cowan, R. and Staudte, R.G. (1986). The bifurcating autoregression model in cell lineage studies. *Biometrics.* 42, 769–783.

[7] Dana C. Crawford, Charles E. Schwartz, Kellen L. Meadows, James L. Newman, Lisa F. Taft, Chris Gunter, W. Ted Brown, Nancy J. Carpenter, Patricia N. Howard-Peebles, Kristin G. Monaghan, Sarah L. Nolin, Allan L. Reiss, Gerald L. Feldman, Elizabeth M. Rohlfs, Stephen T. Warren, and Stephanie L. Sherman (2000) Survey of the Fragile X Syndrome CGG Repeat and the Short-Tandem-Repeat and Single-Nucleotide-Polymorphism Haplotypes in an African American Population. *Am. J. Hum.Genet.* **66** 480–493.

[8] Dierckx, P.(1993). *Curve and surface fitting with splines*, Clarendon: New York.

[9] Eichler, E.E., Macpherson, J.N., Murray, A. Jacobs, P.A., Chakravarti, A. and Nelson, D.L. (1996) Haplotype and interspersion analysis of the FMR1 CGG repeat identifies two different mutational pathways for the origin of the fragile X syndrome. *Human Molecular Genetics.* **5** 319–330.

[10] Gawel B & Kimmel M(1996) The iterated Galton-Watson process. *J.Appl.Prob.* 33: 949-959.

[11] Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**,457-511.

[12] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*, Chapman & Hall: London.

[13] Huggins, R.M. & Loesch, D.Z. (1998) On the analysis of mixed longitudinal growth data. *Biometrics*

[14] Huggins, R.M., Loesch, D.Z., & Sherman, S.L. (1998) A branching nonlinear autoregressive model for the transmission of the fragile X dynamic repeat mutation. *Ann. Hum. Genet.* **62** 337-347.

[15] Huggins, R.M. & Basawa, I.V. (1999) Extensions of the bifurcating autoregressive model for cell lineage studies. *J.Appl.Prob.* **36** 1225–1233.

[16] Morris, A., Morton, N.E., Collins, A., Macpherson, J., Nelson, D. and Sherman, S. (1995) An n-allele model for progressive amplification in the FMR1 locus. *Proc. Natl. Acad. Sci. USA* **92** 4833-4837.

[17] Morton, N.E. and Macpherson, J.N. (1992) Population genetics of the fragile X syndrome: multiallelic model for the FMR1 locus. *Proc. Natl. Acad. Sci. USA* **89** 4215–4217.

[18] Muirhead (1981).*Aspects of multivariate statistical theory*, John Wiley: New York.

[19] Murray, M.A., Macpherson, J.N., Pound, M.C., Sharrock, A., Youings, S.A., Dennis, N.R., McKechnie, N., Lineham, P., Morton, N.E., Jacobs, P.A.(1997). The role of size, sequence and haplotype in the stability of FRAXA and FRAXE alleles during transmission. *Hum. Mol. Genet.* **6**:173-84.

[20] Park, Jeong-gun & Basawa, I.V. (2000) Optimal estimating equations for mixed effects models with dependent observations. Technical report 2000-16, Department of Statistics, University of Georgia.

[21] Robert, C.P. (1994) *The Bayesian Choice.* New York: Springer-Verlag

[22] Robert, C.P. (1998). *Discretization and MCMC Convergence Assessment*, Springer-Verlag: New York.

[23] Staudte, R.G., Zhang, J., Huggins, R.M. and Cowan, R. (1996). A reexamination of the cell lineage data of E. O. Powell. *Biometrics.*, 52, 1214-1222.

Appendices

In the appendices we derive the joint and conditional densities required to estimate the model parameters using MCMC, and then give the MCMC algorithm.

A Joint Densities

Using the notations introduced in the paper, it follows that the joint conditional density of $g(\mathbf{Y})$ given Θ, γ and the observed ancestor of each family is

$$
\begin{aligned}
p\{g(\mathbf{Y})|\Theta,\gamma\} \quad & p\{g(\mathbf{Y})|\Theta\}\prod_{h=1}^{H}\prod_{f=1}^{N_h}\prod_{k=1}^{n_{h=f}}\prod_{l=1}^{w_{hfk}} p\{g(Y_{hfkl})|\beta,\beta_h,\beta_f^{(h)},\sigma^2,\delta_F^2\} \\
& \prod_{h=1}^{H}\prod_{f=1}^{N_h}\prod_{k=1}^{n_{hf}}[2\pi(\delta_F^2 I_F(hfk)+\sigma^2)]^{w_{hfk}/2} \\
& \times \exp\left\{-\sum_{l=1}^{w_{hfk}}\frac{\{g(Y_{hfkl})-X_{hfk}[\beta+(\beta_h+\beta_f^{(h)})I_F(hfk)]\}^2}{2(\delta_F^2 I_F(hfk)+\sigma^2)}\right\}
\end{aligned}
\tag{A.1}
$$

The joint prior density of the parameter Θ given γ is

$$
\begin{aligned}
\pi(\Theta|\gamma) \quad & \prod_{h=1}^{H}\prod_{f=1}^{N_h} b_1(\beta_f^{(h)}|\Sigma_{f0})\cdot\prod_{h=1}^{H} b_2(\beta_h|\Sigma_{h0})\cdot \\
& b_3(\beta|\beta_0,\Sigma_0)b_4(\sigma^2|\nu,\lambda)b_5(\delta_F^2|\nu_F,\lambda_F) \\
= & \prod_{h=1}^{H}\prod_{f=1}^{N_h}(2\pi)^{-m/2}|\Sigma_{f0}|^{-1/2}\exp\left\{-\frac{1}{2}\beta_f^{(h)t}\Sigma_{f0}^{-1}\beta_f^{(h)}\right\} \\
& \times\prod_{h=1}^{H}(2\pi)^{-m/2}|\Sigma_{h0}|^{-1/2}\exp\left\{-\frac{1}{2}\beta_h^t\Sigma_{h0}^{-1}\beta_h\right\} \\
& \times(2\pi)^{-m/2}|\Sigma_0|^{-1/2}\exp\left\{-\frac{1}{2}(\beta-\beta_0)^t\Sigma_0^{-1}(\beta-\beta_0)\right\} \\
& \times\frac{(\nu\lambda/2)^{\nu/2}}{\Gamma(\nu/2)\sigma^{2(\nu/2+1)}}e^{-\nu\lambda/2\sigma^2}\frac{(\nu_F\lambda_F/2)^{\nu_F/2}}{\Gamma(\nu_F/2)\delta_F^{2(\nu_F/2+1)}}e^{-\nu_F\lambda_F/2\delta_F^2} \\
= & (2\pi)^{-\frac{m}{2}(\sum_{h=1}^{H}N_h+H+1)}|\Sigma_{f0}|^{-\frac{1}{2}\sum_{h=1}^{H}N_h} \\
& \exp\left\{-\frac{1}{2}tr\{\Sigma_{f0}^{-1}(\sum_{h=1}^{H}\sum_{f=1}^{N_h}\beta_f^{(h)}\beta_f^{(h)t})\}\right\}
\end{aligned}
$$

$$\times|\Sigma_{h0}|^{-H/2}\exp\left\{-\frac{1}{2}tr\{\Sigma_{h0}^{-1}(\sum_{h=1}^{H}\beta_h\beta_h^t)\}\right\}$$
$$\times|\Sigma_0|^{-1/2}\exp\left\{-\frac{1}{2}tr\{\Sigma_0^{-1}(\beta-\beta_0)(\beta-\beta_0)^t\}\right\}$$
$$\times\frac{(\nu\lambda/2)^{\nu/2}}{\Gamma(\nu/2)\sigma^{2(\nu/2+1)}}e^{-\nu\lambda/2\sigma^2}\frac{(\nu_F\lambda_F/2)^{\nu_F/2}}{\Gamma(\nu_F/2)\delta_F^{2(\nu_F/2+1)}}e^{-\nu_F\lambda_F/2\delta_F^2} \tag{A.2}$$

The joint prior density of the hyper-parameter γ is

$$\begin{aligned}\pi(\gamma) &= w_{f0}(\Sigma_{f0}|\xi_{f0},R_{f0})\cdot w_{h0}(\Sigma_{h0}|\xi_{h0},R_{h0})\cdot w_0(\Sigma_0|\xi_0,R_0)\cdot b_0(\beta_0|b,\Sigma)\\ &= \frac{(\xi_{f0}/2)^{m\xi_{f0}/2}|R_{f0}|^{\xi_{f0}/2}}{\Gamma_m(\xi_{f0}/2)}|\Sigma_{f0}|^{-(\xi_{f0}+m+1)/2}\exp\left\{-\frac{\xi_{f0}}{2}tr(\Sigma_{f0}^{-1}R_{f0})\right\}\\ &\quad\times\frac{(\xi_{h0}/2)^{m\xi_{h0}/2}|R_{h0}|^{\xi_{h0}/2}}{\Gamma_m(\xi_{h0}/2)}|\Sigma_{h0}|^{-(\xi_{h0}+m+1)/2}\exp\left\{-\frac{\xi_{h0}}{2}tr(\Sigma_{h0}^{-1}R_{h0})\right\}\\ &\quad\times\frac{(\xi_0/2)^{m\xi_0/2}|R_0|^{\xi_0/2}}{\Gamma_m(\xi_0/2)}|\Sigma_0|^{-(\xi_0+m+1)/2}\exp\left\{-\frac{\xi_0}{2}tr(\Sigma_0^{-1}R_0)\right\}\\ &\quad\times(2\pi)^{-m/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\beta_0-b)^t\Sigma^{-1}(\beta_0-b)\right\}\end{aligned} \tag{A.3}$$

where $\Gamma_m(\cdot)$ is the multivariate gamma function (see Muirhead 1981, pp. 61-63).

In the simulations, the value of b was taken to be the estimate of β in the spline regression model $g(Y_{hfkl}) = X_{hfk}\beta$. We also used the following values in the simulation in such a way that they should have very weak effects on the simulation results: $\nu = \lambda = \nu_F = \lambda_F = 10$,= $\Sigma = R_0 = R_{h0} = R_{f0} = 10I_{10}$ with I_{10} being the 10×10 identity matrix, and $\xi_0 = \xi_{h0} = \xi_{f0} = 50$.

B Posterior Conditional Densities

In order to simulate the posterior distributions using MCMC methods we need a number of posterior conditional densities.

The joint posterior density of Θ, γ given the observed ancestor of each family is

$$\pi(\Theta,\gamma|\mathbf{Y}) = \frac{p\{g(\mathbf{Y})|\Theta,\gamma\}\pi(\Theta|\gamma)\pi(\gamma)}{\int\int p\{g(\mathbf{Y})|\Theta,\gamma\}\pi(\Theta|\gamma)\pi(\gamma)d\Theta d\gamma} \tag{B.1}$$

The conditional posterior density of Θ given γ and the observed ancestor of each family is

$$\pi(\Theta|\mathbf{Y},\gamma) = \frac{p(g(\mathbf{Y}),\Theta,\gamma)}{\int p(g(\mathbf{Y}),\Theta,\gamma)d\Theta} = \frac{p\{g(\mathbf{Y})|\Theta,\gamma\}\pi(\Theta|\gamma)}{\int p\{g(\mathbf{Y})|\Theta,\gamma\}\pi(\Theta|\gamma)d\Theta} \tag{B.2}$$

The conditional posterior density of γ given Θ and the observed ancestor of each family is

$$\pi(\gamma|\mathbf{Y},\Theta) = \frac{p\{g(\mathbf{Y})|\Theta,\gamma\}\pi(\Theta|\gamma)\pi(\gamma)}{\int p\{g(\mathbf{Y})|\Theta,\gamma\}\pi(\Theta|\gamma)\pi(\gamma)d\gamma} \tag{B.3}$$

C Simulating the Posterior Densities by MCMC

To simulate the posterior distributions of the β_h's and $\beta_f^{(h)}$'s, we first generate a random sample for (Θ,γ) from the posterior density $\pi(\Theta,\gamma|\mathbf{Y})$. Then we extract those values of $\vec{\beta}_h$ and $\vec{\beta}_f$ in the sample, which clearly comprises a random sample from the posterior distributions of $(\vec{\beta}_h^t,\vec{\beta}_f^t)^t$. So the question is how to generate a random sample from the posterior density $\pi(\Theta,\gamma|\mathbf{Y})$.

To answer this question we apply an Markov chain Monte Carlo method (MCMC) which uses a Metropolis-Hastings acceptance–rejection algorithm alternately to the two blocks of the random quantity (Θ,γ). In other word, this is the so-called MCMC block-at-a-time algorithm (refer to Chib and Greenberg, 1995). We list the algorithm in the following:

- Obtain the initial values $(\gamma^{(0)},\Theta^{(0)})$.
  1. Generate a matrix $\gamma^{(0)}$ from the prior $\pi(\gamma)$.
  2. Generate a vector $\Theta^{(0)}$ from the prior $\pi(\Theta|\gamma^{(0)})$.
- Update $(\gamma^{(j)},\Theta^{(j)})$ to $(\gamma^{(j+1)},\Theta^{(j+1)})$. Repeat for $j = 0,1,\cdots,N-1$.
  1. Use an M-H acceptance–rejection algorithm to generate a matrix $\gamma^{(j+1)}$ from the conditional posterior density $\pi(\gamma|\mathbf{Y},\Theta^{(j)})$:
     - Generate an initial matrix $\gamma_0^{(j)}$ from $\pi(\gamma)$.
     - Repeat for $i = 0,1,\cdots,I-1$.
     - Generate a matrix γ' from $\pi(\gamma)$ and u from the uniform distribution $\mathcal{U}(\prime,\infty)$.
     - If $u \le \alpha_\gamma(\gamma_i^{(j)},\gamma')$, set $\gamma_{i+1}^{(j)} = \gamma'$. The acceptance probability is defined by

$$\alpha_\gamma(\gamma_i^{(j)},\gamma') = \min\left\{\frac{\pi(\Theta^{(j)}|\gamma')}{\pi(\Theta^{(j)}|\gamma_i^{(j)})},1\right\}.$$

     - If $u > \alpha_\gamma(\gamma_i^{(j)},\gamma')$, set $\gamma_{i+1}^{(j)} = \gamma_i^{(j)}$.
     - Finally set $\gamma^{(j+1)} = \gamma_I^{(j)}$.
  2. Use an M-H acceptance–rejection algorithm to generate a vector $\Theta^{(j+1)}$ from the conditional posterior density $\pi(\Theta|\mathbf{Y},\gamma^{(j+1)})$:

- Generate an initial vector $\Theta_0^{(j)}$ from $\pi(\Theta|\gamma^{(j+1)})$.
- Repeat for $q = 0, 1, \cdots, Q-1$.
- Generate a matrix Θ' from $\pi(\Theta|\gamma^{(j+1)})$ and u from the uniform distribution $\mathcal{U}(\prime, \infty)$.
- If $u \leq \alpha_\Theta(\Theta_q^{(j)}, \Theta')$, set $\Theta_{q+1}^{(j)} = \Theta'$. The acceptance probability is defined by

$$\alpha_\Theta(\Theta_q^{(j)}, \Theta') = \min\left\{\frac{p\{g(\mathbf{Y})|\Theta', \gamma^{(j+1)}\}}{p\{g(\mathbf{Y})|\Theta_q^{(j)}, \gamma^{(j+1)}\}}, 1\right\}.$$

- If $u > \alpha_\Theta(\Theta_q^{(j)}, \Theta')$, set $\Theta_{q+1}^{(j)} = \Theta_q^{(j)}$.
- Finally set $\Theta^{(j+1)} = \Theta_Q^{(j)}$.

- Return the sequence $\{(\gamma^{(0)}, \Theta^{(0)}), (\gamma^{(1)}, \Theta^{(1)}), \cdots, (\gamma^{(N)}, \Theta(N))\}$.

As in any MCMC method, the sequence generated by the above algorithm is actually a Markov chain with $\pi(\Theta, \gamma|\mathbf{Y})$ as its invariance density. So by ignoring the first N_0 values the sequence can be approximately regarded as a random sample from the posterior density $\pi(\Theta, \gamma|\mathbf{Y})$, provided that both N_0 and N are sufficiently large. In our simulations we found that when taking $I = Q = 50$ the simulated Markov chain became convergent after $N_0 > 48$ runs.